1ST ANNUAL CISON CONFERENCE PROCEEDINGS

Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges

https://doi.org/10.63255/001-01283.24/017x

CISON ACT 2022

Pre-conference Workshop

11–12 November 2024, National Bureau of Statistics, CBD, Abuja

Main Conference

13–15 November 2024, Exclusive Serene Hotel, Wuye, Abuja

Volume 001

Prepared by the Editorial Team of the
Journal of Chattered Institute of Statisticians of Nigeria (JCISON)

Editorial Disclaimer and Statement of Responsibility

This document is the official Proceedings of the 1st Annual National Conference of the Chartered Institute of Statisticians of Nigeria (CISON), held from 13th to 15th November 2024, at Exclusive Serene Hotel, Wuye, Abuja, Nigeria. The conference brought together statisticians, researchers, academics, and professionals from across the country and Africa to discuss innovative statistical approaches to understanding and solving Nigeria's economic challenges. It also offered an opportunity for brainstorming on the emerging trends and innovations in statistical science and practice during the scientific session.

The articles published in this document were presented during the scientific session, and they represent the scholarly contributions and viewpoints of the individual authors. While each paper has undergone a review process, the views and opinions expressed are solely those of the authors and do not necessarily reflect the official position of CISON, its Conference Planning Committee, or the Governing Council of the Institute.

CHARTERED INSTITUTE OF STATISTICIANS OF NIGERIA (CISON) GOVERNING COUNCIL MEMBERS, 2024

1.	Dr. Godday U. Ebuh	-	President, (Chairman)
2.	Prof. S. U. Gulumbe	-	Vice President
3.	Dr. Michael K. Mba	-	Representative, CBN
4.	Mr. A. C. Anyakorah	-	Representative, NBS
5.	Crown Prince Abubakar B. Afegbua		-Representative, NPC
6.	Mr. Jaafar Bello	-	Council Member
7.	Dr. Umaru Baba	-	Council Member
8.	Dr. (Mrs.) U. P. Ogoke	-	Council Member
9.	Prof. J. I. Mbegbu	-	Council Member
10.	Mr. Stephen Aiyedun	-	Council Member
11.	Dr. Bright Ajibade	-	Council Member
12.	Prof. P. E. Chigbu	-	Editor-in-Chief, JCISON
13.	Dr. N. V. Atoi	-	Managing Editor, JCISON
14.	Dr. O. S. Yaya	-	Associate Editor, JCISON
15.	Christopher M. Okafor	-	Registrar, (Secretary)

Foreword

The establishment of the Chartered Institute of Statisticians of Nigeria (CISON) by the CISON Establishment Act of 2022 marked a significant milestone in the advancement and formal regulation of the statistical profession in Nigeria. The Institute was created to promote excellence in statistical practice, foster adherence to international best practices, and serve as a national reference point for the coordination and issuance of official statistics.

In alignment with its mandate, CISON successfully convened its 1st Annual National Conference from the 13th to 15th of November 2024 at the Exclusive Serene Hotel, Wuye, Abuja. The theme of the conference, "**Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges**," was both timely and forward-looking, reflecting the urgent need to embed statistical thinking into national policy discourse and socioeconomic development strategies.

Preceding the main event, a two-day Pre-Conference Workshop was held from the 11th to 12th of November 2024 at the National Bureau of Statistics, Central Business District, Abuja. The workshop, titled "**Quantitative and Qualitative Approaches to Statistical Analysis**," intensive training in statistical methodologies using R software. It provided participants, comprising early-career statisticians, academics, analysts, and data enthusiasts, with practical tools to apply statistical techniques in diverse domains such as business, healthcare, education, governance, and the social sciences.

The main conference was distinguished by two insightful keynote addresses delivered by eminent experts. The first keynote, presented on behalf of Prof. Benedict Oramah (President of Afreximbank) by Dr. Sampawende Tapsoba (Deputy Chief Economist), explored the critical role of innovative statistical frameworks in diagnosing and addressing Nigeria's complex economic landscape. The second keynote, delivered by Dr. Olorunsola Olowofeso, Director General of the West African Monetary Institute (WAMI), further emphasised the strategic importance of data-driven policymaking in achieving economic resilience and regional integration. Over the course of the scientific sessions, high-quality papers were presented by scholars and practitioners from across the country and beyond.

We commend the dedication of the authors, reviewers, workshop facilitators, and the Conference Planning Committee for their invaluable contributions. We are confident that this publication will not only serve as a record of academic discourse but will also inspire further research, collaboration, and innovation in statistical science in Nigeria and beyond.

Dr. Godday U. Ebuh
CISON President & Chairman, CISON Governing Council.

Preface

The 1st Annual National Conference of the Chartered Institute of Statisticians of Nigeria (CISON) was held from the 13th to 15th of November, 2024, at the Exclusive Serene Hotel, Wuye, Abuja, Nigeria. This landmark event brought together statisticians, academics, policymakers, researchers, and industry professionals from Nigeria and around the globe to share knowledge, exchange ideas, and critically examine the role of statistics in addressing national development issues.

The central theme of the conference, "***Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges***," provided a platform for robust discussions on the integration of advanced statistical methods into evidence-based policy formulation, strategic planning, and sustainable economic development.

The articles contained in this volume were presented during the scientific sessions of the conference and have been accepted for publication in these proceedings. These scholarly contributions reflect diverse perspectives and methodological approaches to pressing economic issues, offering valuable insights for government agencies, development partners, academic institutions, and private sector stakeholders.

It is our expectation that the papers compiled herein will serve not only as a repository of current research but also as a catalyst for continued dialogue, collaboration, and innovation within the statistical community and beyond.

On behalf of the Editorial Board of the Journal of the Chartered Institute of Statisticians of Nigeria (JCISON), I express sincere appreciation to all contributors, discussants, and the conference organizing committee for their unwavering commitment to the success of this inaugural conference and its proceedings.

Prof. P.E. Chigbu
Editor-in-Chief, JCISON

Message from the Managing Editor

It is with great pleasure that I present to you the inaugural volume of the Proceedings of the 1st Annual National Conference of the Chartered Institute of Statisticians of Nigeria (CISON).

This publication represents a significant step forward in the Institute's commitment to fostering academic excellence and professional collaboration within the statistical community in Nigeria and across Africa. The central theme of the CISON maiden conference was timely, addressing the growing demand for data-driven policies in confronting the country's economic realities.

The diverse perspectives shared through keynote addresses, scientific sessions, and workshop interactions are documented in the pages of this publication. A total of nineteen papers are featured in this volume (Volume 001), offering both theoretical insights and applied research across sectors such as economics, governance, health, agriculture, education, and technology.

The publication process involved editorial oversight and coordination to ensure the integrity and quality of content. I extend my heartfelt appreciation to the authors for their scholarly contributions, the discussants for their diligence, and the editorial and technical teams for their unwavering support in making this volume possible.

May this work inspire future research and strengthen the role of statistics in shaping a more informed and prosperous Nigeria.

Dr. Atoi, N. V.
Managing Editor, JCISON

Table of Contents

COMMUNIQUÉ ISSUED BY THE CHARTERED INSTITUTE OF STATISTICIANS OF NIGERIA (CISON) AT THE END OF ITS 1ST ANNUAL CONFERENCE HELD AT EXCLUSIVE SERENE HOTEL, WUYE, ABUJA, NIGERIA FROM WEDNESDAY, 13TH NOVEMBER TO FRIDAY, 15TH NOVEMBER 2024.

1.0 Preamble

The Chartered Institute of Statisticians of Nigeria (CISON) held its 1st Annual Conference from 13th to 15th November 2024, at Exclusive Serene Hotel, Wuye, Abuja. The theme of the Conference was **"Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges."** Prior to the main conference was a Pre-conference workshop on "**Quantitative and Qualitative Approaches to Statistical Analysis,"** held from 11th to 12th November 2024 at the National Bureau of Statistics, Central Business District, Abuja. The workshop provided a comprehensive guide to both quantitative and qualitative statistical methods using R, focusing on practical examples and applications in different sectors such as business, healthcare, education, and social sciences, etc. During the conference, it was pointed out that CISON came into existence by virtue of CISON Establishment Act, 2022, to streamline the practice of the Statistics profession and coordinate the issuance of official Statistics based on international best practices. A total of 19 papers were presented at the conference.

The annual conference was specifically aimed at exploring innovative statistical methodologies and their applications in addressing Nigeria's economic challenges and to:

- Strengthen collaboration among producers, suppliers and users of statistics;
- Equip participants with hands-on experience with innovative statistical techniques and tools;
- Keep members updated on modern statistical methods and models; and
- Elect members of the 1st Council of CISON.

2.0 **Attendance**

The opening ceremony, which was held on 14th November, 2024 was attended by the Statistician General of the Federation and Chief Executive Officer of the National Bureau of Statistics, Prince Adeyemi Adeniran; the Governor of Central Bank of Nigeria (CBN), ably represented by Dr. Michael Mba; the Director General (DG) West African Monetary Institute (WAMI), Dr. Olorunsola E. Olowofeso; The Vice Chancellor, University of Nigeria Nsukka and Editor-in-

Chief, Journal of the Chartered Institute of Statisticians of Nigeria (JCISON), Prof. P. E. Chigbu; Mr. Charles Mordi; Dr. Chris O. Itsede; Mr. Godstime O. Eigbiremolen; the President of CISON, Dr. G. Ebuh; Fellows of the Chartered Institute of Statisticians of Nigeria including Prof. E. C. Nduka, Prof. Peter A. Osanaiye, Prof. (Mrs.) Joy C. Nwabueze, the Registrar of CISON, Christopher M. Okafor; the Chairman, National Organizing Committee (NOC), Dr. B.O. Amobi, and stakeholders from the Academia and Ministries, Departments and Agencies (MDAs)

The pre-conference attracted 58 participants while the main Conference recorded 285 participants.

Earlier on Wednesday, the Registrar, Mr. Christopher M. Okafor reminded the audience that the Chartered Institute of Statisticians of Nigeria (CISON) is the successor body to the Nigerian Statistical Association (NSA). He stated that CISON as a professional body is conceptualized as both a learned institute for Statistics and a professional community of statisticians that would promote the development of Statistical theory and practice in Nigeria. He stated that the motto of the Institute is, "Statistical Science, Knowledge and Professionalism" implying that we drive data to the knowledge domain through the professional application of standard statistical procedures. He also remarked that the Mission of CISON is "To advance statistical science and practice in Nigeria through professional excellence and innovation, while its Vision is "To be Nigeria's leading professional body for statisticians, driving excellence and impact in statistical practice.

The opening session was chaired by the Statistician General of the Federation and Chief Executive of the National Bureau of Statistics (NBS).

Delivering his welcome address, the President of CISON, Dr. Godday Ebuh, welcomed delegates to the 1st annual conference of CISON and highlighted that the discussions and activities from both the pre-conference and main conference will be documented in the Journal of CISON and the conference proceedings. He encouraged participants to beam their searchlight in Machine Learning, Big Data Analytics, Artificial Intelligence (AI) and Geospatial Data Analysis that assist in providing greater insights to data.

He thanked the past executives of the NSA for laying the groundwork that led to the birth of CISON.

Following the welcome address, Goodwill messages were received from corporate members and support organizations, such as the Central Bank of Nigeria (CBN), the SG and CEO of the NBS, and Dr. Kabir Nakaura, former board Chairman, NBS.

Two Keynote addresses were delivered; the first was by Prof. Benedict Oramah, ably represented by Dr. Sampawende Jules Tapsoba, Deputy Chief Economist & Director, Data Management & Model Development, Research and International Cooperation at the Afreximbank who emphasized the importance of innovative statistical approaches in addressing Nigeria's economic challenges. He urged CISON to leverage on technologies and capacity building for statisticians while also embracing Big, data and Artificial Intelligence (AI) in addressing challenges of the Nigerian economy as it proffers better opportunity. The speech highlighted Nigeria's economic growth, significant reforms, and the potential for future development, while also acknowledging the critical role of statistical data in shaping policies and fostering sustainable growth.

Dr. Olorunsola E. Olowofeso, delivered the second keynote address and noted the achievements of CISON's leadership. He also enumerated WAMI's contributions to statistical development in Nigeria to include, harmonization of statistical data in collaboration with CBN, and NBS, and urged the Institute to collaborate with international bodies on capacity building and data generation. He raised the following questions for CISON's consideration.

i. How can CISON support National Statistical System (NSS)?
ii. How can CISON contribute to creating standardized approaches to data transparency and reliability within Nigeria?
iii. How can real-time data collection, perhaps through mobile or satellite technology, be better leveraged to support agile and responsive policymaking? and
iv. What role should CISON play in establishing guidelines for data security and ethical data use? How can the Institute help members navigate the balance between data accessibility for innovation and the need to protect sensitive information?

3.0 Presentations

3.1 Panel Discussions

- Two panel discussions on the theme, **Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges** were held and aptly discussed by panelists drawn from the academia and corporate organizations. *The first*

panel discussion focused on the keynote speech by the representative of Prof. Benedict Oramah, Dr. Sampawende Jules Tapsoba, and was discussed by Prof. Joy Nwabueze, the moderator, Prof. P. E. Chigbu, Dr. Anthony Ayiola, Dr. Chris O. Itsede and Mr. Godstime O. Eigbiremolen. They highlighted the need to combine geospatial data and national surveys to improve data accuracy. The second panel discussion focused on the keynote speech by Dr. Olorunsola Olowofeso and discussed by Dr. Mba Micheal, the moderator, Prof. Peter A. Osanaiye, Dr. Tope Fasua, Mr. Charles Mordi, and Yakubu A. Bello, ably represented by Dr. Micheal Mba. They noted that Nigeria has serious statistical challenges and there is need for data insights and data driven policies. They, however, suggested a need for workshops, capacity building and collaborations.

3.2 Scientific Sessions

Parallel scientific sessions, chaired by seasoned academics and members of CISON, were held, and papers ranging from statistical theories to applications of statistical techniques for solving real-life problems were presented and coordinated by the JCISON Editorial Team.

4.0 Observations

From the presentations at the Conference, the following observations were raised:

a. It was observed that in tertiary institutions, non-statisticians are now teaching Statistics.
b. Statistics as a subject is no longer taught in secondary schools.
c. Statisticians in developing nations in general and Nigeria in particular, were challenged in collecting relevant data.
d. Most statistics produced within the framework of the National Statistical System often vary and are sometimes not being relied upon by users.
e. The venue of the annual conference, Exclusive Serene Hotel, Wuye, Abuja, was conducive with uninterrupted power supply.

5.0. Recommendations:

a. There is need for data visualization and professionalization of Planning, Research, and Statistics (PRS) workshops.
b. CISON should leverage on technologies and encourage capacity building for Statisticians while also embracing Big data, Artificial Intelligence (AI) and Geospatial Data Analysis in addressing challenges of the Nigerian economy as they proffer better solutions.
c. CISON should extend its regulatory power to university activities so as to prevent non statisticians from teaching Statistics especially, in other faculties.

d. CISON should define how statistics should be used and practiced in Nigeria so as to guide policy makers in decision making.
e. CISON should create groups/team of experts from different areas of statistics to formulate guidelines/rules for the practice of statistics in core areas of statistics. Hence, there is need for statistical quality control as a problem-solving technique.
f. There is need for NBS to collaborate with key agencies to generate micro-level data as well as state level GDP.
g. There is the need for synergy between stakeholders in the National Statistical System to ensure timely, accurate and reliable statistics.
h. CISON should work with the National Universities Commission (NUC) to ensure that CISON representatives are included in the accreditation of Statistics Departments in all Universities just as it is currently doing with the National Board for Technical Education.
i. CISON should collaborate with relevant agencies to include the teaching of statistics at primary school level and secondary school level.

Dr. Godday O. Ebuh — President, CISON

Christopher M. Okafor — Registrar, CISON

Keynote Speaker I: Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges

Delivered on behalf of Prof. Benedict Oramah (President of Afreximbank) by Dr. Sampawende Tapsoba (Deputy Chief Economist)

It is my esteemed pleasure to be here today, representing Professor Benedict Oramah, President and Chairman of the Board of Directors of the African Export-Import Bank (Afreximbank) Group, at the inaugural conference of the Chartered Institute of Statisticians of Nigeria (CISON), under the theme *"Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges."*

Prof. Oramah wished to deliver this keynote address personally but could not due to work demands. Nevertheless, he extends his good wishes and congratulations with a special acknowledgment to the CISON President, Dr. Godday Ebuh, and the Chairman of the National Organizing Committee, Dr. Boniface Amobi.

Ladies and Gentlemen,
Today's theme, "Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges," highlights the importance of data on our Continent.
I want to emphasize the importance of our gathering today as we discuss a crucial topic impacting policymaking across Africa. Your insights are vital for our journey together! Data and statistics play a critical role in shaping policies and fostering growth. Seeing the thriving data advocacy is encouraging, especially in Nigeria, Africa's largest oil producer.

Often called the "heartbeat of Africa," the continent's largest economy and most populous nation, Nigeria symbolizes potential and resilience with its unwavering promise for growth and development … Nigeria has experienced significant economic growth, expanding from around 9 trillion Naira in the 1960s to an estimated 78 trillion Naira by 2023. It now ranks among the top three economies in Africa, alongside South Africa and Egypt, reflecting its vital role in Africa's economic future.

Nevertheless, Nigeria faces significant economic challenges, with 38.9 percent of its population living in poverty—about 87 million people—making it the second highest globally after India. The wealth gap is stark, as the five richest Nigerians have a combined wealth of $29.9 billion, which could eradicate extreme poverty in the country. The economy heavily relies on oil exports, which account for over 90 percent of government revenue and 40 percent of GDP, leaving it vulnerable to global oil price fluctuations. Efforts to diversify into manufacturing and services have not yet stabilized the economy.

…Yet, Nigeria has the potential to realize its potential as an emerging economy as it is turning the corner with far-reaching macro-fiscal reforms.

Over the past year, the Nigerian government has implemented significant reforms amidst political challenges. Notably, it has adopted a market-driven exchange rate policy, merged official and parallel rates, and is phasing out costly petrol subsidies. To stabilize inflation expectations and restore confidence in the naira, the central bank raised its policy rate by 850 basis points and stopped financing fiscal deficits. These measures have led to inflation soaring to 32.70 percent in September 2024, the highest in nearly three decades, burdening many

families. These reforms aim to optimize fiscal resource allocation and increase investment in critical sectors like infrastructure, healthcare, education, and economic diversification to improve citizens' quality of life.

…Nigeria has the potential to achieve a one-trillion-dollar economy by 2030.

Such transformation could see Nigeria surpass the Netherlands, Thailand, and Malaysia. The Nigeria Agenda 2050 envisions an impressive annual average real GDP growth rate of 7.0 percent from 2021 to 2050. By 2050, per capita annual income is projected to soar to more than US$33,000, paving the way for the nation to embrace upper-income status confidently.

Ladies and Gentlemen,
With such complex economic challenges and potential, today's theme is timely and more relevant.

Nigeria is emphasizing the transformative power of data for economic growth through its National Digital Economy Policy and Strategy (NDEPS). This initiative promotes digital literacy and innovation, enabling data-driven decision-making in critical healthcare, education, and agriculture sectors. For instance, farmers use sensor data to optimize yields, and data platforms improve access to healthcare and financial services, highlighting the significant potential for progress. Innovative statistical approaches can help Nigeria address challenges and promote sustainable development. By enhancing data collection and analysis, these methods improve policymaking, stimulate business growth, and boost public services in key sectors while supporting marginalized communities.

Ladies and gentlemen,
I want to take a moment to recognize the significant strides the Nigerian National Bureau of Statistics (NBS) achieved.

The NBS has significantly improved Nigeria's statistical capacity through initiatives like the National Strategy for the Development of Statistics (NSDS). By promoting standards across the National Statistical System, they've established a solid foundation for progress. Initiatives such as the General Household Survey, Nigeria Living Standard Survey, and Labor Force Survey provide crucial data that supports policymaking. These achievements enhance data availability and reliability, enabling evidence-based decision-making and supporting national development planning.

Statistics is the vehicle through which we can empower policymakers to design the most accurate policies. In this regard, Nigeria has been exemplary. The National Bureau of Statistics (NBS) has maintained a commendable standard with statisticians nationwide, keeping them accountable for providing reliable data for economic planning, policy formulation, and performance monitoring across sectors. Yet, more remains to be done, which is why this event is significant. CISON can support NBS' efforts by promoting innovative and cutting-edge statistical approaches.

Distinguished colleagues,
Allow me to lay out some fundamental principles to embrace for harnessing data effectively and exploring innovative statistical methods to tackle complex challenges.

Data should be viewed as a **public good** because of its vital role in improving service delivery, informing policymaking, fostering innovation, and tackling urgent issues like poverty and climate change. Open data can drive economic development and enhance government accountability. There's a growing trend of organizations adopting open data policies, with cross-sector partnerships improving data sharing and interoperability. Researchers and policymakers can develop predictive models to address complex issues like disease outbreaks and climate scenarios by combining advanced statistical methods with open datasets. Promoting open data increases transparency, allowing various stakeholders to use government data for tailored solutions, while crowdsourced data enriches official statistics by providing deeper insights into local economic conditions.

Timeliness and accuracy in statistical reporting are crucial for effective decision-making, as they enhance processes, support monitoring, and increase data credibility. Timely information improves resource allocation and accountability and allows meaningful comparisons across countries. This ensures that data remains relevant and actionable, helping individuals, corporations, and policymakers make informed decisions. For example, Nigeria's GDP rebasing in 2014 was a significant step in refining economic statistics, providing a more accurate view of the economy essential for informed policymaking and investment.

Relying on high-quality data alone is not enough. There is a need to leverage the **Big Data and AI landscape**. As you know, advanced statistical methods are vital for deriving actionable insights that make policies effective and problem-specific. These methods help address key economic questions, such as diversifying away from oil and commodities. Statistics are critical in the current Big Data and AI landscape, driving FinTech innovation and transforming banking. Institutions like CISON should embrace these innovations and support Big Data analysis. The solutions to Nigeria and Africa's economic challenges lie in data. Insights from mobile networks and social media reveal economic activities and population movements. Machine learning enhances predictions for inflation and unemployment, assisting policymakers in crisis management and resource allocation.

Integrating **satellite imagery** into statistical analyses significantly benefits addressing economic challenges in Africa, particularly Nigeria. Geospatial data from satellites is crucial for quantifying economic disparities, assessing infrastructure deficits, and analyzing resource distribution. These insights enable targeted interventions to mitigate inequalities in specific regions or communities. Satellite imagery combined with Geographic Information Systems (GIS) allows monitoring agricultural yields, urbanization, and environmental changes. This data is essential for informing rural development strategies and improving food security, leading to more effective decision-making.

Enhancing the **proficiency of statisticians**, data scientists, and policymakers is crucial for the effective application of advanced statistical methodologies. Capacity-building initiatives should emphasize econometrics, machine learning, and data visualization, among other contemporary statistical techniques. Collaborating with international organizations can provide essential technical expertise and training, cultivating a highly skilled workforce capable of addressing Nigeria's complex data requirements.

Ladies and Gentlemen, Institutions like CISON must guide and ensure the collection of high-quality data and produce research that directs policy to effectively address the threatening macroeconomic challenges of youth unemployment, inflation, exchange rate volatility, and fiscal deficit, to mention a few. CISON’s commitment to ensuring the credibility and

standardization of statistical methods in Nigeria, training statisticians, upholding professional ethics, and providing expertise in data collation and analysis aligns with Afreximbank's vision for data management on the Continent.

Afreximbank welcomes collaborations to collect high-quality data to help African countries implement targeted policies to meaningfully reduce poverty and invigorate the private sector. Our financing of trade on the Continent and, more recently, the Caribbean Community and Common Market (Caricom) countries relies on high-quality data and analysis from the Afreximbank Research Department.

The Afreximbank Data Strategy is forward-looking and encompasses emerging technological innovations and advancements. Digital technology and artificial intelligence have transformed statistical sciences, giving rise to methodologies that can handle unprecedented levels of data, identify patterns, and predict outcomes. These innovations offer powerful tools for analyzing complex macroeconomic bottlenecks. The Continent must harness these innovations to stay caught up in the global race for a better and brighter future.

We at Afreximbank leverage these technologies to provide our clients with in-depth market insights. The private sector is critical to fostering resilient economic growth. Despite macroeconomic challenges and global uncertainty, the Nigerian private sector, driven by services, contributes significantly to GDP growth. With more reliable and comprehensive data and advanced statistical methods, the Nigerian private sector can model more accurate financial projections to attract foreign capital.

Therefore, we invite you, CISON, to work with the Afreximbank Research Department to further enhance statistical capacity through training, data collection standardization, and technology in Nigeria and across the Continent and Caricom to attract long-term financing to support economic growth and development through intra and extra-regional trade.

Esteemed members of the Chartered Institute of Statisticians of Nigeria (CISON), honored guests, and distinguished colleagues.

In closing, I urge us all to embrace the vast potential of data and statistics and continue pushing the boundaries. As statisticians, you are not just interpreting numbers. You are crafting Nigeria's growth trajectory. Let us forge a path where policy is borne out of high-quality data, and every data point is a building block for Nigeria's transformation and, indeed, Africa at large.

Thank you, and I look forward to a productive conference ahead.

Keynote Speaker II: Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges

Dr. Olorunsola Olowofeso
Director General of the West African Monetary Institute (WAMI)

Your Excellency, the President of the Federal Republic of Nigeria, Asiwaju Bola Ahmed Tinubu, GCFR; the Chief Host of this event, the Honourable Minister of the Federal Capital Territory, Chief Barrister Ezenwo Nyesom Wike; the President and Chairman of the Board of Directors of the African Export-Import Bank (AFREXIM), Professor Benedict Okey Oramah; the Honourable Minister of Budget and Economic Planning, Senator Abubakar Atiku Bagudu; the Statistician-General of the Federation and CEO of the National Bureau of Statistics, Prince Adeyemi Adeniran; the President of the Chartered Institute of Statisticians of Nigeria (CISON), Dr. G. U. Ebuh; distinguished members of the CISON Executive; esteemed guests; ladies and gentlemen.

I bring you greetings from the city of Accra, Ghana. It is a profound honour to stand before this distinguished gathering at the Inaugural Annual Conference of the Chartered Institute of Statisticians of Nigeria (CISON).

Today's event, centered on the theme *"Mainstreaming Innovative Statistical Approaches to Understanding and Solving Nigeria's Economic Challenges,"* provides an invaluable platform for statisticians, researchers, policymakers, and stakeholders to converge, share insights, and chart new directions for Nigeria's economic future. I commend CISON's leadership and everyone who has contributed to establishing this vital Institute, which I am confident will serve as a cornerstone for advancing statistical science in Nigeria.

I can attest to the critical role statisticians play in shaping and informing policies that drive sustainable development and economic stability. The theme of today's conference is both timely and appropriate. As one of the largest economies in Africa and a cornerstone of the West African Monetary Zone (WAMZ), Nigeria faces an array of complex challenges, including persistent inflation, high unemployment, fiscal pressures, and price instability driven by global oil markets. Addressing these issues requires moving beyond traditional economic models to embrace innovative, data-driven approaches that leverage big data, machine learning, and artificial intelligence.

Mr. President, we are all aware of the commendable efforts your administration has made to advance Nigeria's approach to statistical data, information technology, and innovation. Notable among these efforts are:

a. **Digital Transformation Initiatives**: which is intended to empower young Nigerians and prepare them for the AI-driven global economy.

b. **Public Service Digitization**: With a broader plan to modernize Nigeria's civil service.

c. **Infrastructure Enhancements**: This includes expansion in internet bandwidth to enhance digital connectivity and support data-intensive sectors. Others include business

and government analytics, thereby promoting economic growth through improved data accessibility.

d. **Public-Private Partnerships**: The administration has shown a commitment to collaborating with tech giants, such as Meta, to encourage innovation and support Nigeria's digital economy.

Mr. President, the strides your administration has made in revolutionizing Nigeria's approach to data, information technology, and innovation are both forward-thinking and transformative. These initiatives are paving the way for a future in which the country thrives on a foundation of reliable data, cutting-edge analytics, and sustainable economic growth. I urge all stakeholders present today to lend their full support to these efforts, so that together, we can build a digitally advanced Nigeria, capable of harnessing its potential and driving dynamic progress.

WAMI's Commitment to Statistical Innovation

At WAMI, our mission is to foster regional economic integration and monetary cooperation within the WAMZ. This is achieved through rigorous economic surveillance, financial stability, and macroeconomic convergence, all underpinned by high-quality data and strong statistical frameworks. WAMI's efforts in statistical excellence are evident in publications such as the Annual Report and Statement of Accounts, Convergence Report, and Financial Stability Report, alongside initiatives like the Unique Bank Identification/ Digital Interoperability, Regional Digital Regulatory Sandbox Projects, which enhance data transparency and financial stability.

WAMI has made significant progress in harmonizing statistical data across the region by aligning methodologies, definitions, and reporting standards. This harmonization has created a cohesive dataset for better cross-country comparisons, aiding policymakers in aligning national policies with broader WAMZ goals. These efforts continue to promote data-driven policy development and contribute to regional economic growth and integration.

Nigeria's Economic Challenges and the Need for Statistical Innovation

Nigeria's economy is resilient, but it is also faced with numerous challenges. Inflation, unemployment, and exchange rate volatility are compounded by structural issues like income inequality, infrastructure gaps, and a dependency on oil revenues. Addressing these complex issues demands innovative approaches in data collection, analysis, and interpretation, integrating technology and big data analytics to deepen our understanding of the economic landscape.

Mainstreaming new statistical tools can transform our perspective on the Nigerian economy. Technologies such as machine learning and AI offer robust predictive capabilities, enabling us to analyze vast datasets efficiently. Data science, meanwhile, supports the discovery of otherwise unnoticed patterns, helping policymakers take proactive steps to address potential economic disruptions.

Innovative Approaches to Statistical Analysis: Opportunities and Implications

Emerging technologies like artificial intelligence, satellite imagery, and geospatial analysis are now invaluable tools for economic diagnostics. Satellite data, for example, can provide real-time insights into agricultural productivity, energy consumption, and urbanization, allowing policymakers to respond rapidly to changing circumstances. Similarly, mobile technology offers ways to gather real-time data on consumer behavior and labour trends, providing insights into previously opaque sectors of the economy.

I am encouraged by CISON's commitment to embracing these innovations. As members of this distinguished Institute work to incorporate these methods, I urge you to prioritize capacity-building initiatives that equip Nigerian statisticians with cutting-edge skills in data science, machine learning, and econometrics. Such investments yield benefits that extend far beyond economic policy, impacting public health, environmental planning, and social development—all of which depend on high-quality data for impactful solutions.

The Role of CISON in Nigeria's Statistical Development

The establishment of CISON marks a significant milestone in Nigeria's statistical journey. This Institute stands as a beacon of statistical rigour and integrity, setting standards for data collection and analysis across multiple sectors. By promoting research, facilitating professional development, and advocating for policies that enhance statistical literacy, CISON is poised to play an instrumental role in shaping the future of data and analytics in Nigeria.

I encourage the Institute to forge strong partnerships with academia, government agencies, and international organizations, building a network that fosters knowledge exchange and research dissemination. CISON is also well-positioned to serve as an advisory body to the Nigerian government, offering insights on incorporating innovative statistical approaches into policy formulation and economic management.

A Call to Action: Harnessing Data Innovation for Nigeria's Economic Challenges

Let us use this conference as a platform to explore how statisticians and data innovators can contribute solutions to Nigeria's economic challenges. The focus should be on how advanced tools—particularly big data and AI—can be deployed to provide insights, predict trends, and address Nigeria's pressing economic concerns. I invite you all to propose strategies that will make these technologies accessible and impactful within Nigeria's policy formulation processes.

Four specific areas where your expertise can drive significant impact include:

a) **Big Data and Predictive Analytics**: The wealth of data generated every day presents a unique opportunity to analyze vast datasets in real-time. Predictive analytics can forecast economic trends, anticipate market shifts, and provide models that allow policymakers to respond more effectively to economic disruptions, particularly those caused by volatile global oil prices.

b) **Geospatial Analysis to Address Regional Disparities**: Nigeria's economic landscape is marked by regional disparities. Geospatial analysis offers a powerful tool for visualizing these differences, supporting decisions aimed at balanced regional

development. Mapping infrastructure, economic activity, and resource distribution helps identify development priorities and address regional inequalities effectively.

c) **Machine Learning and AI for Economic Insights**: Machine learning and AI unlock new possibilities for economic analysis, revealing patterns that traditional methods may overlook. These technologies allow us to measure and track informal economic activities, which form a substantial part of Nigeria's economy, ultimately paving the way for policies that reflect the economy's true dynamics.

d) Data security and Data Protection: Data security is essential as Nigeria embraces big data, AI, and machine learning to address economic challenges, ensuring that economic data remains protected from breaches and manipulation. CISON can play a critical role in promoting secure data practices, such as encryption, access controls, and ethical standards, tailored to the country's needs. By implementing these practices and educating statisticians on emerging risks, CISON can safeguard sensitive data and help ensure that statistical innovations contribute positively to Nigeria's economic stability and resilience.

Artificial Intelligence and the Job of the Statistician

Ladies and Gentlemen, one area that is revolutionizing the field of statistics is Artificial Intelligence (AI). AI provides statisticians with powerful tools that simplify data collection, model building, and analysis. And I urge all of you to begin the process of adopting and integrating this new normal into our job processes. Through AI, statisticians can quickly access global data repositories like the IMF and World Bank databases with just a prompt, facilitating the integration of extensive datasets. These tools also enable the automatic generation of complex statistical code for software like Python, R, EViews, and MATLAB, making statistical and econometric analyses more efficient.

Machine Learning (ML), a branch of AI, significantly enhances statistical modeling and forecasting by enabling the creation of predictive models that can analyze extensive datasets and improve accuracy in forecasting economic indicators such as inflation, GDP, and exchange rates. However, using ML effectively requires a solid grounding in statistical principles to avoid inaccurate or misleading outputs. CISON's role is vital in ensuring that Nigerian statisticians are skilled not only in using AI tools but also in applying fundamental statistical methods, so they can use ML models responsibly and generate reliable, meaningful results.

Moreover, AI's integration into statistical work raises new challenges related to data security and ethics. As statisticians increasingly work with sensitive economic data, ensuring that AI systems adhere to strict data protection and privacy standards is essential. Additionally, AI systems must be transparent and explainable, ensuring that their results can be trusted. CISON can play a pivotal role in promoting awareness of these ethical issues by incorporating data security and ethical AI use into its professional development programs.

AI presents vast opportunities for statistical innovation, but it also requires careful handling to maintain the integrity of statistical work. CISON can drive Nigeria's progress in this area by providing its members with the necessary training and resources to effectively use AI while adhering to high standards of statistical rigour and ethics.

Supporting Government Fiscal Operations: The Role of CISON

Mr. President, distinguished guests, statisticians play a critical role in supporting government fiscal operations through enhanced data collection and analysis. Challenges in revenue collection often arise from outdated or incomplete data, creating opportunities for statisticians to make a substantial impact. CISON is well-positioned to establish frameworks that improve data accuracy, consistency, and transparency, enabling revenue authorities to better identify tax bases, monitor economic activities, and uncover new revenue sources—ultimately contributing to the nation's economic growth.

Furthermore, as CISON and its members collaborate closely with government agencies, we can foster more effective data-sharing processes that allow for real-time insights into government spending and resource allocation. With robust and timely data, we empower policymakers to allocate resources more strategically, ensuring that funds reach areas of greatest need.

To achieve CISON's mission, we must support capacity-building initiatives that equip statisticians across Nigeria with advanced skills in big data, predictive analytics, and fiscal econometrics. These skills will enable us to generate more accurate revenue and expenditure forecasts, ultimately strengthening Nigeria's fiscal resilience.

Overcoming Challenges: Capacity Building and Data Quality

While pursuing these advances, we must also address structural challenges within Nigeria's statistical system. *Outdated infrastructure, limited data collection methods, and resource constraints hinder statistical effectiveness. Strengthening Nigeria's statistical framework will require investments in training, enhanced data governance, and collaboration between institutions like WAMI, CISON, and government policymakers.*

Building the capacity of our statisticians and modernizing our statistical infrastructure are essential steps toward fostering a culture of evidence-based policymaking.

Key questions for CISON

Distinguished participants, as we gather to ponder on the theme of this conference, I urge CISON to seek answers to the following **questions:**

i. How can CISON support Nigerian statisticians in adopting big data analytics and machine learning tools to enhance economic forecasting, especially in areas like inflation and employment to guide policy formulation?

ii. How can CISON contribute to creating a standardized approach to data transparency and reliability within Nigeria?

iii. How can real-time data collection, perhaps through mobile or satellite technology, be better leveraged to support agile and responsive policymaking?

iv. What role should CISON play in establishing guidelines for data security and ethical data use? How can the Institute help members navigate the balance between data accessibility for innovation and the need to protect sensitive information?

Conclusion: Building Nigeria's Economic Future with Data

The path to addressing Nigeria's economic challenges is complex but filled with opportunity. By integrating innovative statistical methods, particularly big data and AI, into economic analysis and policy design, we can gain deeper insights, develop more effective policies, and foster sustainable economic growth.

I urge all participants to utilize this conference as a forum for collaboration and networking, a place where ideas can flourish, and a platform to advance Nigeria's data-driven future. Nigeria's statisticians hold the potential to transform the economic landscape, supporting policies that are both equitable and impactful.

Together, let us build Nigeria's economic future with data, innovation, and an unwavering commitment to excellence.

Thank you, Mr. President, and everyone here present. May God bless the Chartered Institute of Statisticians of Nigeria and our beloved country, Nigeria.

Technical Papers Presented at the Conference

Stochastic Modeling of COVID-19: A SEI_AI_SQR-type model

F. Ewere[1,*], J. I. Mbegbu[2] and C. U. Chikwelu[3].

[1,2]Department of Statistics, Faculty of Physical Sciences, University of Benin, Benin City, Nigeria
[3]Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria

*Corresponding author Email: friday.ewere@uniben.edu

Abstract

This paper presents a model for the Corona-Virus (COVID-19) disease taking into account random perturbations. The proposed model is composed of six different classes namely the Susceptible population, the Exposed population, the Asymptomatic infectious population, the Symptomatic Infectious population, the Quarantined population and the Recovered population (SEI_AI_SQR). Using appropriately formulated stochastic Lyapunov functions, we established sufficient conditions for the extinction of the disease. Numerical simulations are applied to illustrate the analytical results obtained herein. The reproduction number was obtained as $R_0^S < 1$ and $R_0^S > 1$ which show that the stability analysis of the equilibrium point is locally asymptotically stable whenever the basic reproduction number $R_0^S < 1$ and unstable whenever $R_0^S > 1$

Keywords: Stochastic Model, COVID-19, SEI_AI_SQR Model, Lyapunov Function, Milstein's Higher Order Method

1. Introduction

The 2019 novel coronavirus has been known to the virologist's community as Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) (Lai, 2020). Coronaviruses are a family of many diverse and numerous viruses that can infect both humans and animals that can cause a number of diseases (WHO, 2020). The name coronavirus, which means "crown virus" is related to the fact that all viruses of this family have a crown-like shape when observed under an electron microscope. The epidemic of novel coronavirus (COVID-19) infections that began in China in late 2019 has rapidly grown and cases have been reported worldwide. The virus appears to be transferred mostly through narrow respiratory droplets by coughing, sneezing, or people's interaction in close proximity (usually less than one meter) with each other for a

certain time frame. These droplets can further be inhaled or can stay on the surfaces that came in contact with the infected person that can now cause infection in others by touching their nose, mouth or eyes. The virus possesses ability to survive on various surfaces commencing several hours (e.g. copper, cardboard), up to a few days (e.g. plastic and stainless steel). Nonetheless, the quantity of the viable virus certainly decays over a time span and might not be present in sufficient quantity for causing the infection (Din *et al* 2020).

Most of the real world phenomenon are not simply deterministic, because in deterministic models, the output of the model is fully determined by the parameter values and the initial conditions. Stochastic models possess some inherent randomness. The same set of parameter values and initial conditions will lead to an ensemble of different outputs or we can say a deterministic model is one that uses numbers as inputs, and produces numbers as outputs. A stochastic model includes a random component that uses a distribution as one of the inputs, and results in a distribution for the output. These distributions may reflect the uncertainty in what the input should be (e.g. a deterministic input plus noise), or may reflect a random process (i.e. a stochastic input) (Perko, 2013; Wu and Mcgoogan, 2020).

Statement of the Problem

Ding *et al* (2021) proposed a Levy noise-driven susceptible-exposed-infected-recovered model incorporating media coverage to study the outbreak of the COVID-19. The model consists of four the susceptible individuals, the second one pertains to the exposed individual, the third is the infected population and the fourth is the recovered individuals. However, the model by Ding *et al*. (2021) has a short coming, in that, it does not take congnisance of the population quarantined and symptomatic infected class. Therefore, in view of the limitation this cannot be used to capture data that has the quarantined and the symptomatic infected class. Hence, this work intends to fill the gap by including the Quarantine and symptomatic infected class. In other words, we shall attempt to modify the existing susceptible-exposed-infected-recovered model by in-cooperating the quarantine and symptomatic infected class.

Objective of the Study are:

1. To develop a stochastic epidemic Susceptible-Exposed-Infected-Quarantine-Recovered (SEIQR) model.
2. To estimate the parameters using stochastic approach.
3. To deduce a stochastic basic reproduction number for the extinction of the disease.

2. Literature Review

Din et al. (2020) studied the dynamics of Covid-19 virus, they proposed a susceptible (S), infected(I), and quarantine(Q) model. Stochastic Lyapunov functions theory was used to derive the existence and the reproduction number for the model, but the model has a shortcoming in that it cannot be used to model disease with the exposed class and asymptomatic infectious class.

Ding et al. (2021) proposed a stochastic Lévy noise-driven Susceptible-Exposed-Infected-Recovered ($SEIR$) model incorporating media coverage to analyze the outbreak of Covid-19. The model did not incorporate the quarantine class thereby cannot be used to model the class of quarantine and asymptomatic class.

Gonzalez et al. (2022) proposed two different mathematical models to study the effect of immigration on the COVID-19 pandemic. The mathematical models considered five different subpopulations: susceptible, exposed, infected, asymptomatic carriers, and recovered (SEIAR). It is important to know that this work did not consider the quarantine class.

Olabode et al. (2021) proposed a deterministic susceptible-exposed-infected-recovered (SEIR) model to study the transmission dynamics of the Coronavirus Disease 2019 (COVID-19)

The result indicates that the basic reproduction number R_0 serves as a sharp disease threshold it is important to know that this approach did not extend to the quarantine approach.

Yavuz et al. (2021) studied the effect of vaccine treatment on covid-19. They proposed a susceptible, infected, exposed, and recovered population. It is important to know that this work did not include the quarantine class.

3. Methodology

In this section, we formulate a stochastic model to study the transmissions dynamics of COVID-19. According to the characteristics of the disease, we propose a Susceptible-Exposed-Asymptomatic Infectious-Symptomatic Infectious-Quarantined-Removed epidemic model. We take into consideration the variations of the population environment in order to study the dynamics of COVID-19.

Some of the assumptions underlying the formulation of the model are:

1. The total population at any time *t* is denoted by $N(t)$ and it is classified into six exclusive groups of individuals: the susceptible class $S(t)$, the Exposed class $E(t)$, the Asymptomatic infectious class $I_A(t)$, the Symptomatic infectious class $I_S(t)$, the

Quarantine class $Q(t)$, and the Recovered $R(t)$. That is, $S(t) + E(t) + I_A(t) + I_S(t) + Q(t) + R(t) = N(t)$ which is changing with time t.

2. The state variables and parameters included in the model are assumed to be nonnegative.
3. The infected individuals move to the quarantined class
4. Once the infection is confirmed, then the quarantined will go back to the infected compartment.

In the light of the assumptions, we obtain the deterministic model:

$$\frac{dS}{dt} = \pi - \rho_1\beta_A I_A S - \rho_2\beta_S I_S S - \mu S$$

$$\frac{dE}{dt} = \rho_1\beta_A I_A S + \rho_2\beta_S I_S S - (\alpha + \mu)E$$

$$\frac{dI_A}{dt} = (1-\delta)\alpha E - (\gamma + \mu_\chi + \mu)\mathrm{I_A} \qquad (1)$$

$$\frac{dI_S}{dt} = \alpha\delta E - (\omega + \mu_\chi + \mu)\mathrm{I_S}$$

$$\frac{dQ}{dt} = \gamma I_A + \omega I_S - (\theta + \mu_\chi + \mu)\mathrm{Q}$$

$$\frac{dR}{dt} = \theta\mathrm{Q} - \mu\mathrm{R}$$

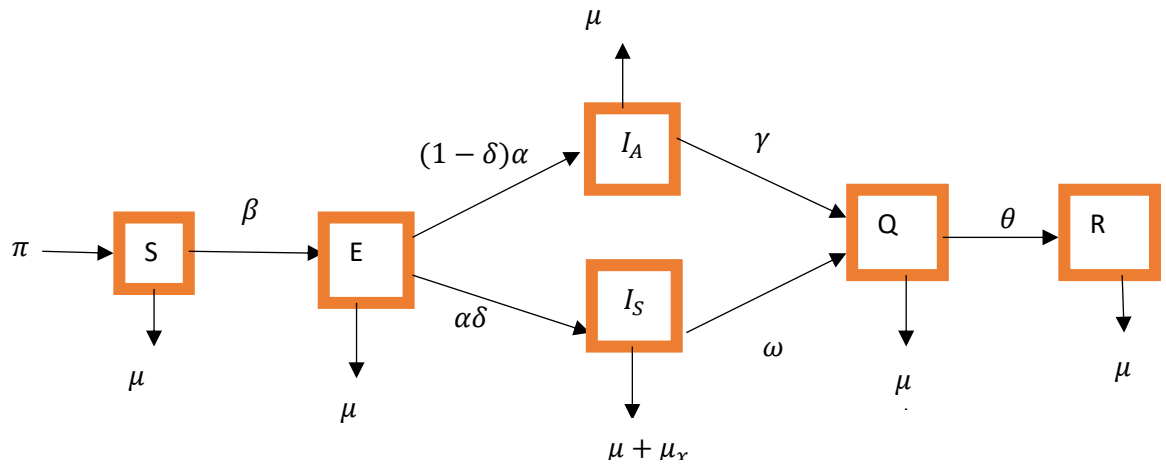


Fig 1: *The schematic diagram of the SEI$_A$I$_S$QR Model*

where π be the recruitment rate into the susceptible class and μ, is the natural death rate, β is the force of infection, ρ_1 *and* ρ_2 are the effective contact rate between the symptomatic and asymptomatic individuals respectively, β_A *and* β_S are the transmission probabilities of the virus from infectious individuals to the susceptible class, μ_χ is death due to covid-19, $(1-\delta)\alpha$ the proportion of individuals moving into the Asymptomatic class, $\alpha\delta$ the rate at which the

individuals move into the symptomatic class, γ the rate at which the Asymptomatic class move into the quarantine class, ω is the rate at which the Symptomatic class are quarantined and θ is the recovery rate.

From the deterministic model in (1), we then formulate the stochastic mode

$$S(t) = (\pi - \rho_1\beta_A I_A(t)S(t) - \rho_2\beta_S I_S(t)S(t) - \mu S(t))dt + \sigma_1 S(t)dW_1(t)$$

$$E(t) = (\rho_1\beta_A I_A(t)S(t) + \rho_2\beta_S I_S(t)S(t) - (\alpha + \mu)E(t))dt + \sigma_2 E(t)dW_2(t)$$

$$\mathrm{I_A}(t) = ((1-\delta)\alpha E(t) - (\gamma + \mu_\chi + \mu)\mathrm{I_A}(t))dt + \sigma_3 \mathrm{I_A}(t)dW_3(t)$$

$$\mathrm{I_S}(t) = (\alpha\delta E - (\omega + \mu_\chi + \mu)\mathrm{I_S})dt + \sigma_4 \mathrm{I_S}(t)dW_4(t) \qquad (2)$$

$$dQ(t) = (\gamma I_A(t) + \omega I_S(t) - (\theta + \mu_\chi + \mu)\mathrm{Q(t)})dt + \sigma_5 Q(t)dW_5(t)$$

$$dR(t) = (\theta\mathrm{Q(t)} - \mu\mathrm{R(t)})dt + \sigma_6 R(t)dW_6(t)$$

where $W_1(t)$, $W_2(t)$, $W_3(t)$, $W_4(t)$, $W_5(t)$, $W_6(t)$ are independent standard Brownian motions, and σ_1, σ_2, σ_3, σ_4, σ_5, σ_6 as the intensities of the standard Gaussian white noises, respectively.

The terms $\sigma_1 S(t)dW_1(t), \sigma_2 E(t)dW_2(t), \sigma_3 \mathrm{I_A}(t)dW_3(t), \sigma_4 \mathrm{I_S}(t)dW_4(t), \sigma_5 Q(t)dW_5(t)$, and $\sigma_6 R(t)dW_6(t)$ are the interactions between the individuals and the environment. The initial conditions are given by ($S(0)$,$E(0)$, $I_A(0)$, $I_S(0)$,$Q(0)$,$R(0)$) $\in \mathbb{R}^6_+$

The reproduction number $\mathfrak{R}_0^S$ for the stochastics system (2) is given by

$$\mathfrak{R}_1^S = \frac{(\rho_1\beta_A \frac{\pi}{\mu})(1-\delta)\alpha}{(\alpha+\mu)(\gamma+\mu_x+\mu)} - \frac{\delta_3^2}{2(\gamma+\mu_x+\mu)} \qquad (3)$$

$$\mathfrak{R}_2^S = \frac{(\rho_2\beta_S \frac{\pi}{\mu})\delta\alpha}{(\alpha+\mu)(\omega+\mu_x+\mu)} - \frac{\delta_4^2}{2(\omega+\mu_x+\mu)}$$

Extinction

In modeling the dynamics of any infectious disease, it is important to study conditions under which the disease will go into extinction or die out from the population. In this section, we will show that if the white noise is sufficiently large, then the solution of associated stochastic system (2) will become extinct with probability one. Let us consider the following notation and results:

$$X(t) = \frac{1}{t}\int_0^t x(\tau)d\tau \quad (4)$$

Lemma 1 (Strong law of large number, Mao (1997)) let $M = \{M_t\}, t \geq 0$ be continuous and real valued local martingale vanishing at $t = 0$ and $\langle M, M\rangle t$ be its quadratic variation. Then

$$\lim_{t\to\infty}(M,M)_t = \infty\ a.s,\ \lim_{t\to\infty}\frac{M_t}{(M,M)_t} = 0\ a.s \quad (5)$$

Also

$$\lim_{t\to\infty} sup\frac{(M,M)_t}{t} < 0\ a.s,\ \lim_{t\to\infty}\frac{M_t}{t} = 0\ a.s$$

Lemma 2

Let $\big(S(t), E(t), I_A(t), I_S(t), Q(t), R(t)\big)$ be the solution of the stochastic system (2) with initial value $\big(S(0), E(0), I_A(0), I_S(0), Q(0), R(0)\big) \in \mathbb{R}_+^6$ then

$$\lim_{t\to\infty}\frac{\big(S(t),E(t),I_A(t),I_S(t),Q(t),R(t)\big)}{t} = 0$$

Moreover,

$$\lim_{t\to+\infty}\frac{1}{t}\int_0^t S(\tau)dW_1(\tau) = 0,\ \lim_{t\to+\infty}\frac{1}{t}\int_0^t E(\tau)dW_2(\tau) = 0,\ \lim_{t\to+\infty}\frac{1}{t}\int_0^t I_A(\tau)dW_3(\tau) = 0, \quad (6)$$

$$\lim_{t\to+\infty}\frac{1}{t}\int_0^t I_S(\tau)dW_4(\tau) = 0,\ \lim_{t\to+\infty}\frac{1}{t}\int_0^t Q(\tau)dW_5(\tau) = 0,\ \lim_{t\to+\infty}\frac{1}{t}\int_0^t R(\tau)dW_6(\tau) = 0$$

The threshold quantity $\mathfrak{R}_1^S$ for the stochastic system (2) can be written as

$$\mathfrak{R}_1^S = \frac{\rho_1\beta_A\frac{\pi}{\mu}(1-\delta)\alpha}{(\alpha+\mu)(\gamma+\mu_x+\mu)} - \frac{\sigma_3^2}{2(\gamma+\mu_x+\mu)} < 1,$$

$$\mathfrak{R}_2^S = \frac{(\rho_2\beta_S\frac{\pi}{\mu})(\delta\alpha)}{(\alpha+\mu)(\omega+\mu_x+\mu)} - \frac{\sigma_4^2}{2(\omega+\mu_x+\mu)} \quad (7)$$

The following theorem gives the necessary conditions for the extinction of infections.

Theorem 1: for any given initial value $\big(S(0), E(0), I_A(0), I_S(0), Q(0), R(0)\big) \in \mathbb{R}_+^6$, the solution $\big(S(t), E(t), I_A(t), I_S(t), Q(t), R(t)\big)$ of the system in (2) has the following properties: if

a(i) $\sigma_3^2 > \frac{((\rho_1\beta_A\frac{\pi}{\mu})(1-\delta)\alpha)^2}{2(\alpha+\mu)^2(\gamma+\mu_x+\mu)}$, then Covid-19 of I_A goes into extinction a.s

(ii) $\sigma_4^2 > \frac{((\rho_2\beta_S\frac{\pi}{\mu})\delta\alpha)^2}{2(\alpha+\mu)^2(\omega+\mu_x+\mu)}$, then Covid-19 of I_S goes into extinction a.s

b(i) $\mathfrak{R}_1^S$<1, then I_A dies out with probability 1

(ii) $\Re_2^S < 1$, then I_S dies out with probability 1

This implies that , if condition (a) and (b) hold, then

$\lim_{t\to\infty} \frac{\langle logI_A\rangle}{t} < 0$ and $\lim_{t\to\infty} \frac{\langle logI_S\rangle}{t} < 0\ a.s$

That is, the disease goes into extinction with probability 1

Moreover,

$\lim_{t\to\infty}\langle S(t)\rangle = \frac{\pi}{\mu}, \lim_{t\to\infty}\langle E(t)\rangle = 0, \lim_{t\to\infty}\langle I_A(t)\rangle = 0, \lim_{t\to\infty}\langle I_S(t)\rangle = 0, \lim_{t\to\infty}\langle Q(t)\rangle =$

$0, \lim_{t\to\infty}\langle R(t)\rangle = 0$ (8)

Proof: Integrating the model in (2), we have

$$\frac{S(t)-S(0)}{t} = \pi - \rho_1\beta_A\langle I_A(t)S(t)\rangle - \rho_2\beta_S\langle I_S(t)S(t)\rangle - \mu\langle S(t)\rangle + \frac{\sigma_1}{t}\int_0^t S(\tau)dW_1(\tau)$$

$$\frac{E(t)-E(0)}{t} = \rho_1\beta_A\langle I_A(t)S(t)\rangle - \rho_2\beta_S\langle I_S(t)S(t)\rangle - (\alpha+\mu)\langle E(t)\rangle + \frac{\sigma_2}{t}\int_0^t E(\tau)dW_2(\tau)$$

$$\frac{I_A(t)-I_A(0)}{t} = (1-\delta)\alpha\langle E(t)\rangle - (\gamma+\mu_X+\mu)\langle I_A(t)\rangle + \frac{\sigma_3}{t}\int_0^t I_A(\tau)dW_3(\tau)$$

$$\frac{I_S(t)-I_S(0)}{t} = \delta\alpha\langle E(t)\rangle - (\omega+\mu_X+\mu)\langle I_S(t)\rangle + \frac{\sigma_4}{t}\int_0^t I_S(\tau)dW_4(\tau) \quad (9)$$

$$\frac{Q(t)-Q(0)}{t} = \gamma\langle I_A(t)\rangle + \omega\langle I_S(t)\rangle - (\theta+\mu_X+\mu)\langle Q(t)\rangle + \frac{\sigma_5}{t}\int_0^t Q(\tau)dW_5(\tau)$$

$$\frac{R(t)-R(0)}{t} = \theta\langle Q(t)\rangle - \mu\langle R(t)\rangle + \frac{\sigma_6}{t}\int_0^t R(\tau)dW_6(\tau)$$

(a) if Ito's formula is applied to the third equation of (2), we have

$$dlogI_A(t) = \left[(1-\delta)\alpha E(t) - (\gamma+\mu_x+\mu)I_{A(t)}\right]\frac{1}{I_A}dt - \frac{\sigma_3^2}{2}dt + \sigma_3 dW_3(t) \quad (10)$$

If we integrate the third equation of (9) within $[0,t]$, then we have

$$logI_A(t) = \int_0^t (1-\delta)\alpha E(\tau)d\tau - (\gamma+\mu_x+\mu)t\int_0^t \frac{\sigma_3^2}{2}dt + \frac{\sigma_3}{t}\int_0^t dW_3(\tau) + logI_A(0)$$

$$\leq \int_0^t (1-\delta)\alpha E(\tau)d\tau - (\gamma+\mu_x+\mu)t\int_0^t \frac{\sigma_3^2}{2}dt + \frac{\sigma_3}{t}\int_0^t dW_3(\tau) + logI_A(0)$$

$$\leq \int_0^t \left(\frac{(1-\delta)\alpha\rho_1\beta_A\left(\frac{\pi}{\mu}\right)}{\alpha+\mu} - \frac{\sigma_3^2}{2}\right)d\tau - (\gamma+\mu_x+\mu)t + \int_0^t \sigma_3 dW_3(\tau) + logI_A(0)$$

Which can be re-written as

$$logI_A \leq -\frac{\sigma_3^2}{2}\int_0^t \left(1-\frac{(1-\delta)\alpha\rho_1\beta_A\left(\frac{\pi}{\mu}\right)}{(\alpha+\mu)\sigma_3^2}\right)^2 d(\tau) - (\gamma+\mu_x+\mu)t$$

$$+\int_0^t \left(1-\frac{(1-\delta)\alpha\rho_1\beta_A\left(\frac{\pi}{\mu}\right)}{(\alpha+\mu)\sigma_3^2}\right)^2 d(\tau) + \int_0^t \sigma_3 dW_3(\tau) + logI_A(0)$$

$$\leq -(\gamma+\mu_x+\mu) - \frac{\left((1-\delta)\alpha\right)^2\left(\rho_1+\beta_A\left(\frac{\pi}{\mu}\right)\right)^2}{2\sigma_3^2(\alpha+\mu)^2}t + \int_0^t \sigma_3 dW_3(\tau) + logI_A(0)$$

Division by t gives

$$\frac{logI_A}{t} \leq -(\gamma+\mu_x+\mu) - \frac{\left((1-\delta)\alpha\right)^2\left(\rho_1+\beta_A\left(\frac{\pi}{\mu}\right)\right)^2}{2\sigma_3^2(\alpha+\mu)^2} + \frac{1}{t}\int_0^t \sigma_3 dW_3(\tau) + \frac{logI_A(0)}{t} \qquad (11)$$

By strong law of large number,

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \sigma_3 dW_3(\tau) = 0\ a.s \text{ as } \sigma_3^2 > \frac{\left((1-\delta)\alpha\right)^2\left(\rho_1+\beta_A\left(\frac{\pi}{\mu}\right)\right)^2}{2(\alpha+\mu)^2(\gamma+\mu_x+\mu)},$$

By taking limit superior on both sides of (11), we obtain,

$$\lim_{t\to\infty} sup\frac{logI_A}{t} \leq -\left((\gamma+\mu_x+\mu) + \frac{\left((1-\delta)\alpha\right)^2\left(\rho_1+\beta_A\left(\frac{\pi}{\mu}\right)\right)^2}{2\sigma_3^2(\alpha+\mu)^2}\right) < 0$$

This implies that, $\lim_{t\to\infty} I_A(0) = 0$

In a similar manner, it can be shown that

$$\lim_{t\to\infty} sup\frac{logI_S}{t} \leq -\left((\omega+\mu_x+\mu) + \frac{\left(\delta\alpha(\rho_1+\beta_A\left(\frac{\pi}{\mu}\right)\right)^2}{2\sigma_4^2(\alpha+\mu)^2}\right) < 0$$

This implies that, $\lim_{t\to\infty} I_S(0) = 0$

(b) Again if equation (10) is integrated within $[0,t]$ and divided by t, then we have

$$\frac{logI_A(t) - logI_A(0)}{t} = (1-\delta)\alpha\langle E(t)\rangle - (\gamma+\mu_X+\mu)\frac{\sigma_3^2}{2}\frac{\sigma_3}{t}\int_0^t dW_3(\tau),$$

$$\leq \frac{(1-\delta)\alpha\rho_1\beta_A\left(\frac{\pi}{\mu}\right)}{\alpha+\mu} - (\gamma+\mu_x+\mu) - \frac{\sigma_3^2}{2}\frac{\sigma_3}{t}\int_0^t dW_3(\tau),$$

$$= (\gamma+\mu_x+\mu)\left(\frac{(1-\delta)\alpha\rho_1\beta_A\left(\frac{\pi}{\mu}\right)}{\alpha+\mu}\right) - \left(\gamma+\mu_x+\mu-\frac{\sigma_3^2}{2(\gamma+\mu_x+\mu)}-1\right) + \frac{\sigma_3}{t}\int_0^t dW_3(\tau),$$

$$= (\gamma+\mu_x+\mu)(\Re_1^S-1) + \frac{\sigma_3}{t}\int_0^t dW_3(\tau), \quad (12)$$

Moreover, $M(t) = \frac{\sigma_3}{t}\int_0^t dW_3(\tau)$ is continuous (locally) and $M(0) = 0$ from equation (2) and $t \to \infty$

$$\lim_{t\to\infty} sup\frac{M(t)}{t} = 0 \quad (13)$$

If $\Re_1^S < 1$ then equation (12) becomes

$$\lim_{t\to\infty} sup\frac{logI_A(t)}{t} \leq (\gamma+\mu_x+\mu)(\Re_1^S-1) < 0\ a.s \quad (14)$$

Equation (14) implies

$$\lim_{t\to\infty} I_A(t) = 0\ a.s \quad (15)$$

Likewise, if Ito's formula is applied to the fourth equation of (2), then we obtain

$$\frac{logI_S(t)-logI_S(0)}{t} \leq (\omega+\mu_x+\mu)\left(\frac{\alpha\delta(\rho_1\beta_A\left(\frac{\pi}{\mu}\right))}{(\omega+\mu_x+\mu)(\alpha+\mu)} - \frac{\sigma_4^2}{2(\omega+\mu_x+\mu)} - 1\right) + \frac{\sigma_4}{t}\int_0^t dW_4(\tau)$$

$$= (\omega+\mu_x+\mu)(\Re_2^S-1)\frac{\sigma_4}{t}\int_0^t dW_4(\tau) \quad (16)$$

Note that $M(t) = \frac{\sigma_4}{t}\int_0^t dW_4(\tau)$, is also locally "continuous martingale" and by lemma (2) and $t \to \infty$, we have $\lim_{t\to\infty} sup\frac{M(t)}{t} = 0$ (17)

$$\lim_{t\to\infty} sup\frac{logI_S(t)}{t} \leq (\omega+\mu_x+\mu)(\Re_2^S-1) < 0\ a.s \quad (18)$$

$$\lim_{t\to\infty} I_S(t) = 0\ a.s \quad (19)$$

4. Result and Discussion (Numerical Simulation)

Using the Milstein's Higher Order Method (Higham, 2001), we conducted numerical simulations for model (2) to support analytical findings discussed earlier. The initial conditions

for the simulations are set as flows: S (0) = 0.6, E (0) = 0.2, $I_A(0)$=0.1, $I_S(0)$=0.1, Q (0) = 0.5 and R (0) = 0.1 over the time range from 0 to 10. To verify the extinction result numerically, we selected the following μ=0.01, π =0.9, β_S=0.02, β_A=0.02, δ =0.05, α=0.5, μ_χ =0.2, γ =0.3, ρ_1=0.8, ρ_2=0.7, ω=0.4, θ =0.1. The noise intensities used were $\sigma_1 = 0.2$, $\sigma_2 = 0.5$, $\sigma_3 = 0.35$, $\sigma_4 = 0.1$, $\sigma_5 = 0.2$ and $\sigma_6 = 0.4$ which gives $\mathcal{R}_0^S = 0.6019$. The long-term dynamics of the model are illustrated in Figures 1 through 6, showing the evolution of the susceptible, exposed, asymptomatic infected, symptomatic infected, quarantined, and recovered populations in the stochastic model. The simulations indicate that the disease eventually dies out, with the infection rate of the novel coronavirus diminishing exponentially as the intensity of the white noise increases. Nevertheless, a portion of the susceptible population remains even as the disease is eradicated.

Table 1: Parameter values

Notation	Values	Reference
S	0.6	Ding *et al* (2021)
E	0.2	Ding *et al*
I_A	0.1	Ding *et al*
I_S	0.1	Ding *et al*
Q	0.5	Din *et al* (2020)
R	0.1	Ding *et al*
π	0.9	assumed
δ	0.05	Ding *et al*
α	0.5	Din *et al*
μ_x	0.2	Din *et al*
μ	0.01	Ding *et al*
γ	0.3	Din *et al*
ω	0.4	Assumed
θ	0.1	Din *et al*
β_S	0.02	Assumed
β_A	0.02	Assumed
ρ_1	0.8	assumed
ρ_2	0.7	assumed

σ_1	0.2	Ding *et al*
σ_2	0.5	Ding *et al*
σ_3	0.35	Ding *et al*
σ_4	0.1	Ding *et al*
σ_5	0.2	Din *et al*
σ_6	0.4	Din *et al*

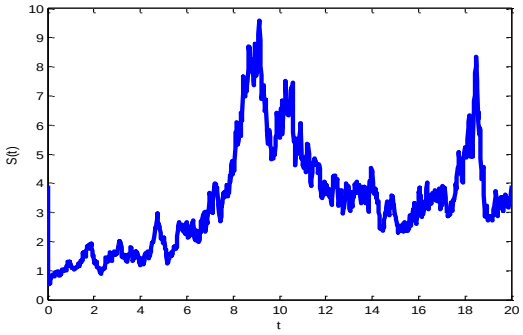


Figure 1: Susceptible population

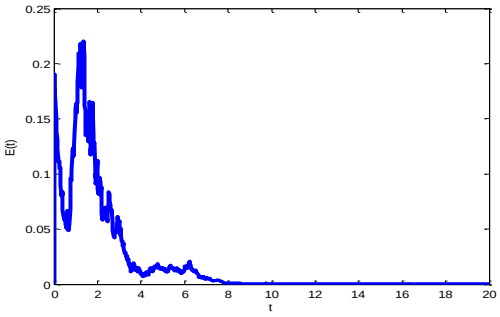


Figure 2: Exposed population

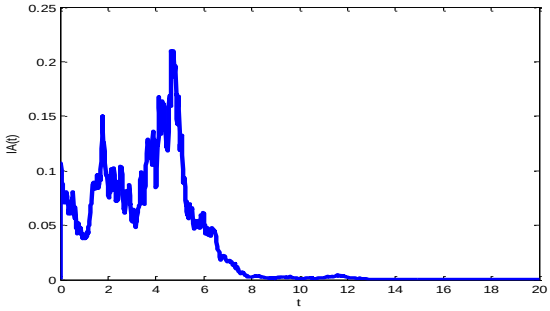


Figure 3: Asymptomatic Infectious

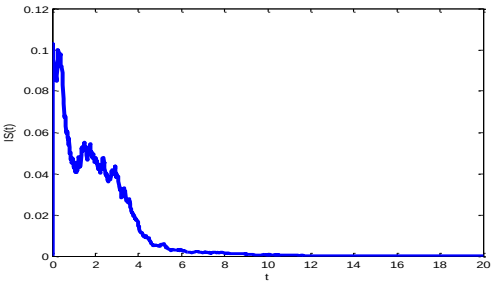


Figure 4: Symptomatic Infectious

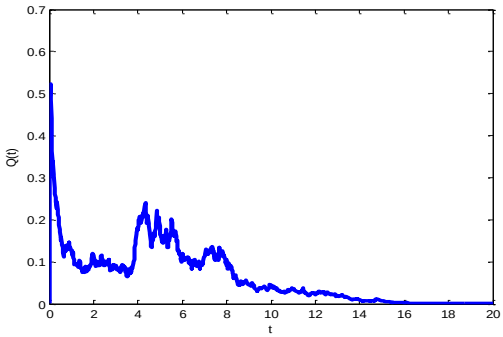


Figure 5: Quarantine population

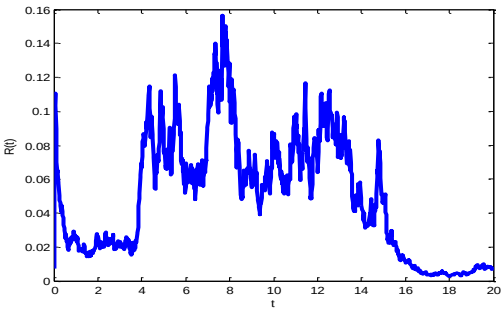


Figure 6: Recovered population

Discussion

Figure 1-6 is the graphical result showing the Extinction of the COVID-19 epidemic

Figure 1, shows the evolution of the susceptible population. The fluctuations suggest that the population's susceptibility varies over time due to interactions with the infected populations and the introduction of random noise (stochastic effects). The spikes correspond to periods when the infection rates drop, causing more people to remain susceptible, followed by declines as the disease spreads again and depletes the susceptible pool. The continued oscillations suggest that while the disease hasn't completely died out in this group, there is some fluctuation due to stochastic factors.

In figure 2, the rapid initial rise followed by a steep drop to near-zero levels suggests that the exposed population (those in the incubation phase) quickly grows as the disease starts to spread, but then sharply declines as the infection progresses or containment measures take effect. The near-zero levels later in the graph indicate that the exposed population is almost entirely diminished, which is due to successful quarantining or the progression of exposed individuals into symptomatic or recovered categories.

Figure 3, shows the dynamics of asymptomatic individuals who are infected but not showing symptoms. Like the exposed population, there is a rise at the onset, indicating rapid infection among this group. The gradual decline to zero suggests that asymptomatic infections eventually decrease, due to recovery and quarantining and the trend toward zero aligns with the overall extinction scenario, where the disease is progressively wiped out over time.

In figure 4, the initial rise is due to disease spread, the eventual decline reflects the diminishing spread of the infection, suggesting that the disease is being brought under control but control measures (like quarantine) and natural recovery lead to a decrease over time, with some variability due to the stochastic model.

Figure 5, shows as that quarantined population also rises with infection detection, fluctuates as more individuals are isolated, the decrease after the initial rise suggests that fewer new individuals are being quarantined, likely because the overall infection rate is declining. Once the symptomatic population decreases, there are fewer individuals left to isolate, leading to a drop in the quarantined population.

Figure 6, recovered population shows more variability, with cycles of increase and decrease, reflecting the complex interactions between infection, recovery, and the stochastic nature of the disease spread. The overall trend, however, suggests an eventual increase in the recovered population as more individuals overcome the disease.

5. Conclusion

The novel coronavirus, also known as COVID-19, is one of the most severe and dangerous viruses that has impacted the world, with significant consequences for both the global economy and social structures in nearly every country. Since it involves real-world complexities, deterministic models are insufficient to describe its behavior, particularly in the presence of randomness. Therefore, stochastic models are a more suitable approach for accurately representing such situations. In this study, we proposed a stochastic model for COVID-19, considering its unique characteristics to analyze the transmission dynamics within a changing population. We explored the conditions under which the virus might become extinct, focusing on the role of noise intensity in influencing transmission. Our findings show that as the intensity of noise increases, the likelihood of COVID-19 extinction also rises. The theoretical results were validated numerically using Milstein's Higher-Order stochastic method. Overall, stochastic analysis proves to be a more effective approach for studying the spread of infectious diseases like COVID-19, as it accounts for the numerous variables that change over time and across different locations.

References

Din, A., Khan, A., & Baleanu, D. (2020). Stationary distribution and extinction of stochastic coronavirus (COVID-19) epidemic model. *Chaos, Solitons & Fractals, 139*, 1–10.

Ding, Y., Fu, Y., & Kang, Y. (2021). Stochastic analysis of COVID-19 by an SEIR model with Lévy noise. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 31*(4), 1–19.

González-Parra, G., Díaz-Rodríguez, M., & Arenas, A. J. (2022). Mathematical modeling to study the impact of immigration on the dynamics of the COVID-19 pandemic: A case study for Venezuela. *Spatial and Spatio-temporal Epidemiology, 43*, 1–12.

Higham, D. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, *43*(3), 525–546.

Olabode, D., Culp, J., Fisher, A., Tower, A., Hull-Nye, D., & Wang, X. (2021). Deterministic and stochastic models for the epidemic dynamics of COVID-19 in Wuhan, China. *Mathematical Biosciences and Engineering*, *18*(1), 950–967.

Perko, L. (2013). *Differential equations and dynamical systems* (3rd ed.). Springer-Verlag.

Yavuz, M., Coşar, F. O., Günay, F., & Özdemir, F. N. (2021). A new mathematical modeling of the COVID-19 pandemic including the vaccination campaign. *Open Journal of Modelling and Simulation, 9*, 299–321.

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72,314 cases from the Chinese Center for Disease Control and Prevention. *Journal of the American Medical Association, 323*(13), 1–15.

Zhu, Q. (2018). Stability analysis of stochastic delay differential equations with Lévy noise. *Systems & Control Letters, 118*, 62–68.

Zhu, Q. (2017). Razumikhin-type theorem for stochastic functional differential equations with Lévy noise and Markov switching. *International Journal of Control, 90*(8), 1703–1712.

Stochastic nonlinear optimal allocations with Weibull cost function when non-response is observed

Ewere, F., Mbegbu, J. I. and Okon, E. J.

Department of Statistics, University of Benin, Edo State, Nigeria

*Corresponding author email: julian.mbegbu@uniben.edu

Abstract

Different cost functions, comprising of linear, quadratic and gamma cost function have been utilized based on their functional form and will be obsolete if additional cost is incurred. In this paper, a Weibull cost function when there is non-response is proposed for multivariate stratified sampling surveys. Hansen and Hurwitz (1946) introduced the concept of non-response by subsampling non-respondents in a more careful second attempt and the variance function was considered to be deterministic. Furthermore, in real life situations, population parameters are considered to be random and uncertain, and it is advisable to use a stochastic programming approach to model such problems. To optimally distribute sample sizes to strata, the problem will be modeled using the multi-objective stochastic integer nonlinear programming problem (MO-SINLPP) technique. A solution method is suggested using D_1 distance technique and fuzzy programming. The results of the comparative study showed that the compromise optimum allocations outperformed the individual optimum allocation of the different characteristics using the Weibull cost function.

Keywords: Stochastic programming, Weibull cost function, Non-response, Optimal allocation, Fuzzy programming

1. Introduction

Multivariate stratified sampling has been the core of sampling surveys in recent studies and primarily, sampling efficiency is determined by how sample sizes are distributed across strata. In surveys, obtaining information from the complete population rather than just a sample could make decision-making for researchers and policy makers more difficult. One way of improving the precision of estimates without increasing the sample size in a survey is by arranging the entire population into subpopulations called strata. In multivariate surveys, the best allocation for one character in a unit may not be convenient for another in that same unit. In these case, a compromise that is, in some way, optimal for all the criteria must be found in order to determine

how samples should be distributed throughout various strata. Optimum allocation problem was considered by Cochran (1977) and Sukhatme et al. (1984), for stratified random sampling using a univariate population. Several optimal compromise allocation methods for surveys that are multipurpose in nature have been developed, either by minimizing the fixed cost for a specified variance or by reducing the coefficient of variation for a given cost.

Stratified sampling surveys present different challenges, one of which is determining the most appropriate stratum size. When allocating sample size to strata, factors such as the cost of collecting data per element in each stratum, variability within each stratum and the population size of each stratum must be considered. Some of these problems are shrouded in different forms of uncertainty. They become stochastic or fuzzy when some of the parameters are random **(*Alshqaq et al. 2022)*.** These problems may be formulated on areas of random probability or uncertainty of outcome. This can be solved using some well-known methods like Stochastic Chanced Constrained Programming (SCCP) technique, which was originally introduced by Charnes and Cooper in (1959). The technique provides a robust approach for modeling stochastic control systems and decision-making (Ding and Wang, 2012).

In sample survey, it is expected that information would not be obtained from all the selected units in the sample, but practically, it's generally not possible due to nonresponse. Nonresponse will exist if the sampled person does not provide the required information in the survey. According to Shafiullah et al. (2014), the degree of nonresponse solely relies on factors like, the nature of the population being targeted, kind of survey considered and the period in which the survey is conducted.

A classical nonresponse concept, designed originally for surveys which first attempt involved mailing of questionnaires and conducting of interviews second attempt to the subsample of the non-respondents was introduced by Hansen and Hurwitz (1946). Additionally, the estimator of the population mean was obtained and the expression for its variance was also derived.

El-Badry (1956) modified the Hansen and Hurwitz method by sending multiple waves of questionnaires to the non-responding units, in other to improve the rate of response. Foradori (1961) further expanded El-Badry's approach to accommodate several sampling designs.

Khan et al. (2008) determined the optimal sample sizes and subsamples into strata when non-response is observed for multivariate stratified sampling survey with Lagrange Multipliers Technique and developed an estimate for stratum mean for the population characteristics that

is unbiased and also derived the variance of the Hansen and Hurwitz (1946) estimator, which accounts for the sub-sample of non-respondents. The linear cost function with nonresponse was utilized to solve the nonlinear programming problem (NLPP), with a convex objective function. To further enhance the concept of nonresponse, new estimators were developed by Ahmed et al. (2016) for Scrambled Response on a second attempt.

Optimum compromise allocations with Integer values are recommended since there are widely used for practical implementation. To obtain integer values, Varshney et al. (2011) applied multivariate stratified sampling in the context of non-response by formulating an integer nonlinear programming problem with linear cost function. A solution method was derived with lexicographic goal programming (LGP) with the aim of securing optimal compromise allocations. The procedure compared Cochran's allocation, Chatterjee's allocation, minimized trace and Khan's compromise allocation with the LGP technique. The LGP approach ensured a feasible solution, which other methods may not provide, and proved to be competitive in terms variability and cost.

Varshney et al. (2012) introduced an optimal sampling design to address non-response when strata weights are unknown. An All Integer Non-Linear Programming Problem (AINLPP) was developed, to estimate unknown population means. A solution method was formulated with the goal programming (GP) technique using the linear cost function to obtain sample sizes to strata and the LINGO optimization software was used to obtain integer solutions. The proposed compromise criterion using goal programming was effective in determining a compromise allocation in multivariate surveys when travel costs within strata are taken into account and auxiliary information is available.

In real life scenario, sampling cost and variance are known to be stochastic and follow probability distributions (Melaku, 1986). Haseen et al. (2014) developed multivariate stratified sampling technique with non-response, where sampling cost and variance are random, as a multi-objective stochastic programming problem. To achieve the aim, random sampling variances and costs are transformed into a deterministic nonlinear programming problem (NLPP) with chance constraint and a modified E-model. Three different methods were adopted, which are, fuzzy programming, D_1 distance method and goal programming, to obtain the optimal compromise allocation to the developed problem. The linear cost function which incorporates non-response was employed.

Several optimization techniques in multivariate surveys were examined by Raghav et al. (2014) when non-responses are observed with linear cost function. These techniques depend on what is known about the population at first, which can be classified as complete, partial, or null information. The problem is addressed using various methods: the value function technique and goal programming when all information are known, the $\in$-constraint method when only partial information is available, and the distance-based method when insufficient information about the variables exists, all aimed at obtaining the compromise allocation of sample sizes under non-response.

The geometric programming technique in multivariate stratified sampling when non-response is observed was utilized by Shafiullah et al. (2015). The allocation problem was derived in two phases; the first phase problem was worked out as a geometric programming problem and in the second phase the primal-dual relationship theorem. A dual solution was obtained using the formulated geometric programming problem with the help of the quadratic cost function when nonresponse is observed.

Okon and Yahaya (2018) proposed a new gamma cost function in the presence of non-response in multivariate stratified sampling design. This was further modified by Varshney and Mradula, (2019), were they formulated a gamma cost function when non-response is considered and solved it using the LGP method. Contextually, the last known cost approximation in the presence of nonresponse is the gamma cost function proposed by (Okon and Yahaya 2018).

Optimum allocation techniques are best formulated with integer bounds so as to obtain true optimal solution for any given problem. The cost of a survey plays a critical role in determining the allocation of the sample across different strata. Determining the true functional form of the cost function is of great essence. Problem solving comes with real life cost, sometimes from hired labour to equipment maintenance, transportation and several other financial constraints. The linear cost function was developed to take care of enumerating cost; the quadratic cost function is considered if transportation cost is significant and the gamma cost function is relevant if the time taken to obtain data from the units of the population is taken into consideration. The existing cost functions were all proposed based on their functional form. From literature, if additional cost is incurred in the survey the existing cost functions becomes obsolete. Hence, developing a new cost function that will approximate the existing cost functions will be valid in some sense.

Modern electronic devices are currently used in enumeration of household units during surveys. Advancement in modern technology has made it possible to design complex systems whose safety and operations, relies on the reliability of the components that the system is made up of. For instance, a fuse might burn out, a steel column may buckle, or a heat-sensing device or battery might fail. It is widely established that identical components, even when exposed to the same environmental conditions, will fail at different and unpredictable times (Ronald et al. 2007). The linear, quadratic and gamma cost function may not be suitable if survey device fails. This failure may affect enumeration of sampling units and will prompt an additional cost. In this case, a failure cost can be attached per stratum and the time to failure modeled using the Weibull distribution both at the first attempt and a more careful second attempt.

Taking decisions with uncertainty is tasking and unavoidable when dealing with real-life situations. In practice, due to the uncertainty in the population data, the variances as well as the costs should be treated as random variables (Alshaqq et al. 2022 and Mahfouz et al. 2023). Stochastic optimization, which is also known as probabilistic optimization, occurs when some or all the data of the optimization function follows a probability distribution.

In this light, we suggest a new cost function when nonresponse is observed, known as Weibull cost function which approximates the existing cost functions known in literature. The existing cost functions used by other authors, that are being challenged in this work are only relevant in pencil-to-paper era of making questionnaires and will become obsolete with the advent of technologically based questionnaire system. The Weibull cost function as well as the variances will be considered as random variables. The chance constraint and modified E-model technique will be used respectively to convert the sampling cost and variance which are random variables, into a more non-stochastic nonlinear programming problem. The problem will be constructed as a MO-SINLPP.

There are many practical challenges that are yet to be solved in surveys and censuses here in Nigeria and Africa at large. In Nigeria for example, most surveys are not carried out with the correct statistical model in mind. The Weibull cost function suggested in this paper will be a good approximation to the already existing cost functions.

2. Methodology

Let N represents a finite population which is separated into Q mutually exclusive and exhaustive strata with size $N_k (k = 1, 2, ...Q)$ such that $\sum_{k=1}^{Q} N_k = N$. A simple random sample without replacement (SRSWOR) of size n_k is drawn independently from each stratum in a way that $\sum_{k=1}^{Q} n_k = n$. If the stratums are firstly divided into respondents and non-respondents groups which are mutually exclusive and exhaustive. Let the mutually exclusive strata N_k be divided into N_{k1} and N_{k2}, which represents the respondents and non-respondents group respectively in the k^{th} stratum, where $N_{k2} = N_k - N_{k1}$. The values of N_{k1} and N_{k2} are not known before the sample observations are obtained. Let n_k units be drawn from N_k in the k^{th} stratum and let n_{k1} units be members of the respondents group and the remaining n_{k2} units be members of the non-respondents group, that is $n_{k2} = n_k - n_{k1}$.

To achieve a comprehensive result, a subsample of size r_k is obtained from n_{k2} non-respondents in a more careful second attempt. The necessary information is now obtained from the r_k units and this time around we assume that all the r_k units responded.

Let the subsample sizes at the second attempt be given as

$$r_k = \frac{n_{k2}}{h_k}, \quad k = 1, 2, ...Q \tag{1}$$

Equation (1) above are subsample sizes drawn from n_{k2} non-respondents group of the k^{th} stratum, with sample fraction denoted by $1/h_k$ where $h_k \geq 1$.

The estimator for non-response in the study variable is estimated as;

$$\bar{y}_{jk(w)} = \frac{n_{k1}\bar{y}_{jk1} + n_{k2}\bar{y}_{jk2(r_k)}}{n_k} \tag{2}$$

where $\bar{y}_{jk(w)}$ is the weighted mean, with $E(\bar{y}_{jk(w)}) = \bar{Y}_{jk}$.

The corresponding variance of $\bar{y}_{jk(w)}$ will be obtained as

$$V(\bar{y}_{jk(w)}) = \left(\frac{1}{n_k} - \frac{1}{N_k}\right) S_{jk}^2 + \frac{W_{k2}^2 S_{jk2}^2}{r_k} - \frac{W_{k2} S_{jk2}^2}{n_k} \quad (3)$$

where W_{k2} is the non-respondent weights.

The estimate of total population variance of $\bar{y}_{jk(w)}$ which is unbiased is obtained as

$$V(\bar{y}_{jk}) = \sum_{k=1}^{Q} W_k^2 V(\bar{y}_{jk(w)}) \quad (4)$$

$$V(\bar{y}_{jk}) = \sum_{k=1}^{Q} W_k^2 \left[\left(\frac{1}{n_k} - \frac{1}{N_k}\right) S_{jk}^2 + \frac{W_{k2}^2 S_{jk2}^2}{r_k} - \frac{W_{k2} S_{jk2}^2}{n_k} \right] \quad (5)$$

W_k is the stratum weights. If finite population correction (FPC) is ignored in Equation (5), then the following equation will be obtained, that is,

$$V(\bar{y}_{jk}) = \sum_{k=1}^{Q} \frac{W_k^2 \left(S_{jk}^2 - W_{k2} S_{jk2}^2\right)}{n_k} + \sum_{k=1}^{Q} \frac{W_k^2 W_{k2}^2 S_{jk2}^2}{r_k} \quad (6)$$

3. $\bar{Y}_{jk} = N_k^{-1} \sum_{i=1}^{N_k} y_{jki}$: stratum mean in the k^{th} stratum for j^{th} characteristics;

$S_{jk}^2 = \left(N_k - 1\right)^{-1} \sum_{i=1}^{N_k} \left(y_{jki} - \bar{Y}_{jk}\right)^2$: stratum variance in the k^{th} stratum for j^{th} characteristics;

$S_{jk2}^2 = \left(N_{k2} - 1\right)^{-1} \sum_{i=1}^{N_{k2}} \left(y_{jki} - \bar{Y}_{jk2}\right)^2$: stratum variance for non-response group in the k^{th} stratum.

$\bar{y}_{jk1} = n_{k1}^{-1} \sum_{i=1}^{n_{k1}} y_{jki}$: sample mean of the first attempt units in the k^{th} stratum.

$\bar{y}_{jk2(r_k)} = r_k^{-1} \sum_{i=1}^{r_k} y_{jki}$: sample mean of units respond on second call in the k^{th} stratum.

2.1 Cost Function with Non-Response Stratified Sampling

The linear cost function if nonresponse is considered is given as

$$C = \sum_{k=1}^{Q} c_{k0} n_k + \sum_{k=1}^{Q} c_{k1} n_{k1} + \sum_{k=1}^{Q} c_{k2} n_{k2} \quad (7)$$

The quantity $W_{k1}n_k$ is used as expected value since n_{k1} is not known until the first attempt is made. A cost function of the form is expected;

$$C = \sum_{k=1}^{Q}\left(c_{k0} + c_{k1}W_{k1}\right)n_k + \sum_{k=1}^{Q} c_{k2}r_k \tag{8}$$

where C is the total budget, c_{k0} is the cost attributed to selecting the n_k units.

In real life events, when determining survey cost, both measurement unit costs and travel costs within strata are key factors. If travel costs can be substantially measured, the cost function becomes non-linear. A non-linear cost function, which accounts for measurement unit costs and travel costs within strata, provides a better approximation of the survey's actual budget. Beardwood et al. (1959) derived the distance between m randomly dispersed destinations within an area to be asymptotically proportional to $\sqrt{m}$ for large m. Therefore, the cost of visiting the n_k and r_k units in the k^{th} stratum dispersed all over the stratum can be taken as $t_{ko}\sqrt{n_k}$ and $t_{k2}\sqrt{r_k}$ $k = 1, 2, ..., Q$, approximately. t_k is the cost of travelling around selected units in the k^{th} stratum. The quadratic cost function considered is,

$$C = \sum_{k=1}^{Q}\left(c_{k0} + c_{k1}W_{k1}\right)n_k + \sum_{k=1}^{Q} c_{k2}r_k + \sum_{k=1}^{Q} t_{k0}\sqrt{n_k} + \sum_{k=1}^{Q} t_{k2}\sqrt{r_k} \tag{9}$$

where t_{ko} is per unit traveling cost at first attempt, t_{k2} is travel cost for visiting the non-respondent unit within the k^{th} stratum and r_k are samples gotten from the non-respondent unit at second attempt.

If the labor cost which represents the time taken to obtain data from all selected units is significant, then the cost function considered by Abubakar and Okon (2018) is given as

$$C = \sum_{k=1}^{Q}\left(c_{k0} + c'_{k1}W_{k1}\right)n_k + \sum_{k=1}^{Q} c'_{k2}r_k + \sum_{k=1}^{Q} t_{k0}\sqrt{n_k} + \sum_{k=1}^{Q} t_{k2}\sqrt{r_k} + \alpha\sum_{k=1}^{Q}\frac{n_k}{\lambda} + \omega\sum_{k=1}^{Q}\frac{r_k}{\lambda^*} \tag{10}$$

where α and ω are the effect of labour cost at the first and second attempt respectively with rates λ and λ^*. $c'_{k1} = c_{k1} + u_{k1}$ and $c'_{k2} = c_{k2} + u_{k2}$ are the enumerating and processing unit cost with reward paid to respondents and non-respondents in k strata.

2.2 Weibull Cost Function with Non-Response

In survey enumeration, failure of device is inevitable. Survey devices have tendencies of wearing out with time. However, in many complex systems or real life situations, failure rate exhibits a bathtub shape, it shows a form of decrease in the beginning, then it stabilizes and towards the end of its life cycle it enters the wear out phase, at this point the failure rate begins to rise. A constant failure rate means that the component's lifetime T will lead to an exponential distribution, with the rate parameter equal to the constant failure rate λ. The probability distribution function and cumulative distribution function is given as

$$\left.\begin{aligned} f(t) &= \lambda e^{-\lambda t} \\ F(t) &= 1-e^{-\lambda t} \end{aligned}\right\} \qquad (11)$$

for an irreparable component, the mean time to failure (MTTF) is equal to expected lifetime

$$E(t) = \int_0^{\infty} t f(t) dt = \frac{1}{\lambda} \qquad (12)$$

In some circumstances, the simplifying assumption is not suitable, particularly during the early stages and wear-out phases of a component's lifespan. In such cases, a Weibull distribution will be preferred. Now if the time to failure for a device affecting sampling units follows a Weibull distribution with state space $t \geq 0$, then the probability distribution function for time to failure is approximately

$$f(t,\theta,\gamma) = \frac{\gamma}{\theta}\left(\frac{t}{\theta}\right)^{\gamma-1} \ell^{-\left(\frac{t}{\theta}\right)^{\gamma}}, \qquad t \geq 0 \qquad (13)$$

with $\gamma > 0$ as shape parameter and $\theta > 0$ scale parameter for the distribution (Karuna and Khan 2017). The equation given in the Equation (15) is a two parameter Weibull distribution.

Proposition 1: If the time to failure attributed to the sampling units, that is, the selected units of respondents at the first attempt and more careful second attempt for the subsample of non-respondents, with rates θ_1^* and θ_2^* having an exponential distribution. The sum of independent identically exponential random variables will follow Weibull distributions with parameters (n_k, θ_1^*) and (r_k, θ_2^*) respectively. Summing up to all strata, the Weibull functions will have parameters $\left(\sum_{k=1}^{Q} n_k, \theta_1^*\right)$ and $\left(\sum_{k=1}^{Q} r_k, \theta_2^*\right)$ respectively.

The proposed Weibull cost function with non-response is given as

$$C=\sum_{k=1}^{Q}\left(c_{k0}+c'_{k1}W_{k1}\right)n_k+\sum_{k=1}^{Q}c'_{k2}r_k+\sum_{k=1}^{Q}t_{k0}\sqrt{n_k}+\sum_{k=1}^{Q}t_{k2}\sqrt{r_k}+\sum_{k=1}^{Q}\beta_k\frac{n_k}{\lambda}+\sum_{k=1}^{Q}\beta_k^*\frac{r_k}{\lambda^*}$$
$$\varphi_k\frac{\sum_{k=1}^{Q}n_k}{\theta_1^*}\int_0^{\infty}\left(\frac{t_1}{\theta_1^*}\right)^{\sum_{k=1}^{Q}n_k-1}\ell^{-\left(\frac{t_1}{\theta_1^*}\right)^{\sum_{k=1}^{Q}n_h}}dt_1+\varphi_k^*\frac{\sum_{k=1}^{Q}r_k}{\theta_2^*}\int_0^{\infty}\left(\frac{t_2}{\theta_2^*}\right)^{\sum_{k=1}^{Q}r_k-1}\ell^{-\left(\frac{t_2}{\theta_2^*}\right)^{\sum_{k=1}^{Q}n_k}}dt_2 \quad (13)$$

β_k, β_k^* and φ_k, φ_k^* are the cost of a unit time of labour and failure for the respondent and non-respondents units respectively. In Equation (13) the cost of a unit time of labour is taken over all strata. Let,

$$\left.\begin{aligned}\mathrm{K}_1&=\varphi_k\frac{\sum_{k=1}^{Q}n_k}{\theta_1^*}\int_0^{\infty}\left(\frac{t_1}{\theta_1^*}\right)^{\sum_{k=1}^{Q}n_k-1}\ell^{-\left(\frac{t_1}{\theta_1^*}\right)^{\sum_{k=1}^{Q}n_k}}dt_1\\ \mathrm{K}_2&=\varphi_k^*\frac{\sum_{k=1}^{Q}r_k}{\theta_2^*}\int_0^{\infty}\left(\frac{t_2}{\theta_2^*}\right)^{\sum_{k=1}^{Q}r_k-1}\ell^{-\left(\frac{t_2}{\theta_2^*}\right)^{\sum_{k=1}^{Q}n_k}}dt_2\end{aligned}\right\} \quad (14)$$

Then Equation (13) will become

$$C=\sum_{k=1}^{Q}\left(c_{k0}+c'_{k1}W_{k1}\right)n_k+\sum_{k=1}^{Q}c'_{k2}r_k+\sum_{k=1}^{Q}t_{k0}\sqrt{n_k}+\sum_{k=1}^{Q}t_{k2}\sqrt{r_k}+\sum_{k=1}^{Q}\beta_k\frac{n_k}{\lambda}+\sum_{k=1}^{Q}\beta_k^*\frac{r_k}{\lambda^*}+ \quad (15)$$
$$\varphi_k\mathrm{K}_1+\varphi_k^*\mathrm{K}_2.$$

from Equation (13), let

$$\left.\begin{aligned}\sum_{k=1}^{Q}\varphi_k E(T_k)&=\sum_{k=1}^{Q}\varphi_k\left[\frac{n_k}{\theta_1^*}\int_0^{\infty}\left(\frac{t_1}{\theta_1^*}\right)^{n_k-1}\ell^{-\left(\frac{t_1}{\theta_1^*}\right)^{n_k}}dt_1\right]\\ \sum_{k=1}^{Q}\varphi_k^* E(T_k^*)&=\sum_{k=1}^{Q}\varphi_k^*\left[\frac{r_k}{\theta_2^*}\int_0^{\infty}\left(\frac{t_2}{\theta_2^*}\right)^{r_k-1}\ell^{-\left(\frac{t_2}{\theta_2^*}\right)^{n_k}}dt_2\right]\end{aligned}\right\} \quad (16)$$

Then the MTTF for Equation (16) will be

$$\left.\begin{aligned}\sum_{k=1}^{Q}\varphi_k E(T_k)&=\sum_{k=1}^{Q}\varphi_k\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)\\ \sum_{k=1}^{Q}\varphi_k^* E(T_k^*)&=\sum_{k=1}^{Q}\varphi_k^*\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right)\end{aligned}\right\} \quad (17)$$

The cost function in Equation (13) can be rewritten as

$$C=\sum_{k=1}^{Q}\left(c_{k0}+c'_{k1}W_{k1}\right)n_k+\sum_{k=1}^{Q}c'_{k2}r_k+\sum_{k=1}^{Q}t_{k0}\sqrt{n_k}+\sum_{k=1}^{Q}t_{k2}\sqrt{r_k}+\sum_{k=1}^{Q}\beta_k\frac{n_k}{\lambda}+\sum_{k=1}^{Q}\beta_k^*\frac{r_k}{\lambda^*}+\sum_{k=1}^{Q}\varphi_k\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)+\sum_{k=1}^{Q}\varphi_k^*\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right) \quad (18)$$

2.3 Stochastic Form of the Weibull Cost Function with Non-Response

Theorem 1: Assuming the stochastic vector $\tau=(\mathrm{c}_{k0},c'_{k1},c'_{k2},t_{k0},t_{k1},\beta_k,\beta_k^*,\varphi_k,\varphi_k^*,C)$, where $h=1,2,...,L$ and the function $g(n_k,r_k;\tau)$ has the form;

$$g(n_k,r_k;\tau)=\sum_{k=1}^{Q}\left(c_{k0}+c'_{k1}W_{k1}\right)n_k+\sum_{k=1}^{Q}c'_{k2}r_k+\sum_{k=1}^{Q}t_{k0}\sqrt{n_k}+\sum_{k=1}^{Q}t_{k2}\sqrt{r_k}+\sum_{k=1}^{Q}\beta_k\frac{n_k}{\lambda}+\sum_{k=1}^{Q}\beta_k^*\frac{r_k}{\lambda^*}+\sum_{k=1}^{Q}\varphi_k\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)+\sum_{k=1}^{Q}\varphi_k^*\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right)-C \quad (19)$$

If the stochastic vector contains cost that are normally distributed and independent random variables then, $P\{g(n_k,r_k;\tau)\le C\}\ge p_0$ if and only if $E(g(n_k,r_k;\tau))+G_\alpha\sqrt{V(g(n_k,r_k;\tau))}\le C$.

where p_0, $0\le p_0\le 1$ is a specified probability.

Proof: If $c_{k0}\sim N(\mu_{c_{k0}},\sigma^2_{c_{k0}})$, $c'_{k1}\sim N(\mu_{c'_{k1}},\sigma^2_{c'_{k1}})$, $c'_{k2}\sim N(\mu_{c'_{k2}},\sigma^2_{c'_{k2}})$, $t_{k0}\sim N(\mu_{t_{k0}},\sigma^2_{t_{k2}})$, $\beta_k\sim N(\mu_{\beta_k},\sigma^2_{\beta_k})$, $\beta_k^*\sim N(\mu_{\beta_k^*},\sigma^2_{\beta_k^*})$, $\varphi_k\sim N(\mu_{\varphi_k},\sigma^2_{\varphi_k})$, and $\varphi_k^*\sim N(\mu_{\varphi_k^*},\sigma^2_{\varphi_k^*})$

The cost constraint can be written as

$$p\left[\mathrm{g}(n_k,r_k;\tau)\le C\right]\ge p_0 \quad (20)$$

$$\Rightarrow p\left[\frac{\mathrm{g}(n_k,r_k;\tau)-E(\mathrm{g}(n_k,r_k;\tau))}{\sqrt{V(\mathrm{g}(n_k,r_k;\tau))}}\le\frac{C-E(\mathrm{g}(n_k,r_k;\tau))}{\sqrt{V(\mathrm{g}(n_k,r_k;\tau))}}\right]\ge p_0 \quad (21)$$

Equation (21) is a the standard normal variate with $Z\sim N(0,1)$. The standard normal variate is evaluated at z, with $\Phi(z)$ as cumulative density function. If the standard normal variate at which $\Phi(G_\alpha)=p_0$, is represented as G_α. We can write the constraints as

$$\Phi\left[\frac{C-E(\mathrm{g}(n_k,r_k;\tau))}{\sqrt{V(\mathrm{g}(n_k,r_k;\tau))}}\right]\ge\Phi(G_\alpha) \quad (22)$$

Φ is the standardize normal distribution function. The inequality will be satisfied if and only if

$$\left[\frac{C-E(\mathrm{g}(n_k,r_k;\tau))}{\sqrt{V(\mathrm{g}(n_k,r_k;\tau))}}\right]\geq G_\alpha \tag{23}$$

This also implies that

$$\left.\begin{aligned}E\left(\mathrm{g}(n_k,r_k;\tau)\right)=E\left(\sum_{k=1}^{Q}c_{k0}n_k\right)+E\left(\sum_{k=1}^{Q}c'_{k1}W_{k1}n_k\right)+E\left(\sum_{k=1}^{Q}c'_{k2}r_k\right)+\\ E\left(\sum_{k=1}^{Q}t_{k0}\sqrt{n_k}\right)+E\left(\sum_{k=1}^{Q}t_{k2}\sqrt{r_k}\right)+E\left(\sum_{k=1}^{Q}\beta_k\frac{n_2}{\lambda}\right)+\\ E\left(\sum_{k=1}^{Q}\beta_k^*\frac{r_k}{\lambda^*}\right)+E\left[\sum_{k=1}^{Q}\varphi_k\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)\right]+E\left[\sum_{k=1}^{Q}\varphi_k^*\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right)\right]\end{aligned}\right\} \tag{24}$$

with variance obtained as

$$=\sum_{k=1}^{Q}\left(\begin{aligned}&n_k\mu_{c_{k0}}+W_{k1}n_k\mu_{c'_{k1}}+r_k\mu_{c'_{k2}}+\mu_{t_{k0}}\sqrt{n_k}+\mu_{t_{k2}}\sqrt{r_k}\\&+\frac{n_k}{\lambda}\mu_{\beta_k}+\frac{r_k}{\lambda^*}\mu_{\beta_k^*}+\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)\mu_{\varphi_k}+\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right)\mu_{\varphi_k^*}\end{aligned}\right) \tag{25}$$

$$\begin{aligned}V\left(\mathrm{g}(n_k,r_k;\tau)\right)=&\sum_{k=1}^{Q}n_k^2\sigma_{c_{k0}}^2+\sum_{k=1}^{Q}n_k^2W_{k1}^2\sigma_{c'_{k1}}^2+\sum_{k=1}^{Q}r_k^2\sigma_{c'_{k2}}^2+\sum_{k=1}^{Q}n_k\sigma_{t_{k0}}^2+\sum_{k=1}^{Q}r_k\sigma_{t_{k2}}^2+\\&\sum_{k=1}^{Q}\frac{n_k^2}{\lambda^2}\sigma_{\beta_k}^2+\sum_{k=1}^{Q}\frac{r_k^2}{\lambda^{*2}}\sigma_{\beta_k^*}^2+\theta_1^{2*}\Gamma\left(\frac{1}{n_k}+1\right)^2\sigma_{\varphi_k}^2+\theta_2^{2*}\Gamma\left(\frac{1}{r_k}+1\right)^2\sigma_{\varphi_k^*}^2\end{aligned} \tag{26}$$

we can write Equation (23) as

$$E(g(n_k,r_k;\tau))+G_\alpha\sqrt{V(g(n_k,r_k;\tau))}\leq C \tag{27}$$

Substituting we will obtain the Weibull cost function with nonresponse in its stochastic form as

$$\begin{aligned}&\sum_{k=1}^{Q}\left(\begin{aligned}&n_k\mu_{c_{k0}}+W_{k1}n_k\mu_{c'_{k1}}+r_k\mu_{c'_{k2}}+\mu_{t_{k0}}\sqrt{n_k}+\mu_{t_{k2}}\sqrt{r_k}\\&+\frac{n_k}{\lambda}\mu_{\beta_k}+\frac{r_k}{\lambda^*}\mu_{\beta_k^*}+\theta_1^*\Gamma\left(\frac{1}{n_k}+1\right)\mu_{\varphi_k}+\theta_2^*\Gamma\left(\frac{1}{r_k}+1\right)\mu_{\varphi_k^*}\end{aligned}\right)\\&+G_\alpha\sqrt{\begin{aligned}&\sum_{k=1}^{Q}n_k^2\sigma_{c_{k0}}^2+\sum_{k=1}^{Q}n_k^2W_{k1}^2\sigma_{c'_{k1}}^2+\sum_{k=1}^{Q}r_h^2\sigma_{c'_{k2}}^2+\sum_{k=1}^{Q}n_k\sigma_{t_{k0}}^2+\sum_{k=1}^{Q}r_k\sigma_{t_{k2}}^2+\\&\sum_{k=1}^{Q}\frac{n_k^2}{\lambda^2}\sigma_{\beta_k}^2+\sum_{k=1}^{Q}\frac{r_k^2}{\lambda^{*2}}\sigma_{\beta_k^*}^2+\theta_1^{2*}\Gamma\left(\frac{1}{n_k}+1\right)^2\sigma_{\varphi_k}^2+\theta_2^{2*}\Gamma\left(\frac{1}{r_k}+1\right)^2\sigma_{\varphi_k^*}^2\end{aligned}}\leq C\end{aligned} \tag{28}$$

2.4 Equivalent Deterministic Weibull Cost Constraint

Let X be a feasible space of a multi-objective stochastic optimization problem (MOSOP). If the mean and variance constants are unknown it can be replaced with the estimator of mean $\hat{E}(.)$ and $\hat{V}(.)$ as given below

$$X = \begin{cases} n_k, r_k \in R \,|\, \sum_{k=1}^{Q} (\bar{c}_{k0} + \bar{c}'_{k1} W_{k1}) n_h + \sum_{k=1}^{Q} \bar{c}'_{k2} r_k + \sum_{k=1}^{Q} \bar{t}_{k0} \sqrt{n_k} + \sum_{k=1}^{Q} \bar{t}_{k2} \sqrt{r_k} + \sum_{k=1}^{Q} \bar{\beta}_h \frac{n_h}{\lambda} + \sum_{k=1}^{Q} \bar{\beta}^*_h \frac{r_h}{\lambda^*} + \\ \sum_{k=1}^{Q} \bar{\varphi}_k \theta_1^* \Gamma\left(\frac{1}{n_k} + 1\right) + \sum_{k=1}^{Q} \bar{\varphi}^*_k \theta_2^* \Gamma\left(\frac{1}{r_k} + 1\right) + \\ G_\alpha \sqrt{\sum_{k=1}^{Q} n_k^2 \hat{\sigma}^2_{c_{k0}} + \sum_{k=1}^{Q} n_k^2 W_{k1}^2 \hat{\sigma}^2_{c'_{k1}} + \sum_{k=1}^{Q} r_k^2 \hat{\sigma}^2_{c'_{k2}} + \sum_{k=1}^{Q} n_k \hat{\sigma}^2_{t_{k0}} + \sum_{k=1}^{Q} r_k \hat{\sigma}^2_{t_{k2}} + \\ \sum_{k=1}^{Q} \frac{n_k^2}{\lambda^2} \hat{\sigma}^2_{\beta_k} + \sum_{k=1}^{Q} \frac{r_k^2}{\lambda^{*2}} \hat{\sigma}^2_{\beta^*_k} + \theta_1^{2*} \Gamma\left(\frac{1}{n_k} + 1\right) \hat{\sigma}^2_{\varphi_k} + \theta_2^{2*} \Gamma\left(\frac{1}{r_h} + 1\right) \hat{\sigma}^2_{\varphi^*_h}} \\ for\ k = 1, 2..., Q,\ 2 \le n_k \le N_k,\ 2 \le r_k \le \hat{n}_{k2} \end{cases}$$

3 Modified Expected Value (Modified-E) Model for the Stochastic Sampling Variance

s^2_{jk} and s^2_{jk2} are sample variances which are assumed to be random variables. Until a second attempt is made, the sample size n_{k2} of the non-respondent is not known, but its estimates $\hat{n}_{k2} = W_{k1} n_k$, can be used in this regard.

Consider the random variable ξ_{k2} defined by

$$\xi_{k2} = \frac{1}{\hat{n}_{k2} - 1} \sum_{i=1}^{\hat{n}_{k2}} (y_{jki} - \bar{Y}_{jk2})^2, \tag{29}$$

which is the limiting distribution of the sample variances. Having asymptotic normal distribution with mean given as

$$E[\xi_{k2}] = \frac{1}{\hat{n}_{k2} - 1} \sum_{i=1}^{\hat{n}_{k2}} E(y_{jki} - \bar{Y}_{jk2})^2 = \frac{\hat{n}_{k2}}{\hat{n}_{k2} - 1} S^2_{jk2}$$

(30)

Similarly by independence,

$$\operatorname{var}\left(\xi_{k2}\right)=\frac{\hat{n}_{k2}}{\left(\hat{n}_{k2}-1\right)^{2}}\operatorname{var}\left(y_{ijk}-\bar{Y}_{jk2}\right) \tag{31}$$

$$=\frac{\hat{n}_{k2}}{\left(\hat{n}_{k2}-1\right)^{2}}\left[C_{jk2}^{4}-\left(S_{jk2}^{2}\right)^{2}\right] \tag{32}$$

fourth central moment given as C_{jk2}^{4} in Equation (32) above can be evaluated as

$$C_{jk2}^{4}=\frac{1}{\hat{N}_{k2}}\sum_{k=1}^{\hat{N}_{k2}}\left(y_{ijk}-\bar{Y}_{jk2}\right)^{4} \tag{33}$$

If we observe closely, we can see that

$$s_{jk2}^{2}=\xi_{k}-\frac{\hat{n}_{k2}}{\hat{n}_{k2}-1}\left(\bar{y}_{jk2}-\bar{Y}_{jk2}\right) \tag{34}$$

where $\frac{\hat{n}_{k2}}{\hat{n}_{k2}-1}\xrightarrow{p}1$ and $\left(\bar{y}_{jk2}-\bar{Y}_{jk2}\right)\xrightarrow{p}0$. Then, the sample variances s_{jk2}^{2} have an asymptotical normal distribution, that is, $s_{jk2}^{2}\xrightarrow{a}N(E(\xi_{k}),\operatorname{var}(\xi_{k}))$. Let,

$$V(\bar{y}_{j(w)})=\sum_{k=1}^{Q}W_{k}^{2}\left[\frac{s_{jk}^{2}}{n_{k}}+\frac{W_{k2}^{2}s_{jk2}^{2}}{r_{k}}-\frac{W_{k2}s_{jk2}^{2}}{n_{k}}\right] \tag{35}$$

The variance function in Equation (35) has a normal distribution, with

$$E\left[\hat{V}(\bar{y}_{j(w)})\right]=\sum_{k=1}^{Q}\left(\frac{W_{k}^{2}s_{yjk}^{2}}{n_{k}}+\frac{W_{k}^{2}W_{k2}^{2}s_{yjk2}^{2}}{r_{k}}-\frac{W_{k}^{2}W_{k2}s_{yjk2}^{2}}{n_{k}}\right) \tag{36}$$

$$=\sum_{k=1}^{Q}\left(\frac{W_{k}^{2}E\left(\xi_{k}\right)}{n_{k}}+\frac{W_{k}^{2}W_{k2}^{2}E\left(\xi_{k2}\right)}{r_{k}}-\frac{W_{k}^{2}W_{k2}E\left(\xi_{k2}\right)}{n_{k}}\right) \tag{37}$$

$$=\sum_{k=1}^{Q}\left(\frac{W_{k}^{2}S_{jk}^{2}}{\left(n_{k}-1\right)}+\frac{\hat{n}_{k2}W_{k}^{2}W_{k2}^{2}S_{jk2}^{2}}{r_{k}(\hat{n}_{k2}-1)}-\frac{\hat{n}_{k2}W_{k}^{2}W_{k2}S_{jk2}^{2}}{n_{k}(\hat{n}_{k2}-1)}\right) \tag{38}$$

$$E\left[\hat{V}(\bar{y}_{j(w)})\right]=\sum_{k=1}^{Q}\left(\frac{W_k^2S_{jk}^2}{(n_k-1)}+\frac{n_kW_k^2W_{k2}^3S_{jk2}^2}{r_k(n_kW_{k2}-1)}-\frac{W_{k2}^2W_k^2S_{jk2}^2}{n_kW_{k2}-1}\right) \quad (39)$$

with,

$$V\left[\hat{V}(\bar{y}_{j(w)})\right]=\sum_{k=1}^{Q}V\left(\frac{W_k^2s_{jk}^2}{n_k}+\frac{W_k^2W_{k2}^2s_{jk2}^2}{r_k}-\frac{W_k^2W_{k2}s_{jk2}^2}{n_k}\right) \quad (40)$$

$$=\sum_{k=1}^{Q}\left(\frac{W_k^4V(\xi_k)}{n_k^2}+\frac{W_k^4W_{k2}^4V(\xi_{k2})}{r_k^2}-\frac{W_k^4W_{k2}^2V(\xi_{k2})}{n_k^2}\right) \quad (41)$$

$$=\sum_{k=1}^{Q}\left(\frac{W_k^4\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_k\left(n_k-1\right)^2}+\frac{\hat{n}_{k2}W_k^4W_{k2}^4\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{r_k^2(\hat{n}_{k2}-1)^2}-\frac{\hat{n}_{k2}W_k^4W_{k2}^2\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_k^2(\hat{n}_{k2}-1)^2}\right)$$

(42)

$$=\sum_{k=1}^{Q}\left(\frac{W_k^4\left(C_{jk}^4-\left(S_{jk}^2\right)^2\right)}{n_k\left(n_k-1\right)^2}+\frac{n_kW_k^4W_{k2}^5\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{r_k^2(n_kW_{k2}-1)^2}-\frac{W_k^4W_{k2}^3\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_k(n_kW_{k2}-1)^2}\right)$$

(43)

The objective function in the E-model form is written as;

$$f_j(n_k,r_k)=e_1E(\hat{V}(\bar{y}_{j,(w)}))+e_2\sqrt{V(\hat{V}(\bar{y}_{j,(w)}))} \quad (44)$$

Here $e_1,e_2>0$. Roa (2009) suggested that $e_1+e_2=1$. It illustrate the relative significance of the expectation and variance of $V(\bar{y}_{j(w)})$.

Inserting Equation (39) and Equation (43) into Equation (44), we will obtain

$$f_j(n_k,r_k)=e_1\sum_{k=1}^{Q}\left(\frac{W_k^2S_{jk}^2}{(n_k-1)}+\frac{n_kW_k^2W_{k2}^3S_{jk2}^2}{r_k(n_kW_{k2}-1)}-\frac{W_{k2}^2W_k^2S_{jk2}^2}{n_kW_{k2}-1}\right)$$

$$+e_2\sqrt{\sum_{k=1}^{Q}\left(\frac{W_k^4\left(C_{jk}^4-\left(S_{jk}^2\right)^2\right)}{n_k\left(n_k-1\right)^2}+\frac{n_kW_k^4W_{k2}^5\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{r_k^2(n_kW_{k2}-1)^2}-\frac{W_k^4W_{k2}^3\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_k(n_kW_{k2}-1)^2}\right)} \quad (45)$$

4 Stochastic Individual Allocations

The individual allocation for each characteristic as;

$$
\left.
\begin{aligned}
&\textit{Minimize}\ \ f_j(n_{jk}, r_{jk}) = \mathrm{Z}_j = e_1 \sum_{k=1}^{Q} \left(\frac{W_k^2 S_{jk}^2}{\left(n_{jk}-1\right)} + \frac{n_{jk} W_k^2 W_{k2}^3 S_{jk2}^2}{r_{jk}(n_{jk}W_{k2}-1)} - \frac{W_{k2}^2 W_k^2 S_{jk2}^2}{n_{jk}W_{k2}-1} \right) \\
&+ e_2 \sqrt{\sum_{k=1}^{Q} \left(\frac{W_k^4 \left(C_{jk}^4 - \left(S_{jk}^2\right)^2\right)}{n_{jk}\left(n_{jk}-1\right)^2} + \frac{n_{jk} W_k^4 W_{k2}^5 \left(C_{jk2}^4 - \left(S_{jk2}^2\right)^2\right)}{r_{jk}^2 (n_{jk}W_{k2}-1)^2} - \frac{W_k^4 W_{k2}^3 \left(C_{jk2}^4 - \left(S_{jk2}^2\right)^2\right)}{n_{jk}(n_{jk}W_{k2}-1)^2} \right)} \\
&\textit{subject to} \\
&\sum_{k=1}^{Q} \left(\bar{c}_{k0} + \bar{c}_{k1}' W_{k1}\right) n_{jk} + \sum_{k=1}^{Q} \bar{c}_{k2}' r_{jk} + \sum_{k=1}^{Q} \bar{t}_{k0} \sqrt{n_{jk}} + \sum_{k=1}^{Q} \bar{t}_{k2} \sqrt{r_{jk}} + \sum_{k=1}^{Q} \bar{\beta}_k \frac{n_{jk}}{\lambda} + \\
&\sum_{k=1}^{Q} \bar{\beta}_k^* \frac{r_{jh}}{\lambda^*} + \sum_{k=1}^{Q} \bar{\varphi}_k \theta_1^* \Gamma\left(\frac{1}{n_{jk}}+1\right) + \sum_{k=1}^{Q} \bar{\varphi}_k^* \theta_2^* \Gamma\left(\frac{1}{r_{jk}}+1\right) + \\
&G_\alpha \sqrt{\begin{aligned}
&\sum_{k=1}^{Q} n_{jk}^2 \hat{\sigma}_{c_{k0}}^2 + \sum_{k=1}^{Q} n_{jk}^2 W_{k1}^2 \hat{\sigma}_{c_{k1}'}^2 + \sum_{k=1}^{Q} r_{jk}^2 \hat{\sigma}_{c_{k2}'}^2 + \sum_{k=1}^{Q} n_{jk} \hat{\sigma}_{t_{k0}}^2 + \sum_{k=1}^{Q} r_{jk} \hat{\sigma}_{t_{k2}}^2 + \\
&\sum_{k=1}^{Q} \frac{n_{jk}^2}{\lambda^2} \hat{\sigma}_{\beta_k}^2 + \sum_{k=1}^{Q} \frac{r_{jk}^2}{\lambda^{*2}} \hat{\sigma}_{\beta_k^*}^2 + \theta_1^{2*} \Gamma\left(\frac{1}{n_{jk}}+1\right) \hat{\sigma}_{\varphi_k}^2 + \theta_2^{2*} \Gamma\left(\frac{1}{r_{jk}}+1\right) \hat{\sigma}_{\varphi_k^*}^2
\end{aligned}} \le C \\
&2 \le n_{jk} \le N_k,\ 2 \le r_{jk} \le \hat{n}_{k2},\ n_{jk} \text{ and } r_{jk} \in \square\ \ \forall\ \ k = 1, 2, ..., Q
\end{aligned}
\right\}
$$

(46)

The stochastic individual allocations for j^{th} $(j = 1, 2, ..., p)$ characteristics can be obtained by solving the above MO-SINLPP. Z_j^* will be the objective or variance function values of Z_j under the individual allocations.

5 Compromise Solution with Fuzzy Programming

Consider $\bar{\xi}_j$ to be the solutions obtained individually for the problem in Equation (46), that is, $\bar{\xi}_j = (n_1, ..., n_Q, r_1, ..., r_Q)$. Let $\bar{\xi}_j^B$ be the best solution obtained individually for the j^{th} characteristics subject to the cost constraints.

$$Z_j^B = Z(\bar{\xi}_j^B) = \min Z_j(\bar{\xi}_j);\ \ j = 1, 2, ..., p$$

The corresponding fuzzy goal is of the form

$$Z_j(\bar{\xi}_j) \geq \mathrm{Z}_j^B;\ \ j=1,2,...,p$$

Next, is by constructing a payoff matrix using the individual solution which is best with the variance values at each best solution.

$$\begin{array}{c} \\ \bar{\xi}_1 \\ \bar{\xi}_2 \\ \vdots \\ \bar{\xi}_p \end{array} \begin{array}{c} \begin{array}{cccc} Z_1(\bar{\xi}) & Z_2(\bar{\xi}) & \cdots & Z_p(\bar{\xi}) \end{array} \\ \begin{pmatrix} Z_1(\bar{\xi}_1^B) & Z_2(\bar{\xi}_1) & \ldots & Z_p(\bar{\xi}_1) \\ Z_1(\bar{\xi}_2) & Z_2(\bar{\xi}_2^B) & \ldots & Z_p(\bar{\xi}_2) \\ \vdots & \vdots & \ddots & \vdots \\ Z_1(\bar{\xi}_p) & Z_2(\bar{\xi}_p) & \ldots & Z_p(\bar{\xi}_p) \end{pmatrix} \end{array}$$

The minimum value of each column of $Z_j(\bar{\xi})$ ($j=1,2,...,p$) gives the lower tolerance limit and the maximum value gives the least acceptable level of achievement. Let Z_j^B and Z_j^L represent the upper and the lower tolerance limit respectively. Now, let $d_j = Z_j^L - Z_j^B$.

Then the solution method using fuzzy programming with MO-SINLPP in Equation (46) is given as

$$\left.\begin{array}{l} \textit{Minimize}\ \ \rho \\ \textit{subject to} \\ e_1\sum_{k=1}^{Q}\left(\frac{W_k^2 S_{jk}^2}{(n_{kca}-1)}+\frac{n_{kca}W_k^2W_{k2}^3S_{jk2}^2}{r_{kca}(n_{kca}W_{k2}-1)}-\frac{W_{k2}^2W_k^2S_{jk2}^2}{n_{kca}W_{k2}-1}\right)+ \\ e_2\sqrt{\sum_{k=1}^{Q}\left(\frac{W_k^4\left(C_{jk}^4-\left(S_{jk}^2\right)^2\right)}{n_{kca}\left(n_{kca}-1\right)^2}+\frac{n_{kca}W_k^4W_{k2}^5\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{r_{kca}^2(n_{kca}W_{k2}-1)^2}-\frac{W_k^4W_{k2}^3\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_{kca}(n_{kca}W_{k2}-1)^2}\right)}-\rho d_j \leq Z_j^B \\ \sum_{k=1}^{Q}\left(\bar{c}_{k0}+\bar{c}_{k1}'W_{k1}\right)n_{kca}+\sum_{k=1}^{Q}\bar{c}_{k2}'r_{kca}+\sum_{k=1}^{Q}\bar{t}_{k0}\sqrt{n_{kca}}+\sum_{k=1}^{Q}\bar{t}_{k2}\sqrt{r_{kca}}+\sum_{k=1}^{Q}\bar{\beta}_k\frac{n_{kca}}{\lambda}+ \\ \sum_{k=1}^{Q}\bar{\beta}_k^*\frac{r_{kca}}{\lambda^*}+\sum_{k=1}^{Q}\bar{\varphi}_k\theta_1^*\Gamma\left(\frac{1}{n_{kca}}+1\right)+\sum_{k=1}^{Q}\bar{\varphi}_k^*\theta_2^*\Gamma\left(\frac{1}{r_{kca}}+1\right)+ \\ G_\alpha\sqrt{\begin{array}{l}\sum_{k=1}^{Q}n_{kca}^2\hat{\sigma}_{c_{k0}}^2+\sum_{k=1}^{Q}n_{kca}^2W_{k1}^2\hat{\sigma}_{c_{k1}'}^2+\sum_{k=1}^{Q}r_{kca}^2\hat{\sigma}_{c_{k2}'}^2+\sum_{k=1}^{Q}n_{kca}\hat{\sigma}_{t_{k0}}^2+\sum_{k=1}^{Q}r_{kca}\hat{\sigma}_{t_{k2}}^2+ \\ \sum_{k=1}^{Q}\frac{n_{kca}^2}{\lambda^2}\hat{\sigma}_{\beta_k}^2+\sum_{k=1}^{Q}\frac{r_{kca}^2}{\lambda^{*2}}\hat{\sigma}_{\beta_k^*}^2+\theta_1^{2*}\Gamma\left(\frac{1}{n_{kca}}+1\right)\hat{\sigma}_{\varphi_k}^2+\theta_2^{2*}\Gamma\left(\frac{1}{r_{kca}}+1\right)\hat{\sigma}_{\varphi_k^*}^2\end{array}} \leq C \\ \rho \geq 0;\ 2 \leq n_{kca} \leq N_k,\ 2 \leq r_{kca} \leq \hat{n}_{k2},\ n_{kca} \text{ and } r_{kca} \in \square\ \ \forall\ \ k=1,2,...,Q \end{array}\right\} \quad (47)$$

The worst deviation level is represented by the decision variable ρ. "ca" represents compromise allocation.

6 D1 -Distance Technique

A preemptive based method for solving multi-objective optimization problems by ranking the goals according to the order of importance is often called Lexicographic Goal Programming (LGP). In certain situations solutions that are considered best are chosen over better preemptive structures. D_1 distance technique chooses a proper preemptive structure which in turn provides that are better.

If we take p characteristic variances into consideration, we can rank or form a hierarchy of goals for each characteristic variance. If preference be given to variance function representing the first characteristics, we may obtain a compromise solution in the form, in some sense as $n_{kca}^{(1)} = n_{1ca}^{(1)}, n_{2ca}^{(1)}, n_{3ca}^{(1)}, ..., n_{Qca}^{(1)}$. Correspondingly, we produce results for all the lexicographic goal programming problems with priority structures and determine the optimal solution for each one. The solutions that corresponds to the second character variance will be $n_{kca}^{(2)} = n_{1ca}^{(2)}, n_{2ca}^{(2)}, n_{3ca}^{(2)}, ..., n_{Qca}^{(2)}$ if we do this continuously we will obtain $n_{kca}^{(\mathrm{R})} = n_{1ca}^{(\mathrm{R})}, n_{2ca}^{(\mathrm{R})}, n_{3ca}^{(\mathrm{R})}, ..., n_{Qca}^{(\mathrm{R})}$.

If we consider the above solution the ideal solution which corresponds to different preemptive structures are clearly stated Table 1. The solution can be described as

$$n_{kca}^{*} = \left\{ \max\left(n_{1ca}^{(1)}, n_{1ca}^{(2)}, ..., n_{1ca}^{(\mathrm{R})}\right), \max\left(n_{2ca}^{(1)}, n_{2ca}^{(2)}, ..., n_{2ca}^{(\mathrm{R})}\right), ..., \max\left(n_{Qca}^{(\mathrm{R})}, n_{Qca}^{(\mathrm{R})}\right), ..., n_{Qca}^{(\mathrm{R})} \right\} = n_{1ca}^{*}, n_{2ca}^{*}, ..., n_{Qca}^{*}$$

and

$$r_{kca}^{*} = \left\{ \max\left(r_{1ca}^{(1)}, r_{1ca}^{(2)}, ..., r_{1ca}^{(\mathrm{R})}\right), \max\left(r_{2ca}^{(1)}, r_{2ca}^{(2)}, ..., r_{2ca}^{(\mathrm{R})}\right), ..., \max\left(r_{Qca}^{(\mathrm{R})}, r_{Qca}^{(\mathrm{R})}\right), ..., r_{Qca}^{(\mathrm{R})} \right\} = r_1^{*}, r_2^{*}, ..., r_Q^{*}.$$

Ideal solutions are not obtainable in practice. The best compromise solution is chosen from the solution with shortest distance to the ideal solution. The preemptive pattern that corresponds is established as best preemptive pattern from the preemptive patterns considered.

Table 1: Computation of Ideal Solution

Preemptive Structure	n_{1ca}	n_{2ca}	$\cdots$	n_{Qca}	r_{1ca}	$\cdots$	r_{Qca}
$Z^{(1)}$	$n_{1ca}^{(1)}$	$n_{2ca}^{(1)}$	$\cdots$	$n_{Qca}^{(1)}$	$r_{1ca}^{(1)}$	$\cdots$	$r_{Qca}^{(1)}$
$Z^{(2)}$	$n_{1ca}^{(2)}$	$n_{2ca}^{(2)}$	$\cdots$	$n_{Qca}^{(2)}$	$r_{1ca}^{(2)}$	$\cdots$	$r_{Qca}^{(2)}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$		$\vdots$	$\cdots$	
$Z^{(\mathrm{R})}$	$n_{1ca}^{(\mathrm{R})}$	$n_{2ca}^{(\mathrm{R})}$	$\cdots$	$n_{Qca}^{(\mathrm{R})}$	$r_{1ca}^{(\mathrm{R})}$	$\cdots$	$r_{Qca}^{(\mathrm{R})}$
Ideal Solution	n_{1ca}^{*}	n_{2ca}^{*}	$\cdots$	n_{Qca}^{*}	r_{1ca}^{*}	$\cdots$	r_{Qca}^{*}

Table 2: Distance Calculation

	$n_{1ca}...n_{Qca}$	$r_{1ca}...r_{Qca}$	$(\mathrm{D}_1)^r$
$Z^{(1)}$	$\left\|n_{1ca}^{*}-n_{1ca}^{(1)}\right\|...\left\|n_{Qca}^{*}-n_{Qca}^{(1)}\right\|$	$\left\|r_{1ca}^{*}-r_{1ca}^{(1)}\right\|...\left\|r_{Qca}^{*}-r_{Qca}^{(1)}\right\|$	$\sum_{k=1}^{Q}\left\|\left(n_{kca}^{*}-n_{kca}^{(1)}\right)+\left(r_{1ca}^{*}-r_{1ca}^{(1)}\right)\right\|$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z^{(\mathrm{r})}$	$\left\|n_{1ca}^{*}-n_{1ca}^{(\mathrm{r})}\right\|...\left\|n_{Qca}^{*}-n_{Qca}^{(\mathrm{r})}\right\|$	$\left\|r_{1ca}^{*}-r_{1ca}^{(\mathrm{r})}\right\|...\left\|r_{Qca}^{*}-r_{Qca}^{(\mathrm{r})}\right\|$	$\sum_{k=1}^{Q}\left\|\left(n_{kca}^{*}-n_{kca}^{(\mathrm{r})}\right)+\left(r_{1ca}^{*}-r_{1ca}^{(\mathrm{r})}\right)\right\|$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z^{(\mathrm{R})}$	$\left\|n_{1ca}^{*}-n_{1ca}^{(\mathrm{R})}\right\|...\left\|n_{Qca}^{*}-n_{Qca}^{(\mathrm{R})}\right\|$	$\left\|r_{1ca}^{*}-r_{1ca}^{(\mathrm{R})}\right\|...\left\|r_{Qca}^{*}-r_{Qca}^{(\mathrm{R})}\right\|$	$\sum_{k=1}^{Q}\left\|\left(n_{kca}^{*}-n_{kca}^{(\mathrm{R})}\right)+\left(r_{1ca}^{*}-r_{1ca}^{(\mathrm{R})}\right)\right\|$

The goal programming technique defined below will be used to obtain the compromise allocation

$$
\begin{aligned}
&\min_{1\le r\le R}\sum_{h=1}^{L}(d_{kr}^{+},d_{kr}^{-})\\
&\textit{subject to}\\
&\left.\begin{aligned}
&e_1\sum_{k=1}^{Q}\left(\frac{W_k^2S_{jk}^2}{(n_{kca}-1)}+\frac{n_{kca}W_k^2W_{k2}^3S_{jk2}^2}{r_{kca}(n_{kca}W_{k2}-1)}-\frac{W_{k2}^2W_k^2S_{jk2}^2}{n_{kca}W_{k2}-1}\right)+\\
&e_2\sqrt{\sum_{k=1}^{Q}\left(\begin{array}{l}\dfrac{W_k^4\left(C_{jk}^4-\left(S_{jk}^2\right)^2\right)}{n_{kca}\left(n_{kca}-1\right)^2}+\dfrac{n_{kca}W_k^4W_{k2}^5\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{r_{kca}^2(n_{kca}W_{k2}-1)^2}\\ -\dfrac{W_k^4W_{k2}^3\left(C_{jk2}^4-\left(S_{jk2}^2\right)^2\right)}{n_{kca}(n_{kca}W_{k2}-1)^2}\end{array}\right)}+d_{kr}^{-}-d_{kr}^{+}\le Z_j^{(R)}\\
&\sum_{k=1}^{Q}\left(\bar{c}_{k0}+\bar{c}_{k1}'W_{k1}\right)n_{kca}+\sum_{k=1}^{Q}\bar{c}_{k2}'r_{kca}+\sum_{k=1}^{Q}\bar{t}_{k0}\sqrt{n_{kca}}+\sum_{k=1}^{Q}\bar{t}_{k2}\sqrt{r_{kca}}+\sum_{k=1}^{Q}\bar{\beta}_k\frac{n_{kca}}{\lambda}+\\
&\sum_{k=1}^{Q}\bar{\beta}_k^{*}\frac{r_{kca}}{\lambda^{*}}+\sum_{k=1}^{Q}\bar{\varphi}_k\theta_1^{*}\Gamma\left(\frac{1}{n_{kca}}+1\right)+\sum_{k=1}^{Q}\bar{\varphi}_k^{*}\theta_2^{*}\Gamma\left(\frac{1}{r_{kca}}+1\right)+\\
&G_\alpha\sqrt{\begin{array}{l}\sum_{k=1}^{Q}n_{kca}^2\hat{\sigma}_{c_{k0}}^2+\sum_{k=1}^{Q}n_{kca}^2W_{k1}^2\hat{\sigma}_{c_{k1}'}^2+\sum_{k=1}^{Q}r_{kca}^2\hat{\sigma}_{c_{k2}'}^2+\sum_{k=1}^{Q}n_{kca}\hat{\sigma}_{t_{k0}}^2+\sum_{k=1}^{Q}r_{kca}\hat{\sigma}_{t_{k2}}^2+\\ \sum_{k=1}^{Q}\frac{n_{kca}^2}{\lambda^2}\hat{\sigma}_{\beta_k}^2+\sum_{k=1}^{Q}\frac{r_{kca}^2}{\lambda^{*2}}\hat{\sigma}_{\beta_k^{*}}^2+\theta_1^{2*}\Gamma\left(\frac{1}{n_{kca}}+1\right)\hat{\sigma}_{\varphi_k}^2+\theta_2^{2*}\Gamma\left(\frac{1}{r_{kca}}+1\right)\hat{\sigma}_{\varphi_k^{*}}^2\end{array}}\le C\\
&\left(n_{kca}^{*}-n_{kca}^{(r)}\right)+\left(r_{1ca}^{*}-r_{1ca}^{(r)}\right)+d_{kr}^{-}-d_{kr}^{+}=0\\
&d_{kr}^{+}\ge 0,d_{kr}^{+}\ge 0,\ \ 1\ge r\ge R\\
&2\le n_{kca}\le N_k,\ 2\le r_{kca}\le\hat{n}_{k2},\ n_{kca}\text{ and }r_{kca}\ \textit{are}\text{ integers }\forall\ \ k=1,2,...,Q
\end{aligned}\right\}
\end{aligned}
\tag{48}
$$

7 Numerical Illustrations

A simulation study will be carried out to show a section of computational procedure. The R software is used to simulate data for different characteristics, see Reddy and Khan (2020), while the LINGO-20 software for optimization problem will be employed to obtained numerical solution to the integer nonlinear programming problems. LINGO is an efficient and comprehensive modeling tool built to solve linear and nonlinear optimization problems. The version used is LINGO-20 with user's guide (2017) or visit http://www.lindo.com.

Let $u_{k1}=5$ unit and $u_{k2}=5$ unit represent the reward given to the respondents and sub-sample of the non-respondent in a much more careful second attempt. Let λ and λ^{*} be the reciprocal of average time taken to obtain data from the selected units be (say 30min, 35min, 40min etc on the average for an individual). Let also θ and θ^{*} be the reciprocal of average time to failure in enumerating the selected units be (say 30min, 35min, 40min etc. per unit),

with the total fixed cost $C_0 = 6000$ units. Let the expected cost per unit time of labour and failure taken over all strata be $E(\beta_k, \beta_k^*) = \underset{\forall k}{E}(10,15)$ and $E(\varphi_k, \varphi_k^*) = \underset{\forall k}{E}(10,20)$ with $V(\beta_k, \beta_k^*) = \underset{\forall k}{V}(2.5, 4.5)$ and $V(\varphi_k, \varphi_k^*) = \underset{\forall k}{V}(2.5, 6.0)$. Let the percentage of non-response in each stratum be represented by 20, 30, 26 and 28. The objective $j = 1, 2$ is carried out in four strata, that is, $k = 1,\ 2,\ 3,\ 4$.

The R-software is used to generate data for two different characteristics. If the population size of N = 1544 is considered to be divided into four strata, then we can obtain the population means, the fourth central moments and standard deviations for all characteristics that are specified on each unit through R- simulation (Reddy and Khan 2020). Let 99% probability satisfy the chance constraint required for the problem. Then we obtain G_α is such a way that $\phi(G_\alpha) = 0.99$. At 99% confidence limit the corresponding value of the standard normal variate G_α is 2.33 from tables. The information on the simulation study is given in Table 1.

Table 1: Data obtained through R-simulation

k	N_k	W_k	S_{1k}^2	S_{2k}^2	C_{1k}^4	C_{2k}^4	W_{k1}
1	410	0.2655	97.8546	58.8913	24904.31	9399.36	0.80
2	496	0.3212	103.4125	66.1549	29691.97	13817.88	0.70
3	321	0.2079	96.2571	67.8723	29455.70	12849.46	0.74
4	317	0.2053	107.2381	59.0672	33676.30	10242.08	0.72

Table 2: Data for non-response obtained through R-simulation

k	N_k	W_k	S_{1k2}^2	S_{2k2}^2	C_{1k2}^4	C_{2k2}^4	W_{k1}	W_{k2}
1	410	0.2655	110.9813	73.6280	29587.08	12693.93	0.80	0.20
2	496	0.3212	109.9064	64.0851	29395.70	13159.13	0.70	0.30
3	321	0.2079	79.3220	52.2284	17700.00	7013.00	0.74	0.26
4	317	0.2053	134.7390	51.6128	56857.70	7177.71	0.72	0.28

Table 3: Survey cost values.

$E(c_{k0})$	$E(c'_{k1})$	$E(c'_{k2})$	$E(t_{k0})$	$E(t_{k2})$	$V(c_{k0})$	$V(c'_{k1})$	$V(c'_{k2})$	$V(t_{k0})$	$V(t_{k2})$
1.0	2.0	6.0	0.5	1.5	0.25	0.25	1.25	0.15	0.30
1.0	3.0	7.0	0.5	1.5	0.25	0.50	1.50	0.15	0.30
1.0	5.0	9.0	0.5	1.5	0.25	0.75	1.75	0.15	0.30
1.0	6.0	10.0	0.5	1.5	0.25	1.00	2.00	0.15	0.30

Table 4: Individual optimum allocation (IOA) values

	Individual Optimal at:	
	$j=1$	$j=2$
Z_1^*	0.0082347	0.5330028
Z_2^*	0.8768263	0.0064172
n_{j1}	210	167
n_{j2}	252	252
n_{j3}	47	11
n_{j4}	109	200
r_{j1}	2	3
r_{j2}	6	10
r_{j3}	2	2
r_{j4}	18	2

Table 5: Results for ideal solution

Preemptive Structure	n_{1ca}	n_{2ca}	n_{3ca}	n_{4ca}	r_{1ca}	r_{2ca}	r_{3ca}	r_{4ca}
$\left(Z_1^{(1)}, Z_2^{(2)}\right)$	187	156	57	13	13	7	3	6
$\left(Z_2^{(2)}, Z_1^{(1)}\right)$	195	163	40	54	14	6	3	15
Ideal Solution	195	163	57	54	14	7	3	15

Table 6: D_1 distance of the compromise solution

Preemptive levels	Distances of compromise solutions								Sum of Distances
	n_{1ca}	n_{2ca}	n_{3ca}	n_{4ca}	r_{1ca}	r_{2ca}	r_{3ca}	r_{4ca}	
$\left(Z_1^{(1)}, Z_2^{(2)}\right)$	0	0	17	0	0	1	0	0	18
$\left(Z_2^{(2)}, Z_1^{(1)}\right)$	8	7	0	41	1	0	0	9	66

Table 7: Compromise allocation obtained using different methods

	Fuzzy	D_1 Distance
$\hat{Z}_1$	0.06411706	0.05567108
$\hat{Z}_2$	0.13992440	0.11234679
n_{1ca}	198	195
n_{2ca}	159	165
n_{3ca}	22	57
n_{4ca}	31	54
r_{1ca}	12	14
r_{2ca}	3	7
r_{3ca}	1	3
r_{4ca}	7	16

Table 8: Trace and cost incurred.

Approach	$\hat{Z}_1$	$\hat{Z}_2$	Trace	Cost incurred
D_1 Distance	0.06411706	0.13992440	0.20404146	4733.969
Fuzzy	0.05567108	0.11234679	0.16801787	4414.410
IOA $j=1$	0.00823470	0.53300280	0.54123750	5992.618
IOA $j=2$	0.00641717	0.87682630	0.88324347	5728.483

Table 1 and Table 2, shows data obtained using the R-software (Stratified-R) for two characteristics in four strata. Table 3 shows the assumed survey cost values. Table 4 shows the optimum allocation obtained individually for $j=1$ and $j=2$ respectively. In this case, the optimum allocation of $j=1$ is used for $j=2$ and vice versa. The ideal solution of the D_1 distance technique is presented in Table 5 with distances from the evaluated ideal solution presented in Table 6. The compromise allocations using D_1 distance method and fuzzy programming is given in Table 7. The traces and cost incurred are presented in Table 8. The traces here represent the total variance with respect to all the variables in the variance-covariance matrices, which are estimates of finite population means of the characteristics considered. Since the characteristics under study are assumed to be independent the covariances are assumed to be zero. The trace value of the different approaches is given in Table 8. This paper, have implemented the Weibull cost function which approximates every existing cost function when nonresponse is considered. The gamma cost function was also

sufficiently modified. We observed closely that cost incurred did not surpass the total budgeted allocation given for the survey and also the compromise allocation technique outperformed the individual optimum allocations using the proposed Weibull cost function when nonresponse is considered.

8 Conclusion

The proposed Weibull cost function when nonresponse is observed approximates the linear, quadratic and gamma cost functions. The compromise allocation using the integer nonlinear programming problem (INLPP) with proposed Weibull cost function has solved the issue observed in stratified sampling survey when more than one variable is considered for optimum allocation. Furthermore, two compromise allocation methods were compared with the individual allocations. The compromised allocation method outperformed the individual allocation with lower trace value and less cost incurred as shown in Table 8. The Weibull cost function will fit into any convex optimization problem.

References

Abubakar, Y. and Okon, E. J. (2018). Optimum compromise allocation in multivariate stratified sampling design with gamma cost function: A case of non-response. *Journal of the Nigerian Association of Mathematical Physics*, 48, 93 - 102.

Ahmed, S., Shabbir, J., and Hafeez, W. (2016). Hansen and Hurwitz estimator with scrambled response on second call in stratified random sampling. *Journal of Statistics Applications and Probability*, 5(2), 357-369.

Alshqaq, S. S. A., Ahmadini A. A. and Ali I. (2022). Nonlinear stochastic multiobjective optimization problem in multivariate stratified sampling design. *Hindawi-Mathematical Problems in Engineering*, Volume 2022, 1-16.

Beardwood, J., Halton, J. H., and Hammersley J. M. (1959). The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55; 299–327.

Charnes, A. and Cooper, W. W. (1959) "Chance-constrained programming," *Management Science*, 6(1), 73–79.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition, John Wiley and Sons, New York.

Ding X. and Wang C. (2011). A novel algorithm of stochastic chance-constrained linear programming and its application. *Hindawi-Mathematical Problems in Engineering*, Volume 2012, 1-17.

El-Badry, M. A. (1956). A sampling procedure for mailed questionnaires. *Journal of the American Statistical Association*, 51, 209-227.

Foradori, G. T. (1961). *Some non-response sampling theory for two stage designs.* Mimeograph Series No. 297, North Carolina State University, Raleigh.

Hansen, M. H., and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of American Statistical Association*, 41, 517–529.

Haseen, S., Ali, I., and Bari, A. (2014). multi-objective stochastic multivariate stratified sampling in presence of non-response. *Communications in Statistics-Simulation and Computation*, 45(8), 2810-2826.

Karuna G. Reddy and Khan M.G.M. (2018). optimal stratification in stratified designs using weibull-distributed auxiliary information. *Communications in Statistics - Theory and Methods*, 1–17.

Khan, M. G. M., Khan, E. A., and Ahsan, M. J. (2008). optimum allocation in multivariate stratified sampling in presence of non-response. *Journal of the Indian Society of Agricultural Statistics*, 62(1), 42-48.

LINGO-User's Guide (2017). LINGO-User's Guide. LINDO SYSTEM INC., USA.

Mahfouz, M. I., Rashwan M. M. and Khadr Zeinab A. (2023). Optimal stochastic allocation in multivariate stratified sampling. *Mathematics and Statistics*, 11(4), 676 – 684.

Melaku A. (1986). Asymptotic normality of the optimal allocation in multivariate stratified random sampling. *Indian Journal of Statistics. Series B*, 48. 224–232.

Raghav, Y. S., Ali, I., and Bari, A. (2014). Multi-Objective Nonlinear Programming Problem Approach in Multivariate Stratified Sample Surveys in Case of Non-Response. *Journal of Statistical Computation and Simulation*, 84(1); 22-36.

Reddy K. G. and Khan, M. G. M. (2020) "stratifyR: an R package for optimal stratification and sample allocation for univariate populations," *Australian & New Zealand Journal of Statistics*, 62(3), 383–405.

Rao S. (2009) *Engineering optimization theory and practice*. Wiley, Hoboken, New Jersey

Ronald, E. W., Raymond, H. M., Sharon, L. M. and Keying, Y. (2007). *Probability and statistics for engineers and scientist*. 8th Edition. Upper Saddle River, New Jersey: Pearson Education, Inc.

Shafiullah, Haseen S., and Ali, I. (2015). Geometric programming approach in multivariate stratified sample survey in case of non-response. *Electronic Journal of Applied statistical Analysis*, 8(1), 28-43.

Shafiullah, Khan, M. F., and Ali, I. (2014). Fuzzy geometric programming in multivariate stratified sample surveys in presence of non-response with quadratic cost function. *American Journal of Operations Research*, 4, 173-188.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling theory of surveys with applications*. Iowa State University Press, Ames and Indian Society of Agricultural Statistics, New Delhi

Varshney R, and Mradula (2019). Optimum allocation in multivariate stratified sampling design in the presence of nonresponse with gamma cost function. *Journal Statistical Computing and Simulation*, 89(13), 2454–2467.

Varshney, R., Ahsan, M. J. and Khan, M. G. M. (2011). An optimum multivariate stratified sampling design with non-response: a lexicographic goal programming approach. *Journal of Mathematical Modeling and Algorithms*, 65, 291-296.

Varshney, R., Najmussehar, and Ahsan, M. J. (2012). estimation of more than one parameters in stratified sampling with fixed budget. *Mathematical Method and Operations Research*, 75(2), 185–197.

Statistical analysis of current risk factors associated with stroke types across gender and age: a retrospective hospital-based study in Southern Nigeria

Augustine Iduseri* and Precious Opoggen

Department of Statistics, University of Benin, P.M.B. 1154, Benin City, Nigeria

**Corresponding Author; E-mail:augustine.iduseri@uniben.edu*

Abstract

Stroke is the second leading cause of death worldwide, and remains one of the main causes of death in Nigeria with its economic impact on the rise. Over the years, hospital-based studies in Nigeria often only report the traditional risk factors, with high blood pressure as the leading risk factor for cardiovascular disease, particularly stroke, in the country. In this study, we will statistically assess all probable traditional and infectious risk factors associated with stroke. Our aim is to identify the current significant risk factors associated with the prevalent type of stroke in Nigeria. This was a retrospective study conducted at a tertiary referral hospital, Federal Medical Centre (FMC), in Asaba, Delta State, Nigeria. We conducted a retrospective audit of medical records from 2019 to 2022 for adult stroke patients admitted to the medical ward. Data extracted included demographic characteristics, clinical variables, and relevant data. Identified risk factors associated with stroke types across gender and age were investigated using multinomial logistic regression. From the result, out of the recorded 907 stroke patients, 457 (50.39%) were female, while 450 (49.61%) were males. The age group (61-80) had more recorded cases totaling 426(47%).Hypertension, diabetes mellitus, heart disease, malaria, and pneumonia were the significant risk factors associated with the prevalent stroke type. Therefore, there is a need to step up measures aimed at increasing public awareness of emerging infectious risk factors associated with stroke and effective preventive measures. Prioritizing interventions directed towards cost-effective treatment of hospitalized stroke patients is imperative as this will hopefully improve the overall outcomes in resource-constrained settings such as Nigeria.

Keywords: Stroke types, traditional risk factors, infectious risk factors, multinomial logistics regression

1. Introduction

A stroke is a sudden condition that leads to multiple physical and cognitive impacts, such as difficulty in communication, memory loss, challenges with mobility, depression, and sometimes being bedridden. It is an acute, focal dysfunction of the brain, originating from vessels and lasting 24 hours or longer. (Saengsuwan and Suangpho, 2019). According to the classification of the American Heart Association report, the two main types of strokes are ischemic and hemorrhagic stroke. Ischemic stroke, the most common type, occurs from

reduced or obstructed blood supply to the brain tissues, whereas hemorrhagic stroke occurs due to the rupture of weakened blood vessels within the brain (Rink and Khanna, 2011). Stroke remains a leading cause of disability, dementia, and the second leading cause of death globally and is associated with substantial economic costsfor the sufferers, andfamilies (Onwuakagba et al., 2023; Teh et al., 2018).Surprisingly, the consequences on patients and their families are often unexplainable and can even be referred to as tragic because of the enormous psychological, social, and financial burden associated with stroke.

In Africa, the annual incidence rate of stroke is about 316 cases per 100,000 person-years, while the prevalence rate is about 1,460 cases per 100,000 person-years (Akinyemi et al., 2021). Also, Africa has a 3-year fatality rate greater than 80%, with stroke-related deaths being about 5.5% to 11% of mortality (Akinyemi et al., 2021). Current evidence indicates that Africa could have up to 2-3-fold greater rates of stroke incidence and higher stroke prevalence than Western Europe and the USA (Tamer et al., 2020).The increase in stroke burden in Africa is being driven mostly by increasing trends in the prevalence of modifiable risk factors (including hypertension, smoking, dyslipidemia, hyperlipidemia, poor diet, abdominal obesity, diabetes, physical inactivity, atrial fibrillation, high salt intake),and also regulated by non-modifiable risk factors (such as age, sex, race, geographic locations, and heredity)(Onwuakagba et al., 2023). Nigeria, similar to many other low-income and middle-income countries in Sub-Saharan Africa, has been experiencing a significant increase in stroke burden with an estimated yearly incidence rate of 1.14 per 1000 persons/year (Moghtaderi and Alavi-Naini, 2012).The increasing number of cardiovascular diseases (CVD), especially stroke, is partly attributed to the growing presence of both modifiable and non-modifiable risk factors. Additionally, urbanization and the adoption of Western lifestyles, characterized by sedentary lifestyles, harmful habits like tobacco and alcohol consumption, and high fat/cholesterol diets, also contribute to this rise. (Omigbile et al., 2023;Ike and Onyema, 2020).

A few international studies and presentations have linked stroke to infectious risk factors like malaria and pneumonia(Leopoldino et al., 1999; Watila et al., 2019; Grossmann et al., 2021; Periyasamy et al., 2023).Therefore, our work is motivated by these studies and presentations. Hence, we will statistically assess the traditional and infectious risk factors associated with stroke.Our aim is to identify the current significant risk factors associated with the prevalent type of stroke in Nigeria.The specific objectives of this study are: (1) To evaluate the association between different age groups and various types of strokes. (2) To evaluate the

association of sex and various types of strokes. (3) To statistically ascertain the current traditional and infectious risk factors associated with the types of strokes in Nigeria.The rest of the article is organized as follows: First, we review the existing literature on traditional, infectious, and environmental risk factors of stroke. This is followed by a description of the research materials and methods used in the study. Next, we discuss the results of our inquiry. Finally, we present our concluding remarks.

2. Literature Review

Several hospital-based studies in Nigeria have documented that cardiovascular disease, particularly stroke, is the leading cause of neurological admissions and medical coma (Eze and Kalu, 2014; Ekenze et al., 2010;Kayode-Iyasere et al., 2019; Ojini and Danesi, 2003; Ogun et al., 2005; Talabi, 2003; Arabambi et al., 2021; Osuegbu et al., 2021; Nwazor et al., 2024). The common risk factors identified by these hospital-based studies in Nigeria are high blood pressure, diabetes, tobacco use, unhealthy diet, physical inactivity, and aging.While traditional cardiovascular risk factors are responsible for most strokes in Nigeria, Sub-Saharan Africa and other parts of the world, infectious pathogens have also been reported to increase the risk, and in some cases, have a direct causal role (Moghtaderi and Alavi-Naini, 2012; Jennifer and Fugate, 2020; Hameed et al., 2024).Important examples of infectious pathogensor infectious diseases of stroke (i.e. parasitic, bacterial, viral, or fungal infections) that can lead to stroke includecerebral malaria (Carod-Artal, 2007;Leopoldino et al., 1999;Periyasamy et al., 2023), tuberculous meningitis (Wasay et al., 2018), pneumonia (Ingeman et al., 2011;Kim et al., 2017; Grossmann et al., 2021; Watilaet el., 2019; Olajide et al., 2021),HIV/AIDS (Grau et al., 2010; Thakur et al., 2016),cysticercosis(Alarcón et al., 1992a; Alarcón et al., 1992b), syphilis (Jennifer and Fugate,2020), chronic hepatitis infection (Cojocaru et al., 2005; Sung et al., 2007),Chagas disease(Lage et al., 2022),and most recently, coronavirus disease 2019 (COVID-19) (Khan et al., 2023). Besides the infectious risk factors, environmental risk factors are also emerging as non-traditional risk factors associated with stroke (Hameed et al., 2024). Important examples include air pollution (Wellenius et al., 2012; Feigin et al., 2016), climate change (Chu et al., 2018; Atwoli et al., 2021), and high altitude (Jaillard et al., 1995; Lu et al., 2020; Liu et al., 2021). Air pollution which is more pronounced in developing countries is a known risk factor for cardiovascular disease and reportedly accounts for 14% of all stroke-associated mortality (Verhoeven et al., 2021).

Nigeria, like many other Sub-Saharan African (SSA) countries, is going through an epidemiological transition. This transition is characterized by increasing urbanization and lifestyle changes, resulting in a higher prevalence of cardiovascular diseases, as well as infectious and environmental risk factors associated with stroke.However, infectious risk factors or tropical diseases that are often not among the top risk factors in developed nations, such as malaria, pneumonia, and HIV/AIDS are endemic in Nigeria. In 2021, there were an estimated 68 million cases and 194,000 deaths from malaria in Nigeria (WHO, 2022).Some case presentations on stroke have indicated a link with the most severe complication of malaria caused by Plasmodium falciparum, also known as cerebral malaria (CM), which could be responsible for up to 10% of strokes in endemic regions like Nigeria(Carod-Artal, 2007;Leopoldino et al., 1999; Periyasamy et al., 2023).Also, a serious and common complication of stroke is pneumonia.Warren-Gash et al. (2018) in their study found that individuals with pneumonia could have an increased risk of heart attack for up to one week after infection, and stroke for up to 28 days.In Nigeria, community-acquired pneumonia (CAP) accounts for 2.5% to 5.7% of medical admissions and between 15.3% and 24.9% of respiratory admissions in tertiary health facilities. Previous surveyshave shown that mortality rates among admitted patients with CAP range from 7.4% to 26% in the country.

Although there are a lot of hospital-based studies confirming that cardiovascular disease, particularly stroke, is the leading cause of neurological admissions and medical comain Nigeria (Eze and Kalu, 2014; Ekenze et al., 2010; Kayode-Iyasere et al., 2019; Ojini and Danesi, 2003; Ogun et al., 2005; Talabi, 2003; Arabambi et al., 2021; Osuegbu et al., 2021; Nwazor et al., 2024); however, these studies often only report the traditional risk factors, with high blood pressure as the leading risk factor for stroke in Nigeria.Against this background, this study employedmultinomial logistic regression to ascertain if there are significant relationships between any infectious risk factors with stroke types in Southern Nigeria; relying on the evidence from retrospective hospital data.

3. Materials and Methods

Study Design and Data Sources

This was a hospital-based retrospective study conducted at theFederal Medical Centre (FMC) Asaba, Delta State.Asaba town is the capital of Delta State, Southern Nigerialocated at the western bank of the Niger River. Asaba with a population of 149,603 as at the 2006 census is close to Onitsha, Owerri, and Enugu which are major cities in Southeastern Nigeria. This makes

FMC, Asaba a referral center for major cities in both South-South and South-East regions of Nigeria. The study population included patients presenting with acute stroke at FMC, Asaba between August 2019 and September 2022. Case files of 907 stroke patients admitted during the study were retrieved, and relevant information was extracted using a structured questionnaire. The questionnaire included relevant information such as a unique case identifier, age, sex, duration to presentation, duration to outcome, length of hospitalization, clinical indications (stroke type), risk factors (both vascular and infectious risk factors), and access to neuroimaging.Eight categories of stroke type were identified as reported in their admission records. These include Transient Ischemic Stroke (TIS), Ischemic Stroke (IS), Acute Stroke (AS), Hemorrhagic Stroke (HS), Left Hemispheric Stroke (LHS), Left Hemispheric Ischemic Stroke (LHIS), Right Hemispheric Stroke (RHS), and Right Hemispheric Ischemic Stroke (RHIS).

Ethical clearance for the study was obtained from the Research and Ethical Committee, Federal Medical Centre, Asaba. The ethical reference number was FMC/ASB/A81 VOL XII/268. Permission to collect data from patient's medical records was also obtained from the Department of Medical Records.During the research, we ensured anonymity and confidentiality were upheld to the highest standard.

Statistical Analysis

The obtained data wasimputed into IBM Statistical Package for the Social Sciences (SPSS) Statistics Version 26 Data Editor for analysis.Before starting the analysis, we checked for missing values and used the SPSS multiple imputation technique to handle them.SPSS descriptive statistics;frequencies (percentages) for categorical variables were used to describe the data distributionsof some sociodemographic characteristics and other variables of the sample data for presentation. We then applied SPSS multinomial logistic regression to assess the association between different age groups and sex with the types of stroke, and to identify the current significant risk factors associated with the types of stroke.Because we are analyzing a set of categorical variables, and one of them is a "response" while the others are predictors, we adopted the multinomial logistic regression to achieve our objectives. The k-th logistic regression model for the variables applicable in this study is given by

$$\ln\left[\frac{P\left(Y_i = k \mid X_{1,i},\ldots,X_{P,i}\right)}{P\left(Y_i = K \mid X_{1,i},\ldots,X_{P,i}\right)}\right] = \beta_{0,k} + \beta_{1,k}x_1 + \beta_{2,k}x_2 + \cdots + \beta_{P-1,k}x_{P-1} = \beta_k x_i$$

(3.1)

Where Y_i is the value of the multinomial response variable for the ith unit, $X_{1,i},\cdots,X_{P,i}$ are the predictor variables, $\beta_0,\beta_1,\cdots,\beta_{P-1}$ are the coefficients of the predictor variables, and *kth* category is the reference category. The probability of having success in the *k*-th category is given in Equation 2

$$p_{i,k} = P\left(Y_i = k \mid X_{1,i},\cdots,X_{P,i}\right) = \frac{\exp\left(\beta_k x_i\right)}{1+\sum_{k-1}^{K-1}\exp\left(\beta_k x_i\right)} \qquad (3.2)$$

4. Results and Discussion

4.1. *Differences in stroke type prevalence across patients' sex*

The results in Table 1 showed that there are 457(50.39%) female patients with stroke, and 450 (49.61%) male patients with stroke. Our observed trend of higher stroke prevalence among women aligns with a previous study conducted at FMC, Asaba, Delta State, Nigeria (Ezunu et al., 2023).Upon reviewing Table 1, it is evident that ischemic stroke was more common in females (46.6%), followed by hemorrhagic stroke (11.6%), and right hemispheric stroke (11.2%). Ischemic stroke (55.1%) was also the most common for males, followed by right hemispheric stroke (12.2%).In the present study, it was found that 461 (50.83%) patients had ischemic stroke, the highest incidence among the patients. This finding is consistent with several other studies conducted in different African countries. The similarity in the genetic makeup of the different race/ethnicity may account for this observation. However, only 39 (4.30%) of the patients had acute stroke, which is the lowest type of stroke recorded in this study.

Table 1: Stroke types across sex and their proportions

Stroke Types	No. of Patients	Female	Males
Transient ischemic stroke (TIS)	46 (5.07%)	30 (6.6%)	16 (3.6%)
Ischemic stroke (IS) (55.1%)	461 (50.83%)	213 (46.6%)	248
Acute stroke (AS)	39 (4.30%)	11 (2.4%)	28 (6.2%)
Hemorrhagic stroke (HS)	84 (9.26%)	53 (11.6%)	31 (6.9%)
Left hemispheric stroke (LHS)	47 (5.18%)	28 (6.1%)	19 (4.2%)
Left hemispheric ischemic stroke (LHIS)	48 (5.29%)	34 (7.4%)	14 (3.1%)
Right hemispheric stroke (RHS)	106 (11.69%)	51 (11.2%)	55 (12.2%)
Left hemispheric ischemic stroke (LHIS)	76 (8.38%)	37 (8.1%)	39 (8.7%)
		457(50.39%)	
		450(49.61%)	

This shows that ischemic stroke was more likely to cause death in the studied population. This finding was consistent with a previous study conducted in Nigeria and the rest of the world(Jowi and Mativo, 2008; Zhanget al., 2017; Mulugeta et al., 2020; Arabambi et al., 2021;).On the other hand, in other studies in Nigeria and the rest of the world, hemorrhagic stroke was the most common subtype (Gedefa et al., 2017; Zewdie et al., 2018; Maskey et al., 2011; Osuegbu et al., 2021). This difference may be attributed to underlying differences in risk factors (Zhang et al., 2017).

4.2. Differences in stroke type prevalence across patients' different age groups

Table 2 shows the distribution of stroke patients by age group. There are 19 patients with stroke in the age group of 20 or younger, 45 patients in the age group 21-40, 286 patients in the age group 41-60, 426 patients in the age group 61-80, and 131 in the age group 81-100. This data indicates that older age groups have a higher prevalence of stroke. The research conducted by Teh et al. in 2018 and Osuegbu et al. in 2021 indicates that as individuals get older, they are more likely to experience a stroke. This could be attributed to the physiological changes that

occur in the elderly, as well as the greater complexity and severity of the condition, which tend to increase with age.

Table 2: Stroke types across different age group and their proportions

Stroke Types	**≤ 20**	**21-40**	**41-60**	**61-80**	**81-100**
TIS	0 (0.0%)	2 (4.4%)	12 (4.2%)	25 (5.9%)	7 (5.3%)
IS	7 (36.8%)	19 (42.2%)	143(50.0%)	241 (56.6%)	51 (38.9%)
AS	2 (10.5%)	0 (0.0%)	1 (0.3%)	20 (4.7%)	16 (12.2%)
HS	6 (31.6%)	6 (13.3%)	40 (14.0%)	23 (5.4%)	9 (6.9%)
LHS	0 (0.0%)	5 (11.1%)	20 (7.0%)	18 (4.2%)	4 (3.1%)
LHIS	2 (10.5%)	2 (4.4%)	14 (4.9%)	20 (4.7%)	10 (7.6%)
RHS	0 (0.0%)	4 (8.9%)	34 (11.9%)	40 (9.4%)	28 (21.4%)
RHIS	2 (10.5%)	7 (15.6%)	22 (7.7%)	39 (9.2%)	6 (4.6%)
	19 (2.09%)	45 (4.96%)	286 (31.54%)	426 (46.97)	131 (14.44%)

Consequently, older individuals may require more inpatient health services due to these factors.A cursory look at Table 2 shows that the distribution of stroke types in the age 61-80 is relatively higher in terms of incidence than in any other age group. This means thatolder adults dominated the stroke incidence and were more likely to die from a stroke irrespective of the type of stroke. However, the prevalence of ischemic stroke was higher across the age groups. This again shows that ischemic stroke was more likely to cause death in the studied population.

4.3. Predominant risk factors across patients' sex

Table 3 presents the distribution of the predominant risk factors among the studied population. Out of all the potential risk factors listed in the structured questionnaire used for data collection, including cardiovascular diseases and infectious risk factors, only the six risk factors shown in Table 3 were predominantly documented as clinical diagnoses and/or laboratory diagnoses in the medical records of stroke patients.

Table 3: Predominant risk factors across sex and their proportions

Risk Factors	No. of Patients	Male	Females
Hypertension	81	34(41.98%)	47 (58.02%)
Diabetes Mellitus	56	16 (28.75%)	40 (71.43%)
Malaria	22	14 (63.64%)	8 (36.6%)
Pneumonia	20	9 (45.00%)	11 (55.00%)
Heart Disease	24	5 (20.83%)	19 (79.17%)
Chronic Anemia	5	5 (100.00%)	0 (0.00%)

While other risk factors listed in the structured questionnaire had only one or two recorded cases in the medical records of the studied population.

In Table 3, hypertension and diabetes mellitus, were found to have the highest recorded risk factors for stroke patients in the studied population, which is consistent with previous studies(Mulugeta et al., 2020; Ojo et al., 2024). The results in Table 3 also indicate that heart disease, malaria, and pneumonia were the second-highest recorded infectious risk factors. Malaria was observed more frequently in males (63.64%) while pneumonia was more commonly observed in females (55.00%).Previous studies have linked severe complications of malaria caused by plasmodium falciparum and pneumonia as possible infectious risk factors associated with stroke (Carod-Artal, 2007; Leopoldino et al., 1999; Warren-Gash et al., 2018; Periyasamy et al., 2023).

To determine if the predominant risk factors listed in Table 3 are significantly associated with different types of stroke, we used multinomial logistic regression. In this analysis, the risk factors were used as the multinomial response variable, while the stroke types were used as the predictor variables.

Table 4: Multinomial logistic regression model for stroke types and risk factors

		Parameter Estimates					
Risk factor[a]		**B**	**S.E**	**Wald**	**df**	**Sig**	**Exp(B) [95% C.I]**
Hypertension	Intercept	19.735	0.880	503.013	1	0.000	
	TIS	-18.348	1.423	166.314	1	0.000	1.075E-8 [6.613E-10; 1.748E-7]
	IS	-17.687	0.746	562.131	1	0.000	2.083E-8[4.827E-9 ; 8.988E-8]
Diabetes mellitus	Intercept	19.378	0.904	459.557	1	0.000	
	TIS	-18.279	1.466	155.379	1	0.000	1.152E-8 [6.504E-10; 2.040E-7]
	IS	-17.720	0.784	510.782	1	0.000	2.016E-8 [4.336E-9; 9.372E-8]
Malaria	Intercept	18.818	0.961	383.155	1	0.000	
	TIS	-18.818	1.710	121.101	1	0.000	6.719E-9 [2.354E-10; 1.918E-7]
	IS	-17.566	0.864	413.786	1	0.000	2.352E-8 [4.329E-9 ; 1.278E-7]
Pneumonia	Intercept	18.818	0.961	383.155	1	0.000	
	IS	-18.007	0.886	412.891	1	0.000	1.512E-8 [2.662E-9; 8.587E-8]
Heart disease	Intercept	18.531	0.584	1007.259	1	0.000	
	TIS	-16.921	1.241	185.817	1	0.000	4.479E-8 [3.932E-9 ; 5.103E-7]

a. The reference category is: chronic anemia

Table 4 displays the SPSS output of the multinomial logistic regression analysis for only the stroke types that have a significant relationship with the identified predominant risk factors.

A cursory look at Table 4 reveals that transient ischemic stroke (TIS) and ischemic stroke (IS) have a significant relationship with hypertension (p=0.000), diabetes mellitus (p=0.000), and malaria (p=0.000) respectively. Table 4 also reveals that IS has a significant relationship with pneumonia (p=0.000), and TIS has a significant relationship with heart disease (p=0.000). This result not onlyconfirms our earlier assertion on Table 3 results, but further reveals that all the identified predominant risk factors have a significant relationship with ischemic stroke. This suggest that among all stroke types, ischemic stroke (whether TIS or IS) has the highest prevalence in this study which is consistent with previous studies(Mulugeta et al., 2020; Osuegbu et al., 2021). In addition to the commonest risk factors such as hypertension, diabetes mellitus, and heart disease identified in this study, the significant relationship of malaria and pneumonia with ischemic stroke in Table 4 suggests the two infectious risk factors as a significant emerging risk factors for stroke in Nigeria.This finding is corroborated by previous studies and presentation outside Nigeria that have linked severe complications of malaria caused by plasmodium falciparum and pneumonia as possible infectious risk factors associated

with stroke, and others that even suggested they could have a direct causal role (Carod-Artal, 2007;Leopoldino et al., 1999; Warren-Gash et al., 2018; Periyasamy et al., 2023).

5. Conclusion

This study found that ischemic stroke was a common indication for medical admissions, in particular for age group 61-80,with hypertension and diabetes mellitus as a leading risk factor, followed by heart disease, malaria, and pneumonia. Therefore, there is a need to step up measures aimed at increasing public awareness of emerging infectious risk factors(such as malaria and pneumonia) associated with stroke and providing effective preventive measures. Prioritizing interventions directed towards cost-effective treatment of hospitalized stroke patients is imperative as this will hopefully improve the overall outcomes in resource-constrained settings such as Nigeria.

This study, being a retrospective review of records, had some limitations. Many of the case files had substantial missing data and could not be included in the data analysis. This low yield of data may distort the accurate picture of stroke presentations at the hospital. Also, the measured risk factors were likely underreported as they were absent in some of the case files reviewed. A multi-centered prospective and retrospective, registry-based study that will provide more reliable data regarding stroke types with current associated risk factors in Nigeria will in doubt help establish this study results.

References

Akinyemi, R. O., Ovbiagele, B., Adeniji, O. A., et al. (2021). Stroke in Africa: Profile, progress, prospects, and priorities. *Nature Reviews Neurology, 17*(11), 634–656.

Alarcón, F., Hidalgo, F., Moncayo, J., & Viñán, I. (1992a). Cerebral cysticercosis and stroke.Stroke, 23(2), 224–228.

Alarcón, F., Vanormelingen, K., Moncayo, J., & Viñán, I. (1992b). Cerebral cysticercosis as a risk factor for stroke in young and middle-aged people. *Stroke, 23*(11), 1563–1565.

Arabambi, B., Oshinaike, O., Akilo, O. O., Yusuf, Y., & Ogun, S. A. (2021). Pattern, risk factors, and outcome of acute stroke in a Nigerian university teaching hospital: A 1-year review. *Nigerian Journal of Medicine, 30*(3), 252–258.

Atwoli, L., Baqui, A. H., Benfield, T., Bosurgi, R., Godlee, F., Hancocks, S., Horton, R. C., Laybourn-Langton, L., Monteiro, C. A., Norman, I., et al. (2021). Call for emergency action to limit global temperature increases, restore biodiversity, and protect health. *BMJ Nutrition, Prevention & Health, 4*(1), 362–364.

Carod-Artal, F. J. (2007). Strokes caused by infection in the tropics. *Revista de Neurología, 44*, 755–763.

Chu, S. Y., Cox, M., Fonarow, G. C., Smith, E. E., Schwamm, L., Bhatt, D. L., Matsouaka, R. A., Xian, Y., & Sheth, K. N. (2018). Temperature and precipitation associate with ischemic stroke outcomes in the United States. *Journal of the American Heart Association, 7*(13), e010020.

Cojocaru, I. M., Cojocaru, M., Iacob, S. A., & Burcin, C. (2005). Cerebral ischemic attack secondary to hepatitis C virus infection. *Romanian Journal of Internal Medicine, 43*, 255–260.

Ekenze, O., Onwuekwe, I., & Ezeala, B. (2010). Profile of neurological admissions at the University of Nigeria Teaching Hospital, Enugu. *Nigerian Journal of Medicine, 19*(4), 419–422.

Eze, C. O., & Kalu, U. A. (2014). Pattern of neurological admissions in the tropics: Experience at Abakaliki, South-Eastern Nigeria. *Nigerian Journal of Medicine, 23*(4), 302–305.

Ezunu, O. E., Akpekpe, J., Yusuf, Y., Ogunniyi, A., Ige, F., Ezunu, N. E., Ogbutor, U. G., & Guobadia, E. K. (2023). Pattern and distribution of neurological disorders among patients that attended Federal Medical Centre, Asaba, Nigeria between 2015–2019. *International Journal of Research and Innovation in Social Science (IJRISS)*, 7(6).

Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., Mensah, G. A., Norrving, B., Shiue, I., Ng, M., et al. (2016). Global burden of stroke and risk factors in 188 countries, during 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet Neurology, 15*(9), 913–924.

Gedefa, B., Menna, T., Berhe, T., & Abera, H. (2017). Assessment of risk factors and treatment outcome of stroke admissions at St. Paul's Teaching Hospital, Addis Ababa, Ethiopia. *Journal of Neurology and Neurophysiology, 8*(3), 1–6.

Grau, A. J., Urbanek, C., & Palm, F. (2010). Common infections and the risk of stroke. *Nature Reviews Neurology, 6*(12), 681–694.

Grossmann, I., Rodriguez, K., Soni, M., Joshi, P. K., Patel, S. C., Shreya, D., Zamora, D. I., Patel, G. S., & Sange, I. (2021). Stroke and pneumonia: Mechanisms, risk factors, management, and prevention. *Cureus, 13*(11), e19912.

Hameed, S., Karim, N., Wasay, M., & Venketasubramaniam, N. (2024). Emerging stroke risk factors: A focus on infectious and environmental determinants. *Journal of Cardiovascular Development and Disease, 11*(1), 19.

Ike, S., & Onyema, C. (2020). Cardiovascular diseases in Nigeria: What has happened in the past 20 years? *Nigerian Journal of Cardiology, 17*, 21–26.

Ingeman, A., Andersen, G., Hundborg, H. H., Svendsen, M. L., & Johnsen, S. P. (2011). In-hospital medical complications, length of stay, and mortality among stroke unit patients. *Stroke, 42*(11), 3214–3218.

Jaillard, A. S., Hommel, M., & Mazetti, P. (1995). Prevalence of stroke at high altitude (3380 m) in Cuzco, a town of Peru: A population-based study. *Stroke, 26*(4), 562–568.

Jennifer, E., & Fugate, D. O. (2020). Infectious causes of stroke. *Practical Neurology.* https://practicalneurology.com/articles/2020-jan/infectious-causes-of-stroke

Jowi, J., & Mativo, P. (2008). Pathological sub-types, risk factors, and outcome of stroke at the Nairobi Hospital, Kenya. *East African Medical Journal, 85*(12), 572–581.

Kayode-Iyasere, E. O., Obasohan, A. O., & Odiase, F. E. (2019). Medical coma in a secondary health centre in Benin City, Nigeria: A 3-year review. *Port Harcourt Medical Journal, 13*(2), 58–62.

Khan, M., Hameed, S., Soomro, B. A., Mairaj, S., Malik, A., Farooq, S., Rukn, S. A., & Wasay, M. (2023). COVID-19 independently predicts poor outcomes in acute ischemic stroke: Insights from a multicenter study from Pakistan and United Arab Emirates. *Journal of Stroke and Cerebrovascular Diseases, 32*, 106903.

Kim, B. R., Lee, J., Sohn, M. K., Kim, D. Y., Lee, S. G., Shin, Y. I., Oh, G. J., Lee, Y. S., Joo, M. C., Han, E. Y., & Kim, Y. H. (2017). Risk factors and functional impact of medical complications in stroke. *Annals of Rehabilitation Medicine, 41*(5), 753–760.

Lage, T. A. R., Tupinambás, J. T., Pádua, L. B., Ferreira, M. O., Ferreira, A. C., Teixeira, A. L., & Nunes, M. C. P. (2022). Stroke in Chagas disease: From pathophysiology to clinical practice. *Revista da Sociedade Brasileira de Medicina Tropical, 55*, e0575.

Leopoldino, J. F. S., Fukujima, M. M., & Gabbai, A. A. (1999). Malaria and stroke: Case report. *Arquivos de Neuro-Psiquiatria, 57*(4), 1024–1026.

Liu, M., Yan, M., Guo, Y., Xie, Z., Li, R., Li, J., Ren, C., Ji, X., & Guo, X. (2021). Acute ischemic stroke at high altitudes in China: Early onset and severe manifestations. *Cells, 10*, 809.

Lu, Y., Zhuoga, C., Jin, H., Zhu, F., Zhao, Y., Ding, Z., He, S., Du, A., Xu, J., Luo, J., et al. (2020). Characteristics of acute ischemic stroke in hospitalized patients in Tibet: A retrospective comparative study. *BMC Neurology*, *20*, 380.

Maskey, A., Parajuli, M., & Kohli, S. C. (2011). A study of risk factors of stroke in patients admitted in Manipal Teaching Hospital, Pokhara. *Kathmandu University Medical Journal (KUMJ), 9*(36), 244–247.

Moghtaderi, A., & Alavi-Naini, R. (2012). Infective causes of stroke in tropical regions. *Iranian Journal of Medical Sciences, 37*(3), 150–158.

Mulugeta, H., Yehuala, A., Haile, D., Mekonnen, N., Dessie, G., Kassa, G. M., Kassa, Z. S., & Habtewold, T. D. (2020). Magnitude, risk factors and outcomes of stroke at Debre Markos Referral Hospital, Northwest Ethiopia: A retrospective observational study. *Egyptian Journal of Neurology, Psychiatry and Neurosurgery, 56*, 41.

Nwazor, E., Obi, P., Madueke, O., Ajuonuma, B., & Chukwuocha, I. (2024). A 2-year review of stroke admissions and short-term outcome predictors in a teaching hospital, Southeast, Nigeria. *Nigerian Medical Journal, 65*(2), 185–194.

Ogun, S., Ojini, F., Ogungbo, B., Kolapo, K., & Danesi, M. (2005). Stroke in South West Nigeria: A 10-year review. *Stroke, 36*, 1120–1122.

Ojini, F., & Danesi, M. (2003). The pattern of neurological admissions at the Lagos University Teaching Hospital. *Nigerian Journal of Clinical Practice, 6*, 38–41.

Ojo, A. E., Ojji, D. B., Grobbee, D. E., Huffman, M. D., & Peters, S. A. E. (2024). The burden of cardiovascular disease attributable to hypertension in Nigeria: A modelling study using summary-level data. *Global Heart, 19*(1), 50.

Olajide, O. O., Adebayo, P. B., Taiwo, F. T., Owolabi, M. O., & Ogunniyi, A. (2021). The contribution of dysphagia to stroke morbidity and mortality among Nigerians. *Annals of Health Research, 7*(2), 118–131.

Omigbile, O., Oni, A., Abe, E., Lawal, E., Oyasope, B., & Mayaki, T. (2023). Prevalence and risk factors of cardiovascular diseases among the Nigerian population: A new trend among adolescents and youths. In *Novel pathogenesis and treatments for cardiovascular disease*. IntechOpen.

Onwuakagba, I. U., Okoye, E. C., Kanu, F. C., Kalu, C. M., Akaeme, D. C., Obaji, O. C., & Akosile, C. O. (2023). Population-based stroke risk profile from a West-African community. *eNeurologicalSci, 33*, 100483.

Osuegbu, O. I., Adeniji, F. O., Owhonda, G. C., Kanee, R. B., Alblihed, M., Batiha, G. E.-S., & Aigbogun, E. O. (2021). Non-modifiable risk factors (age and sex), stroke types, and outcomes in Rivers State, Nigeria: A retrospective hospital-based study. *Preprints*, 2021070446.

Periyasamy, M., Bharathi, D. P., & Vinatha, Devi, M. K. (2023). Typical presentation of malaria with stroke. *Journal of Research in Medical and Dental Science, 11*(1), 299–300.

Rink, C., & Khanna, S. (2011). Significance of brain tissue oxygenation and the arachidonic acid cascade in stroke. *Antioxidants & Redox Signaling, 14*(10), 1889–1903.

Saengsuwan, J., & Suangpho, P. (2019). Self-perceived and actual risk of further stroke in patients with recurrent stroke or recurrent transient ischemic attack in Thailand. *Journal of Stroke and Cerebrovascular Diseases, 28*, 632–639.

Sung, J., Song, Y. M., Choi, Y. H., Ebrahim, S., & Davey Smith, G. (2007). Hepatitis B virus seropositivity and the risk of stroke and myocardial infarction. *Stroke, 38*, 1436–1441.

Talabi, O. (2003). A 3-year review of neurologic admissions in University College Hospital Ibadan, Nigeria. *West African Journal of Medicine, 22*, 150–151.

Tamer, Y., Mohammad, H. J., Safi, U. K., Reed, M., Syed, Z. H., Michael, J. B., Ron, B., Salim, S. V., Michelle, C. J., Farhaan, V., Miguel, C., & Khurram, N. (2020). Stroke in young adults: Current trends, opportunities for prevention and pathways forward. *American Journal of Preventive Cardiology*.

Teh, W. L., Abdin, E., Vaingankar, J. A., Seow, E., Sagayadevan, V., Shafie, S., Shahwan, S., Zhang, Y., Chong, S. A., Ng, L. L., & Subramaniam, M. (2018). Prevalence of stroke, risk factors, disability and care needs in older adults in Singapore. *BMJ Open, 8*(3), Article e020285.

Thakur, K. T., Lyons, J. L., Smith, B. R., Shinohara, R. T., & Mateen, F. J. (2016). Stroke in HIV-infected African Americans: A retrospective cohort study. *Journal of Neurovirology, 22*(1), 50–55.

Verhoeven, J. I., Allach, Y., Vaartjes, I. C. H., Klijn, C. J. M., & de Leeuw, F. E. (2021). Ambient air pollution and the risk of ischemic and haemorrhagic stroke. *The Lancet Planetary Health, 5*, e542–e552.

Warren-Gash, C., Blackburn, R., Whitaker, H., McMenamin, J., & Hayward, A. C. (2018). Laboratory-confirmed respiratory infections as triggers for acute myocardial infarction and stroke: A self-controlled case series analysis of national linked datasets from Scotland. *European Respiratory Journal, 51*, 1701794.

Wasay, M., Khan, M., Farooq, S., Khowaja, Z. A., Bawa, Z. A., Mansoor Ali, S., Awan, S., & Beg, M. A. (2018). Frequency and impact of cerebral infarctions in patients with tuberculous meningitis. *Stroke, 49*, 2288–2293.

Watila, M. M., Nyandaiti, Y. W., Balarabe, S., Bakki, B., Alkali, N. H., Ibrahim, A., Elijah, T., & Chiroma, I. (2019). Aspiration pneumonia in patients with stroke, Northeast Nigeria. *National Journal of Neurology, 1*(3), 48–55.

Wellenius, G. A., Burger, M. R., Coull, B. A., Schwartz, J., Suh, H. H., Koutrakis, P., Schlaug, G., Gold, D. R., & Mittleman, M. A. (2012). Ambient air pollution and the risk of acute ischemic stroke. *Archives of Internal Medicine, 172*, 229–234.

World Health Organization Regional Office for Africa. (2022). *Report on malaria in Nigeria 2022*. https://www.afro.who.int/countries/nigeria/publication/report-malaria-nigeria-2022

Zewdie, A., Debebe, F., Kebede, S., Azazh, A., Laytin, A., Pashmforoosh, G., et al. (2018). Prospective assessment of patients with stroke in Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia. *African Journal of Emergency Medicine, 8*(1), 21–24.

Zhang, F. L., Guo, Z. N., Wu, Y. H., Liu, H. Y., Luo, Y., & Sun, M. S., et al. (2017). Prevalence of stroke and associated risk factors: A population-based cross-sectional study from Northeast China. *BMJ Open, 7*(9), e015758.

Comparative Analysis of Goal Programming Techniques in Bank's Investment Optimization

Ezra, Precious Ndidiamaka [*] ***and Chigbu, Polycarp Emeka***

Department of Statistics, University of Nigeria, Nsukka.

[*]Corresponding Author's e-mail- precious.ezra@unn.edu.ng

Abstract

This study explores and compares two prominent optimization techniques, lexicographic goal programming (LGP) and weighted sum goal programming (WSGP), in the context of a financial institution's investment decision-making. The objective is to evaluate their effectiveness in aiding a bank's investment strategy by maximizing profits, meeting capital-adequacy ratios and limiting risk-asset ratios while considering diverse constraints and goals. A hypothetical financial scenario involving a bank's investment allocation is used as a case study to demonstrate the contrasting outcomes between LGP and WSGP. The analysis involves optimizing investment allocations across various options while adhering to constraints on fund availability, investment diversification, liquidity requirements, and community-based lending. The results obtained using lingo 17.0x64 version software highlight that LGP excels in achieving individual goals sequentially, potentially reaching optimal values for each goal. However, it overlooks interdependencies between objectives, leading to suboptimal overall performance. On the other hand, WSGP provides a balanced approach considering multiple objectives simultaneously, offering a broader perspective on trade-offs but potentially sacrificing the precision of achieving specific goal values.

Keywords: Investment, Goal optimization, Financial institution, Management, Decision-making

1. Introduction

Investment management in banking institutions is a complex endeavour that requires a delicate balance between technical optimization methods and human decision-making. While technical tools and models play a crucial role in analyzing data and optimizing investment portfolios, the importance of human judgment cannot be overstated. Numerous studies have highlighted the significant impact of human decision-making on investment outcomes, alongside the application of technical optimization methods: see Barberis & Thaler (2003) and Kahneman & Tversky (1979).

Human decision-makers bring unique insights, intuition, and expertise to the investment process, enabling them to identify opportunities and risks that may not be captured by quantitative models alone. However, human decision-making is not immune to biases and

heuristics that can influence investment outcomes: see Markowitz (1952 & 1959), and Kristina (2010). According to Tversky & Kahneman (1974) and Barber & Odean (2000), research in behavioural finance has demonstrated the presence of cognitive biases such as overconfidence, anchoring, and loss aversion, which can lead to suboptimal investment decisions. For example, investors may exhibit a tendency to overweight recent performance when making investment decisions, leading to herding behavior and market inefficiencies: see DeBondt & Thaler (1985).

In addition to human biases, decision-making heuristics also play a significant role in shaping investment outcomes. According to Kahneman et al. (1982), heuristics are mental shortcuts or rules of thumb that individuals use to simplify complex decision-making tasks. While heuristics can be efficient in certain contexts, they can also lead to systematic errors in judgment and decision-making. For instance, the availability heuristic, whereby individuals base their judgments on readily available information, can lead investors to overweight the importance of recent news or events in their investment decisions, potentially leading to suboptimal outcomes: see Tversky & Kahneman (1973).

To complement human decision-making and mitigate the impact of biases and heuristics, banking institutions are increasingly turning to advanced technological tools and analytical methods. The integration of technology, such as artificial intelligence (AI) algorithms and machine learning models, into investment decision-making processes holds the promise of enhancing decision-making efficiency and effectiveness: see Lo & Mueller (2010) and Gu & Tang (2014). AI algorithms can analyze vast amounts of financial data, identify patterns, and generate insights to assist human decision-makers in evaluating investment options and managing risks: see Makridakis et al. 2018. According to Rudin (2019), machine learning models can adapt and improve over time, learning from past data and refining their predictions to support more informed investment decisions. Zhang et al. (2019) and Choy & Ng (2021) stated that by integrating technology into investment decision-making processes, banking institutions can leverage the complementary strengths of human judgment and technological capabilities, leading to more robust and effective investment management practices. However, Leung et al. (2020) noted that it is essential to recognize that technology is not a panacea and should be used in conjunction with human expertise to achieve optimal results. Collaborative decision-making processes that combine the insights of human decision-makers with the analytical power of technology have the potential to revolutionize investment management practices and drive superior investment outcomes in banking institutions.

The study adopts a socio-technical framework that acknowledges the intricate interplay between human decision-makers, technological tools, and organizational structures in the context of investment management. This framework recognizes that successful investment outcomes result not only from technical optimization techniques but also from the effective collaboration and synergy among these key elements. In the socio-technical framework, human decision-makers are viewed as essential components of the investment process, bringing valuable expertise, intuition, and judgment to the table. Their role is not replaced but rather augmented by technological tools, which serve as enablers for efficient data analysis, risk assessment, and decision support. Furthermore, organizational structures play a critical role in facilitating communication, coordination, and alignment of objectives among different stakeholders involved in the investment management process. In order to optimize investment decisions, human expertise is leveraged in conjunction with goal programming techniques, a methodology that enables decision-makers to prioritize and balance multiple conflicting objectives: see Charnes & Cooper (1977) and Abha (2015). Human decision-makers provide input regarding investment goals, risk preferences, and constraints, which are then incorporated into the goal programming model. This collaborative approach ensures that investment decisions align with the broader strategic objectives of the organization while accounting for the preferences and insights of human decision-makers.

In this study, the concerned investment bank is set to maximize her return (profit) and capital adequacy ratio (CAR), and at the same time minimize the risk asset ratio. Capital adequacy ratio determines the investment viability and reliability of the bank. A minimum capital adequacy ratio is critical in ensuring that banks have enough cushion to absorb a reasonable amount of losses before they become insolvent and consequently lose depositors' funds. CAR is therefore an indicator of how well a bank can meet its obligations. Also known as the capital-to-risk weighted assets ratio (CRAR), the ratio compares capital to risk-weighted assets and is watched by decision-makers to determine a bank's risk of failure. It's used to protect depositors and promote the stability and efficiency of financial systems around the world. Charnes & Cooper (1977) and Dinesh & Sridha (2016) noted that goal programming is the best operational research technique that is capable of handling investment managerial challenges with many investment criteria.

According to Zadeh & Khalili (2017), goal programming is an extension of linear programming in which management's objectives are treated as goals to be attained as closely as possible,

within the practical constraints of the problem. Ezra et al. (2020) stated that due to the possibilities of non-achievement of stated goal, the deviation variables, δ_i^- and δ_i^+ are added to the goal constraints. Nabendu & Manish (2012), Nasruddin et al. (2013) and Sen & Nandi (2012) noted that in goal optimization, interest is in minimizing the unwanted goal deviations. According to Safiai et al. (2012), a deviation variable assumes zero value when there is an exact achievement of a set goal. It was also affirmed that if complete attainment of any goal is not possible, that goal programming provides the decision-maker with the information on the amount by which the goal is not achieved. According to Kumar (2019), goals can be prioritized or weighted by assigning a priority factor, p_i (i = 1, 2, …, n) or weight, w_i (i = 1, 2, …, n) to the deviation variables, to indicate subjective importance attached to the goals. Hence, in this paper, two major goal programming techniques are explored and compared for an optimal allocation of investment funds to the various investment options to ensure an optimal solution to the bank's set goals. Such goal programming techniques include: Lexicographic goal programming (LGP) and Weighted sum goal programming (WSGP). LGP is a sequential optimization approach that prioritizes objectives hierarchically and optimizes goals in a predetermined order, while weighted sum goal programming tackles multiple objectives simultaneously by assigning weights to each goal within a single optimization problem. It aims to strike a balance between conflicting objectives by considering their relative importance through weighted aggregation. This approach offers a comprehensive view of the trade-offs among various goals but may lack the precision of achieving specific goal values when compared to lexicographic goal programming.

he contemporary literature on optimization in banking investment decisions reflects a diverse range of perspectives, methodologies, and empirical studies. Notably, Agarana et al. (2014) & Tanwar et al. (2021) have delved into practical applications of goal programming approaches in financial decision contexts, shedding light on the trade-offs and complexities inherent in employing LGP and WSGP. However, the selection between LGP and WSGP remains an ongoing debate, highlighting the need for a nuanced understanding of their comparative strengths, limitations, and implications for optimizing investment decisions in banking institutions.

2. Literature Review

Goal Programming (GP) is a multi-criteria decision-making technique. It is traditionally seen as an extension of linear programming to include multiple objectives, with goal target values

attached to each objective: see Dylan and Mehrdad (2003). Although goal programming (GP) is itself a development of the 1950s, it has only been discovered that it has finally received truly substantial and widespread attention since 1970s. Much of the reason for such interest is due to the fact that GP has demonstrated ability to serve as an efficient and effective tool for the modelling, solution, and analysis of mathematical models that involve multiple and conflicting goals; the type of models that most naturally represent real-world problems: see Ignizio (1985).

Goal programming as the most important technique among other multi-criteria decision making techniques can be applied in various field of study such as Academic Management, Agricultural Management, Classical Operation Research Application, Engineering sectors, Environmental and Waste Management, Finance, Health Planning, Information Technology and Computing, Management and Strategic Planning, Military, Natural Resource Management, Production Planning, Socio-Economic Planning, Theoretical, Transportation and Distribution. For instance, Charnes et al. (1963) used goal programming to show managerial method of analysing the relationship between the sales volume and the profit of a firm. The result showed that the implementation of goal programming in decision making among conflicting objectives in financial analysis could help in maximizing a company's wealth.

Rustagi (1976) applied goal programming to forest management planning for timber production and the proposed model could help in preparing and revising management plans for forests which are likely to remain permanently under timber production. The proposed model provides a reforestation plan based on the future expectations for different product output and return from the land, capital and other resources, and the possible reforestation alternatives which may be available. Rehman and Romero (1984) used goal programming to optimize the diet cost, nutrient imbalances and the bulk intake by the animal. Tahir and Carlos (1986) applied goal programming to livestock ration formulation as a method of overcoming restraint rigidities in generating optimal diets. Subsequently. Fabozzi and Drake (2009) used goal programming as a method of achieving the goal of maximizing a company's shareholders' wealth and it was discovered that goal programming could aid in developing strategies which serve as a guide for the company to attain their optimal solution among conflicting objectives. Grace and Olasumbo (2018) applied goal programming to production planning in small and medium bottled water manufacturing enterprise and discovered an optimal production plan that could be used to meet the strategic objectives of a Small and Medium Bottled Water Production

enterprise. This was achieved by developing a mathematical model for the allocation of capital and land among the various timber production alternatives.

Gan, et al. (2021) applied goal programming to selecting suitable, green port crane equipment for international commercial ports. In their study, they formulate a new data envelopment analysis method for collecting real-world CO_2 emissions data at their source and the method was applied to collect real-world operating data from a large container-handling company in Taiwan. They discovered the optimal port selection that provides the best balance between environmental protection and profitability. Ali and Teg (2021) used lexicographic approach in solving goal programming problem. They considered six goals level, namely; maximize asset, minimize liability, maximize equity, maximize operating income, maximize net income, and maximizing total goal achievements and discovered an optimal banking model for the financial management in the Kingdom of Saudi Arabia.

The previous literatures revealed the singular applications of the goal programming techniques. Hence, in this study, the two major goal programming techniques, namely: the Lexicographic Goal Programming (LGP) model and the Weighted sum Goal Programming (WGP) model were applied in solving the same problem, and the results compared, with the bid to know which is better than the other in goal achievements.

3. Material and Methodology

This section of the work discusses the materials (data) and the methods of analyses used for the study.

3.1. Material

In this study, human expertise is incorporated into the investment decision-making process through a participatory approach, whereby decision-makers provide input regarding investment goals, risk preferences, and constraints. Goal programming techniques are therefore employed to formulate an optimization model that aligns investment decisions with the specified objectives while addressing any trade-offs or conflicts among them. This collaborative process ensures that investment decisions reflect the collective wisdom and expertise of human decision-makers within the banking organization. The bank in question has eight investment options with a challenge of how much of its fund should be channelled into each option in order to maximize profits, capital-adequacy ratio and minimize risk-assets ratio. The bank under study is also facing constraints on the availability of funds to be used in its investment activities and management policies. This bank has about twenty million dollars (20m USD) as the capital,

one hundred and fifty million dollars (150m USD) as demand deposits and eighty million dollars (80m USD) as time deposits. Having known the bank's preferred areas of investment, the distribution of the available funds among the investment areas to achieve maximum return (profit) under an acceptable level of risk becomes a major challenge. The decision on the apportioning of funds should not be done arbitrarily since there may not be a guarantee of achieving the set goals. A scientific approach therefore becomes an acceptable option.

The bank in question, according to its records has never used any scientific approach for its investment decisions and as a result cannot tell whether it has been making the desired investment progress or not, over the past years. Hence, the need for this study. The needed information for this study is as provided in Table 1.

Table 1: Bank's investment information

	Bank Investment option (j)	Return rate (%)	Liquid part (%)	Required capital (%)	Risk asset (%)
1	Cash (x_1)	0.0	100.0	0.0	No
2	Short term (x_2)	4.0	99.5	0.5	No
3	Government:2 years (x_3)	4.5	96.0	4.0	No
4	Government:4 years (x_4)	5.5	90.0	5.0	No
5	Government:6 years (x_5)	7.0	85.0	7.5	No
6	Instalment loans (x_6)	10.5	0.0	10.0	Yes
7	Mortgage loans (x_7)	8.5	0.0	10.0	Yes
8	Commercial loans (x_8)	9.2	0.0	10.0	Yes

The following pieces of information from the bank help to generate the relevant constraints of the model:

(i) Total investment fund must sum to the available capital and deposit funds.

(ii) Cash reserves must be at least 14% of demand deposit plus 4% of time deposits.

(iii) The portion of investments considered liquid should be at least 47% of demand deposits plus 36% of time deposits.

(iv) At least 5% of funds should be invested in each of the eight categories, for diversity.

(v) At least 30% of funds should be invested in commercial loans, to maintain the bank's community status.

Moreover, the bank aims at achieving three specific goals, namely:

Goal 1: Making profit of at least 18.5 million USD

Goal 2: Achieving capital-adequacy ration of at least 0.8 million USD

Goal 3: Having risk-asset ratio of at most 7.0 million USD

3.2. Methodology

The bank's investment problem above was solved using both lexicographic goal programming and weighted sum goal programming techniques to effectively allocate funds across various investment options in the bank and also reveal the extent of achievement of the bank's set goals. Also, technological tool such as lingo 17.0x64 version is utilized to facilitate data analysis and decision support in the investment management process. These tools enable the efficient processing of financial data, identification of investment opportunities, and assessment of risk factors. By leveraging advanced analytics and predictive modeling techniques, decision-makers can gain valuable insights into market trends, asset performance, and portfolio dynamics, thereby enhancing the effectiveness and efficiency of investment decision-making.

3.2.1. Lexicographic Goal Programming (LGP) Model

LGP provides a flexible approach to tackling decision-making problems with multiple conflicting objectives, allowing decision-makers to explicitly incorporate preferences and priorities into optimization process while seeking a balanced solution. In the lexicographic goal programming approach, the optimization prioritizes individual goals in a sequential manner, aiming to optimize one objective at a time according to its predetermined priority order. Each goal is solved individually, and the solution is adjusted to meet the criteria of the subsequent goals without compromising the achievement of the higher-priority goals. Goals are prioritized to represent their relative importance in the decision-making process. LGP enables decision-makers to efficiently address objectives based on their significance, leading to feasible and optimized solutions. The general model associated with lexicographic goal programming is as given in (1):

$$Minimize\ Z = \sum_{i=1}^{n} p_i\,(\delta_i^+ + \delta_i^-) \quad (1)$$

Subject to

Goal Constraints and Structural Constraints.

where n is the number of goals;

m is the number of decision variables;

δ_i^+ is the amount of positive deviation from the i^{th} goal;

δ_i^- is the amount of negative deviation from the i^{th} goal;

p_i is the priority level attached to the i^{th} goal.

Hence, the lexicographic goal programming (LGP) model for the bank's investment problem formulated based on the established objective functions and constraints, is as given in (2):

Minimize Z = $\sum_{i=1}^{n} p_i \, (\delta_i^+ + \delta_i^-)$ (2)

Subject to

$$\sum(a_{ij}x_j + \delta_1^- - \delta_1^+) = G_1$$

$$\sum(a_{ij}x_j + \delta_2^- - \delta_2^+) = G_2$$

$$\sum(a_{ij}x_j + \delta_3^- - \delta_3^+) = G_3$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}x_j = T$$

$$x_1 \geq CR$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}x_j \geq L$$

$$x_j \geq D, j = 1,2,\ldots,m$$

$$x_8 \geq C$$

where

G_i, i =1 , 2, ..., n are the set goals; x_j, $j = 1,2,\ldots,m$ are the investment options; a_{ij};i =1 , 2, ..., n and $j = 1,2,\ldots,m$ are the weights attached to the investment options which serve as the coefficients; T is the total investment fund; CR is the least cash reserve; L is the total investment fund considered liquid; D denotes the least amount of fund that can be invested in each investment option for diversification purposes; C is the least amount that can be given out as a commercial loan;

δ_i^+ is the amount of positive deviation from the i^{th} goal;

δ_i^- is the amount of negative deviation from the i^{th} goal;

p_i is the priority level attached to the i^{th} goal.

3.2.2. Weighted Sum Goal Programming (WSGP) Model

WSGP considers multiple conflicting objectives simultaneously, assigns weight to each goal and aims to strike a balance between conflicting objectives within a single optimization problem. Goals are weighted to represent their relative importance in the decision-making process. The higher the weight, the higher the priority. The objective function is optimized by adjusting decision variables to achieve a balance among the multiple objectives, considering their weighted importance. The solution obtained provides a trade-off among multiple objectives based on the assigned weights, aiming to reach an overall optimal solution that balances the different objectives. The general model associated with lexicographic goal programming is as given in (3)

$Minimize\ Z = \sum_{i=1}^{n} w_i\ (\delta_i^+ + \delta_i^-)$ (3)

Subject to

Constraints as in the LGP model in (2)

where : w_i is the weight assigned to the i^{th} goal. Every other variable is as defined earlier.

Using the investment information in Table 1, the following problems are formulated:

With the rates of return from the bank's investment chart, the profit objective function becomes

$Max\ P = 0.040x_2 + 0.045x_3 + 0.055x_4 + 0.070x_5 + 0.105x_6 + 0.085x_7 + 0.092x_8$ (4)

In order to quantify the investment risk, we employ the two commonly used ratio measures. The first is the capital adequacy ratio, which measures a bank's financial strength, using its capital and assets. It is used to protect depositors and promote the stability and efficiency of bank's financial system. It is expressed as the ratio of the required capital to the actual capital. A high value of capital adequacy ratio indicates minimum risk and vice versa. The objective of the concerned bank is to maximize the capital adequacy ratio. The second objective function therefore becomes.

$Max\ C = (20)^{-1}(0.005x_2 + 0.040x_3 + 0.050x_4 + 0.075x_5 + 0.100x_6 + 0.100x_7 + 0.100x_8)$ (5)

Another type of risk that the concerned investment bank faces as seen on the supplied investment information focuses on illiquid assets risk. Illiquid assets are those assets that cannot be easily converted to cash. Such assets have more risk attached to them than other assets such that they can be difficult or even impossible to sell without a heavy discount. This implies that one could ultimately get back less than what one paid for them. This type of risk is measured by the ratio of risk asset to the actual capital and a low ratio indicates a financially secure institution. The objective of the investment management unit of this bank is to minimize the illiquid assets risk. Hence, the third objective function becomes:

$Min\ R = (20)^{-1}\ (x_6 + x_7 + x_8)$ (6)

where x's are the investment options.

The bank is set to make a profit (return) of at least 18.5 million USD with the capital- adequacy ratio of 0.8 million USD and illiquid asset risks of 7.0 million USD. Hence, the three set goals are:

Goal 1: Profit = 18.5

Goal2: Capital-adequacy ratio = 0.8

Goal3: Risk-asset ratio = 7.0.

With the supplied information above, the multiple objective linear programming (MOLP) model of this bank's investment can be stated as follows:

$Max\ P = 0.040\ x_2 + 0.045\ x_3 + 0.055\ x_4 + 0.070\ x_5 + 0.105\ x_6 + 0.085\ x_7 + 0.092\ x_8 \leq 18.5$

$Max\ C = (20)^{-1}\ (0.005x_2 + 0.040\ x_3 + 0.050\ x_4 + 0.075\ x_5 + 0.100\ x_6 + 0.100x_7 + 0.100\ x_8) \leq 0.8$

$Min\ R = (20)^{-1}\ (x_6 + x_7 + x_8) \geq 7.0$

Subject to:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 250 \quad (7)$$

$x_1 \geq 24.2$

$x_1 + 0.995\ x_2 + 0.960\ x_3 + 0.900\ x_4 + 0.850\ x_5 \geq 99.3$

$x_j \geq 12.5$ for j= 1,2,…,8

$x_8 \geq 75$

$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \geq 0$

For goals 1 and 2, interest is to minimize the underachievement of the profit goal and capital adequacy ratio goal, since overachievement of each is allowed for the concerned bank. For goal 3, interest is to minimize the overachievement of the illiquid assets' risk since higher illiquid assets risk may be hazardous to the operation of the bank.

3.3. Data Analysis for LGP

$Minimize\ Z_1 = p_1(\delta_1^-) = G_1$

$Minimize\ Z_2 = p_2(\delta_2^-) = G_2$

$Minimize\ Z_3 = p_3(\delta_3^+) = G_3$

Subject to

$0.040\ x_2 + 0.045\ x_3 + 0.055\ x_4 + 0.070\ x_5 + 0.105\ x_6 + 0.085\ x_7 + 0.092\ x_8 - p_1(\delta_1^+ - \delta_1^-) = 18.50.$

$(20)^{-1}\ (0.005x_2 + 0.040\ x_3 + 0.050\ x_4 + 0.075\ x_5 + 0.10\ x_6 + 0.10x_7 + 0.10\ x_8) - p_2(\delta_2^+ - \delta_2^-) = 0.8$

$(20)^{-1}\ (x_6 + x_7 + x_8) - p_3(\delta_3^+ - \delta_3^-) = 7.0$

$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 250$

$$x_1 \geq 24.2 \quad (8)$$

$x_1 + 0.995\ x_2 + 0.960\ x_3 + 0.900\ x_4 + 0.850\ x_5 \geq 99.3$

$x_j \geq 12.5$ for j= 1,2,…,8

$x_8 \geq 75$

$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \delta_i^+, \delta_i^- \geq 0$

The above problems were solved using lingo 17.0x64 version software and the results are displayed in tables 2, 3, 4, and 5.

Table 2: Investment Options with their optimal allocations

Investment Options	Optimal Allocations (in millions of USD)		
	G_1P_1	G_2P_2	G_3P_3
Cash reserve (x_1)	24.20	24.20	24.20
Short term loan (x_2)	19.73	12.50	16.03
Government:2 years (x_3)	12.50	12.50	12.50
Government:4 years (x_4)	12.50	12.50	12.50
Government:6 years (x_5)	37.90	46.37	44.77
Installment loans (x_6)	55.67	45.75	52.50
Mortgage loans (x_7)	12.50	21.18	12.50
Commercial loans (x_8)	75.00	75.00	75.00

Table 3: Set goals and their levels of achievement (for various priorities)

Goal	Amount of Goal Deviation	Level of Achievement
G_1P_1	δ_1^-= 0.00	exactly achieved
	δ_1^+ = 0.00	exactly achieved
	δ_2^- = 0.00	exactly achieved
	δ_2^+ = 0.1192	overachieved
	δ_3^- = 0.00	exactly achieved
	δ_3^+ = 0.1582	overachieved
G_2P_2	δ_1^-=0.00	exactly achieved
	δ_1^+ =0.00	exactly achieved
	δ_2^- = 0.00	exactly achieved
	δ_2^+ = 0.1429	overachieved
	δ_3^- = 0.00	exactly achieved
	δ_3^+ = 1.6286	overachieved
G_3P_3	δ_1^- = 0.00	exactly achieved
	δ_1^+ =0.00	exactly achieved
	δ_2^- =0.00	exactly achieved
	δ_2^+ = 0.1281	overachieved
	δ_3^- = 0.00	exactly achieved
	δ_3^+ = 0.00	exactly achieved

3.3.1. Data Analysis for WSGP (with equal or unequal weights)

The bank's investment information is formulated as WSGP problem in (9)

Minimize $W = w_1(\delta_1^-) + w_2(\delta_2^-) + w_3(\delta_3^+)$ (9)

subject to

$0.040\,x_2 + 0.045\,x_3 + 0.055\,x_4 + 0.070\,x_5 + 0.105\,x_6 + 0.085\,x_7 + 0.092\,x_8 - w_1(\delta_1^+ - \delta_1^-) = 18.50.$

$(20)^{-1}(0.005x_2 + 0.040\,x_3 + 0.050\,x_4 + 0.075\,x_5 + 0.10\,x_6 + 0.10x_7 + 0.10\,x_8) - w_2(\delta_2^+ - \delta_2^-) = 0.8$

$(20)^{-1}(x_6 + x_7 + x_8) - w_3(\delta_3^+ - \delta_3^-) = 7.0$

$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 250$

$x_1 \geq 24.2$

$x_1 + 0.995\,x_2 + 0.960\,x_3 + 0.900\,x_4 + 0.850\,x_5 \geq 99.3$

$x_j \geq 12.5$ for j= 1,2,…,8

$x_8 \geq 75$

$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \delta_i^+, \delta_i^- \geq 0$

In the course of the analysis, using lingo 17.0x64 version, the above problems were solved with both the same and varying weight assignments and the constantly (same) generated results are as shown in Tables (4) and (5).

Table 4: Investment Options with their optimal allocations in (Millions of USD)

Investment Options	Optimal Allocations
Cash (x_1)	24.20
Short term (x_2)	16.03
Government:2 years (x_3)	12.50
Government:4 years (x_4)	12.50
Government:6 years (x_5)	44.77
Installment loans (x_6)	52.50
Mortgage loans (x_7)	12.50
Commercial loans (x_8)	75.00

Table 5: Set goals (with various weights) and their levels of achievement

Set goal	Amount of goal deviation	Level of Goal Achievement
	δ_1^- = 0.0000	exactly achieved
	δ_1^+ = 0.0000	exactly achieved
0.8	δ_2^- = 0.0000	exactly achieved
	δ_2^+ = 0.1281	overachieved
7.0	δ_3^- = 0.0000	exactly achieved
	δ_3^+ = 0.0000	exactly achieved

4. Discussion of Results

This study is concerned with optimizing the allocation of investment funds across various investment options within a bank, considering multiple constraints and goals, while determining the extent of achievement of the set goals . Tables 2,3,4 and 5 show the results obtained using the two optimization approaches: the lexicographic and weighted sum goal programming, each offering optimal allocations for investment options to meet specific goals. Table 2 shows the effects of a goal's priority on the allocation of investment fund to the various investment options while Table 3 shows how one goal's priority affects the achievement of other goals. The result from Table 3 shows that the values for both negative deviation and positive deviation of the profit goal is each zero, indicating that the investment bank completely achieved their profit goal; while at the same time, the positive deviation value for CAR goal is 0.1192, and the positive deviation value for risk goal is 0.1582. This implies that the concerned bank can only achieve their profit goal as their first priority (ρ_1) if their capital adequacy ratio increases by 12%, so as to be able to cushion the effect of risk increament of 16%.

The CAR goal for priority (ρ_2) is also fully achieved since the negative deviation is minimized to zero. Also, the positive deviation of the goal is increased by 0.1429, indicating that the bank should increase their CAR by 14% so as to comfortably achieve the profit goal as the risk ratio increases by 1.6286.

The third priority (ρ_3) which is the risk goal, is completely achieved since both the negative and positive deviations are each minimized to zero. Also, the profit goal is completely achieved

while the CAR goal is overachieved by 0.1281. This implies the need for the concerned bank to increase their CAR goal by 13% if they must achieve their risk and profit goals concurrently.

Finally, the optimal allocation results obtained at the point of achieving the third priority goal in LGP (shown in column 4 of Table 2) are the same with the optimal allocation results obtained using WSGP (shown in column 2 of Table 4), and the levels of goals' achievements considering the third priority (shown in Table 3), are also the same with the ones obtained using the WSGP approach, as shown in Table 5. This confirms the reason for also having the objective function of LGP as a weighted sum of the goal deviations in WSGP. It is important to note that the same results were obtained using (WSGP) both with equal and different weights assignments.

5. Conclusion

The comparative study between lexicographic goal programming (LGP) and weighted sum goal programming (WSGP) within the context of optimizing a bank's investment decision reveals nuanced trade-offs and distinct advantages of each. LGP's sequential strategy demonstrates its capability to achieve the objectives systematically. By prioritizing goals in a predetermined sequence, LGP optimizes each objective successively, potentially yielding optimal solutions for specific goals. However, its narrow focus on one objective at a time may result in overlooking the intricate interdependencies and trade-offs among different objectives. This limitation could lead to suboptimal overall performance, neglecting the holistic perspective crucial in complex decision-making scenarios. Conversely, WSGP's simultaneous consideration of multiple objectives through weighted aggregation offers a more comprehensive view of the trade-offs among different goals. By incorporating relative weights assigned to each objective, WSGP strives to strike a balance among conflicting objectives. This approach enables decision-makers to explore various scenarios, providing a broader understanding of how different goals interact. However, WSGP's inability to precisely target specific goal values, as seen in LGP, might limit its suitability in situations requiring absolute attainment of individual objectives.

The implications of this comparative analysis underscore the significance of selecting an optimization technique that aligns with the nature of the problem, the hierarchy of objectives, and the organizational preferences. In investment decision-making for financial institutions like banks, where multiple goals coexist alongside constraints, the choice between LGP and WSGP depends on the emphasis placed on achieving specific goal values versus obtaining a

balanced compromise across multiple objectives. Ultimately, the selection between LGP and WSGP hinges on the decision-makers' preferences regarding precision in meeting individual goals versus a more holistic assessment considering trade-offs among competing objectives.

The findings of this study have significant implications for enhancing collaboration between human decision-makers and technology in investment management processes within banking institutions. By recognizing the complementary roles of humans and technology, organizations can foster a culture of collaboration and trust, enabling decision-makers to leverage advanced analytics capabilities and data-driven insights to make more informed investment decisions. This collaboration facilitates the alignment of investment strategies with organizational goals and risk appetite, leading to more effective risk management and value creation. Future research could explore hybrid approaches that leverage the strengths of both LGP and WSGP, aiming to mitigate their respective limitations and offer a more robust framework for complex decision optimization in financial setting.

References

Abha, D. (2015). Goal programming applications in agricultural management. International Research Journal of Engineering and Technology (IRJET), 2(6), 883–887.

Agarana, M. C., Bishop, S. A., & Odetunmibi, O. A. (2014). Optimization of banks loan of portfolio management using goal programming technique. International Journal Research in Applied Natural and Social Sciences, 2(8), 2347–4580.

Ali, A., & Teg, A. (2021). Lexicographic goal programming for bank's performance management. Journal of Applied Mathematics.

Barber, B. M., & Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. The Journal of Finance, 55(2), 773–806.

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. Handbook of the Economics of Finance, 1(1), 1053–1128.

Charnes, A., & Cooper, W. W. (1977). Goal programming and multiple objective optimizations: Part 1. European Journal of Operational Research, 1(1), 39–54.

Charnes, A., Cooper, W. W., & Ijiri, Y. (1963). Break-even budgeting and programming to goals. Journal of Accounting Research, 1(1), 16–39.

Choy, S. Y., & Ng, C. S. P. (2021). Machine learning in financial markets: A comprehensive survey. Journal of Financial Data Science, 3(2), 65–105.

DeBondt, W. F., & Thaler, R. H. (1985). Does the stock market overreact? The Journal of Finance, 40(3), 793–805.

Dylan, F. J., & Tamiz, M. (2003). Goal programming in the period 1990–2000. International Series in Operations Research & Management Science, 52, 129–170.

Dinesh, K., & Sridhar, K. (2016). Multiple objective operations decisions by using goal programming for a small-scale enterprise. International Journal of Research in Engineering and Technology, 4(6), 7–12.

Ezra, P. N., Oladugba, A. V., Ohanuba, F. O., & Opara, P. N. (2020). Goal optimization of a pastry company. American Journal of Operational Research, 10(1), 17–21.

Fabozzi, F. J., & Drake, P. P. (2009). Finance: Capital markets, financial management, and investment management. John Wiley & Sons, Inc.

Frank, K. R., & Keith, C. B. (2012). Investment analysis and portfolio management (10th ed.).

Gan, G. Y., Lee, H. S., Tao, Y. J., & Tu, C. S. (2021). Selecting suitable green port crane equipment for international commercial ports. Sustainability, 13(12), 6801.

Grace, M. K., & Olasumbo, A. M. (2018). A goal programming model for production planning in a small and medium bottled water manufacturing enterprise. Proceedings of the International Conference on Industrial Engineering and Operations Management, 1077–1084.

Gu, B., & Tang, L. C. (2014). The impact of artificial intelligence on stock market prediction: A review and empirical evidence. International Journal of Economics, Commerce and Management, 2(8), 1–11.

Ignizio, J. P. (1985). Introduction to linear goal programming (Quantitative Applications in the Social Sciences; No. 07-056). Sage Publications, Inc.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge University Press.

Kristina, L. (2010). Investment analysis and portfolio management (Leonardo da Vinci Programme Project). Vytautas Magnus University.

Kumar, P. P. (2019). Goal programming through bakery production. International Journal of Scientific and Technology Research, 8(10), 3722–3725.

Leung, S. H., Ng, S. K., & Xin, X. (2020). Artificial intelligence in finance: A review and future direction. International Journal of Financial Engineering, 7(2), 2050006.

Lo, A. W., & Mueller, M. T. (2010). Warning: Physics envy may be hazardous to your wealth! International Journal of Theoretical and Applied Finance, 13(8), 1125–1143.

Markowitz, H. (1952). Mean-variance analysis in portfolio choice and capital markets. John Wiley & Sons.

Markowitz, H. (1959). Portfolio selection: Efficient diversification of investments. John Wiley & Sons.

Nabendu, S., & Manish, N. (2012). An optimal model using goal programming for rubber wood door manufacturing factory in Tripura. Department of Mathematics, Assam University, Silchar-788011, India.

Nasruddin, H., Afifah, H. P., Nur, S. I., & Nurul, F. R. (2013). A goal programming model for bakery production. Advances in Environmental Biology, 7(1), 187–190.

Rustagi, K. P. (1976). Forest management planning for timber production: A goal programming approach. Yale University School of Forestry and Environmental Studies Bulletin, 89, 80 p.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Safiai, S., Hassan, N., Mohammad, N. H., & Raduan, Z. A. (2012). Goal programming formulation in nutrient management for chilli plantation in Sungai Buloh Malaysia. Advances in Environmental Biology, 6(12), 4008–4012.

Sen, N., & Nandi, M. (2012). A goal programming approach to rubber plantation planning in Tripura. Applied Mathematical Sciences, 6(124), 6171–6179.

Tahir, R., & Carlos, R. (1986). Goal programming with penalty functions and livestock ration formulation. Agricultural Systems, 23, 117–132.

Tanwar, J., Vaish, A. K., & Rao, N. V. (2021). Optimizing balance sheet for banks in India using goal programming. International Journal of Accounting & Finance Review, 6(2), 81–101.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.

Zadeh, M. A., & Khalili, K. (2017). A goal programming model for liquidity management: A case study. International Journal of Services and Operations Management, 26(3), 332–346.

Zhang, G. P., Qi, M., & Yang, B. (2019). Neural networks for financial forecasting. In Handbook of Neural Computation (pp. 1203–1225).

Evaluation of practical central composite designs for optimum exploration of response surfaces

Eugene C. Ukaegbu[1] and Polycarp E. Chigbu[2]

[1]Department of Physical Sciences, State University of Medical and Applied Sciences, Igbo-Eno, Enugu State, Nigeria

[2]Department of Statistics, University of Nigeria, Nsukka, Nigeria

**Corresponding Author's email: eugene.ukaegbu@sumas.edu.ng*

Abstract

The drawback of the spherical,$\alpha = \sqrt{k}$, and rotatable, $\alpha = \sqrt[4]{f}$, $f = 2^k$ axial distances of the central composite designs (CCD) is the extreme values of the axial distance as the number of experimental factors, k, increases, resulting in impractical axial distances beyond the bounds of the design region. The practical central composite design compensates for this drawback by providing more stable and less extreme axial distance irrespective of the size of k. This study focuses on the evaluation of partially replicated cube and star portions of the variations of the CCD with practical axial distance, $\alpha = \sqrt[4]{k}$. By replicating the cube and star portions, The variation of the partially replicated central composite designs were evaluated using the following single-value optimality criteria: A-, D- and E-efficiencies and V-criterion. Fraction of design space graphs (FDSG) and variance dispersion graphs (VDG) were used to assess the scaled and unscaled prediction variances across the design regions. The replication of the star points yielded small and better distribution of the prediction variance across the design space and better results for the A-, G- and V-criterion. On the contrary, The D-efficiency was not improved by the replication of the cube portion. Lack-of-fit, residual and pure error degrees of freedom of partially replicated CCD were ascertained.

Keywords: axial distance, degrees of freedom, design efficiency, lack-of-fit, replication

1. Introduction

In response surface methodology, most popular and commonly used response surface design is the CCD which was developed by Box and Wilson (1951). This design is characterized by three distinct components: the centre, the star (axial component) and the cube (factorial component). The coordinates of the cube portion is of the form, $(\pm 1, \pm 1, \ldots, \pm 1)$ for the k independent variables, $(x_1, \ldots, x_k)$. This implies that the cube has the highest value set at +1 and the lowest value set at -1 which defines the extreme regions of the unit cube. The star has coordinates of the form, $(\pm\alpha,\ 0,\ \ldots,\ 0), (0,\ \pm\alpha,\ 0,\ \ldots,\ 0), \ldots, (0,\ 0,\ \ldots,\ \pm\alpha)$, such that α defines the position of the star points in the design region. The highest value of the star

points is at $+\alpha$ and lowest value at $-\alpha$. The centre component has coordinates of the form, $(0,\ 0,\ \ldots,\ 0)$.

Placing the star points on the axis of the design space defined by the CCD determines the distance of the star points from the centre of the design space and is important in the performance of the design with respect to prediction of responses. The axial distance, α, is the distance of the star points from the centre of the design space along the axes of the CCD. There are very important axial distances which exist in the literature for the CCD for different experimental conditions. One of such axial distances is the spherical alpha which puts the design points on a sphere of distance, $\alpha = \sqrt{k}$, from the centre of the design region. Another axial distance is the rotatable alpha, with distance, $\alpha = \sqrt[4]{f}$, from the centre of the design space, where $f = 2^k$ factorial runs, which gives a CCD that has equal precision at equal distance from the centre of the design region. Then, the cuboidal alpha at distance, $\alpha = 1$, puts the star points at the centre of the face of the unit cube, ensuring that the star points are distance $\alpha = 1$ from the centre of the cuboidal region.

The problem of the spherical and rotatable axial distances, as pointed out by Li et al (2009), is that as the number of factors increases, the values of the axial distances could be impracticable in response surface exploration. Consider, for example, an experiment involving k = 10, then, $\sqrt{10} = 3.1623$ and $\sqrt[4]{2^{10}} = 5.6569$, respectively, for the spherical and rotatable axial distances, which are not feasible values (see Table 1 for detailed results of the axial distances for k = 6 to 10). Therefore, Anderson and Whitcomb (2005) provided the practical axial distance, $\alpha = \sqrt[4]{k}$, which is a compromise between the spherical and cuboidal axial distances. The practical alpha, which is moderate and less extreme, as well offers reasonable variance inflation factor (VIF) as the number of factors, k increased. Ukaegbu (2018) compared the prediction variances of the practical axial distance of the CCD with the prediction variances of the spherical, face-centered, rotatable and orthogonal axial distances for 3 to 6 design factors using variance dispersion graphs (VDG) and the fraction of design space (FDS) graphs.

In the present study, we employ the advantage that the practical alpha offers in evaluating the CCD with partial replications of the factorial and star portions. To enhance the evaluation and comparisons of the replicated designs variations, we used three popular single-value efficiency criteria, the *A*-, *D*- and *G*-efficiencies, and the *V*-criterion. Graphical methods employed in assessing and comparing the prediction variance properties of the patially replicated design

variations are the variance dispersion graphs (VDG) and the fraction of design space (FDS) graphs. The number of factors under consideration is k = 3 to 6 factors.

2. Literature Review

Anderson-Cook et al. (2009) listed some of the properties of a good design for fitting second-order response surfaces, some of which include providing sufficient information to allow the test of model lack-of-fit; be robust to outliers in the data; be robust to errors in the control of design levels; provide a good distribution of the prediction variance throughout the design region; etc. (see Myers and Montgomery, 2002 and Anderson-Cook, 2005 for further references). The CCD with spherical and rotatable axial distances, have limitations in attaining some of the properties as the number of factors increases due to resulting impractical axial distances.

As already stated, the practical alpha, $\alpha = \sqrt[4]{k}$, was presented by Anderson and Whitcomb (2005) as a compromise between the spherical and cuboidal axial distances. According to Anderson and Whitcomb (2005), this practical axial distance provides (i) acceptable variance inflation factor (VIF), and (ii) design points which are less extreme with increase in the number of design factors. Generally, The VIF is a measure of how much the variances of estimated model parameters are inflated when compared to the predictor variables not being linearly related. This measure is very useful in detecting the presence of multicollinearity in regression models. Li et al. (2009) submitted that the practical axial distance results to near-rotatable central composite designs and offers stable scaled prediction variance across the design region. In the study by Li et al. (2009), the CCD with practical axial distance was shown to perform better than the other competing designs in the cuboidal region considering the scaled and unscaled prediction variances; the design with practical axial distance improved the prediction variance performance of the design throughout the design space by extending the axial points beyond the cuboidal region of interest.

In this study, the partially replicated variations of the CCD with practical axial distance was evaluated in the cuboidal region for k = 3 to 6 factors. These design variations were evaluated using *A, D, G* and *V* efficiency criteria and variance dispersion graphs (VDG) and fraction of design space (FDS) graphs used to measure the spread of the prediction variances.

3. Methods

3.1. Partial replications

The partially replicated variations of the central coposite designs considered in this study, as provided by Draper (1982) are: (i) Two cubes plus one star (C_2S_1); (ii). One cube plus two stars (C_1S_2); (iii). Three cubes plus one star (C_3S_1); (iv). One cube plus three stars (C_1S_3); (v). four cubes plus one star (C_4S_1); and (vi). One cube plus four stars (C_1S_4). These partially replicated design variations are compared with the CCD without replication of the cube or star portions, that is, (vii). One cube plus one star (C_1S_1). Each design variation was augmented with n_0 centre points in order to effectively determine the effect of increasing the number of centre points on the prediction capabilities of the replicated designs.

3.2. Measures of Design Properties

The *D*-efficiency is a useful statistical tool for quantifying the quality of estimated model parameters, and is defined by $D_{eff} = \{|X'X|^{1/p}/N\} \times 100$, where $N = 2^{k-q}n_c + 2n_sk + n_0$, q is the fraction of 2^k, n_c = number of replications of the cube, n_s = number of replications of the star, n_0 = number of replications of the centre point and *k* is the number design factors. The power, $1/p$, is the inverse of the number of model parameter estimates, *p*, being assessed as the determinant of the information matrix is being computed. The *A*-efficiency, given by $A - eff. = 100p/\{trace[N(X'X)^{-1}]\}$, is obtained as a function of the trace of the inverse of the information matrix, $(X'X)^{-1}$. Also, the *G*-efficiency, given by $G - eff. = 100p/[N\ max\ v\ (x)]$, is obtained from the scaled maximum prediction variance of the design. For the *V*-criterion, a design is *V*-optimal if it minimizes the normalized integrated scaled prediction variance, $V - opt. = min\frac{1}{\Psi}\int_{\Omega} v(x)dx$, where $\Psi = \int_{\Omega} dx$ is the volume of the design space. For *A*-, *D*- and *G*-efficiencies, higher values are desirable while smaller values are desirable for the *V*-criterion. MATLAB version 14 software was used to compute the efficiency criteria of the partially replicated designs.

3.3. Graphical methods for comparison

According to Anderson-Cook et al. (2009), single-value criteria such as the *D*- and *G*-efficiencies do not completely show the prediction variance characteristics of response surface designs. That is, the strength and weakness of a response surface design under evaluation cannot be captured completely using such single-value criteria. On the other hand, graphical methods of prediction variance assessment can completely display the distribution of the prediction variances of a design throughout the design space. One of these graphical methods

is the variance dispersion graph (VDG). Giovannitti-Jensen and Myers (1989) developed the VDG as a prediction variance-based graphical method that displays the spread of unscaled and scaled prediction variances of a multi-dimensional design region on a two-dimensional space. Prediction variance assessment using the VDG requires plotting the prediction variances against the radius, *r*, of the sphere from zero up to the outer region of the sphere covering the region of interest. Zahran et al (2003) developed the fraction of design space (FDS) graph to complement the VDG in prediction variance assessment of both spherical and cuboidal design regions. The FDS graphs display the characteristics of the scaled prediction variance (SPV) throughout a multi-dimensional region on a two-dimensional space with a single curve. The volume of the fraction of the design space is the key factor in the concept of the FDS criterion. The stability and prediction capability of a design are determined by how flat and close to the horizontal line the graph is. The FDS graphs were computed and plotted using Design Expert version 12, while the VDGs were computed and plotted using MATLAB version 14.

4. Results and Discussion

4.1. Results of alphabetic criteria

For ease of use in this study, *A, D, G* and *V* were used to refer to the four single value criteria, *A*-, *D*- and *G*-efficiencies and *V*-criterion, respectively. In Table 1, the results for the single-value efficiency criteria and *V*-optimality are displayed. For $k = 3$ factors, the table shows that additional centre point improves *V* with smaller values and *A* with higher percentage efficiency values. Increasing the number of centre points is beneficial for *D* and *G* for the higher replication of the cube, three cubes plus one star and four cubes plus one star, which have higher efficiency values when the number of centre points increases. The best *D* efficiency values are obtained for the cube-replicated design, C_2S_1, which maintained the highest *D* efficiency values even as the number of centre points increases. The unreplicated CCD, C_1S_1, has the overall best values for *A, G* and *V* with the highest percentage efficiency values for *A* and *G* and smallest optimality value for *V*. Also, C_1S_1 is the best for *G* without additional centre point. However, with increase in number of centre points, C_2S_1 has the best *G* efficiency value.

Table 1: Values of the *A*, *D*, *G* and *V* Criteria for the Partially Replicated Practical CCD

k	Design				$n_0 = 1$					$n_0 = 3$				
		F	$2n_sk$	α	*N*	*D*	*A*	*G*	*V*	*N*	*D*	*A*	*G*	*V*
3	C_1S_1	8	6	1.3161	15	55.3	37.5	89.1	5.9864	17	52.5	42.7	79.3	4.7882
	C_2S_1	16	6	1.3161	23	55.5	30.9	84.7	7.7480	25	54.8	38.3	83.9	5.5476
	C_1S_2	8	12	1.3161	21	51.4	34.2	69.6	6.1975	23	49.8	39.1	64.6	5.1268
	C_3S_1	24	6	1.3161	31	53.6	25.4	64.2	9.7021	33	54.0	32.8	67.7	6.5621
	C_1S_3	8	18	1.3161	27	47.6	30.2	57.5	6.6242	29	46.8	34.5	54.7	5.5637
	C_4S_1	32	6	1.3161	39	51.6	21.4	51.5	11.7102	41	52.5	28.3	56.5	7.6494
	C_1S_4	8	24	1.3161	33	44.5	26.8	49.1	7.1141	35	44.1	30.7	47.7	6.0216
4	C_1S_1	16	8	1.4142	25	58.1	41.8	63.0	7.5417	27	55.8	44.2	58.5	6.4203
	C_2S_1	32	8	1.4142	41	32.2	12.3	37.9	25.9217	43	32.0	13.2	37.0	23.8894
	C_1S_2	16	16	1.4142	33	55.0	42.7	75.6	7.0583	35	44.3	44.6	71.7	6.4323
	C_3S_1	48	8	1.4142	57	54.9	25.9	52.9	13.1066	59	54.9	30.3	54.6	9.8632
	C_1S_3	16	24	1.4142	39	51.2	39.9	63.4	7.1739	41	49.8	41.3	60.8	6.7236
	C_4S_1	64	8	1.4142	73	52.6	21.4	42.1	16.0643	75	53.0	25.4	43.9	11.8071
	C_1S_4	16	32	1.4142	49	47.8	36.8	54.7	7.4575	51	46.8	37.8	53.0	7.0890
5	C_1S_1	32	10	1.4954	43	60.2	41.6	90.9	8.4625	45	58.7	43.8	90.9	7.1443
	C_2S_1	64	10	1.4954	75	57.7	29.8	55.1	13.0996	77	57.3	32.6	56.4	10.5355
	C_1S_2	32	20	1.4954	53	58.9	47.4	85.2	6.6552	55	57.5	48.4	82.2	6.0779
	C_3S_1	96	10	1.4954	107	54.8	22.9	39.4	17.8309	109	54.8	25.4	40.7	14.0473
	C_1S_3	32	30	1.4954	63	55.8	47.0	73.5	6.0707	65	54.6	47.4	71.4	5.7248
	C_4S_1	128	30	1.4954	139	52.3	18.5	30.7	22.5881	141	52.6	20.7	31.8	17.5922
	C_1S_4	32	40	1.4954	73	52.7	45.0	64.9	5.8222	75	51.7	45.2	63.3	5.5798
6	C_1S_1	32	12	1.5651	45	61.5	48.1	94.0	10.4743	47	59.6	48.6	90.0	9.7810
	C_2S_1	64	12	1.5651	77	61.1	37.4	68.0	13.9348	79	60.2	39.1	70.7	12.3548
	C_1S_2	32	24	1.5651	57	57.7	50.7	57.3	9.6140	59	56.1	50.4	73.9	9.4249
	C_3S_1	96	12	1.5651	109	59.0	29.7	50.5	17.7553	111	58.5	31.5	51.5	15.3747
	C_1S_3	32	36	1.5651	69	53.1	47.8	64.6	9.8328	71	51.9	47.4	62.9	9.7848
	C_4S_1	128	12	1.5651	141	57.0	24.6	39.5	21.6692	143	56.9	26.3	40.5	18.4661
	C_1S_4	32	48	1.5651	81	49.2	44.2	56.1	10.3236	83	48.2	43.8	54.8	10.3343

For the four-factor CCD, $k = 4$, additional centre point improves *A* and *V* but the effect is inconsistent for *D* and *G*, where, for the cube-replicated options, C_3S_1 and C_4S_1, additional points at the centre improves *D* and *G* which is not the case with the replication of the star. The best value for *A* and *G* are obtained with C_1S_2 even with additional centre point. However, the values of the four alphabetic criteria indicated that further replications of the star does not in any way improve the performances of the designs. The unreplicated CCD, C_1S_1, gives the best

values for *D* even with additional point at the centre. Also, with additional centre point, C_1S_1 gives the best value for *V* but was outperformed by C_1S_2 when there is only one centre point. It is important to notice that replicating the cube portion from C_2S_1 to C_3S_1 improves the design's alphabetic criteria, which begins to deteriorate with further replication of the cube. For k = 3 and 4, further replication of the CCD does not improve any of the four alphabetic criteria.

The effects of additional centre point on the four alphabetic criteria for k = 3 and 4 factors are also true for k = 5 and 6. However, for k = 5, with or without additional centre point, C_1S_1 gives the best values for *D* and *G* while C_1S_2 offers the best values for *A*. Unlike the cases of k = 3 and 4, the higher the star is replicated, the better is the value of *V* with or without additional centre points. For k = 6, C_1S_1 gives the best for *G* with or without additional centre point and also the best for *D* with one centre point while C_2S_1 performs better with additional centre point. Also, C_1S_2 gives the best *A* and *V* values with and without an additional centre point.

4.2. Results of graphical methods

If the interest of a practitioner is to understand the prediction variance distribution of a design throughout the entire design region, that is, to understand the stability of the prediction variance throughout the entire design region and/or where in the region the design has the best and worst prediction variance, the two graphical methods are formidable tools for exploring the prediction variance properties of competing designs. In this section, the VDG and FDS graphs were plotted for the unscaled and scaled prediction variances in the spherical region for n_0 = 1 and 3. The graphs are displayed in Figures 1 to 4 for the VDGs and Figures 5 to 9 for the FDS plots.

The VDG and FDS graphs show that replicating the star resulted in minimum and stable spread of the prediction variances throughout the entire design space for both unscaled and scaled prediction variances. The VDGs show that the star-replicated practical CCD options have the lowest distribution of the unscaled and scaled prediction variances. All the designs maintain slightly stable prediction variance from the origin up to the point where the radius, r = 1.0, then the prediction variance increases rapidly as r increases towards the extreme of the design region. Additional centre point tends to reduce the prediction variance for both the unscaled and scaled prediction variance.

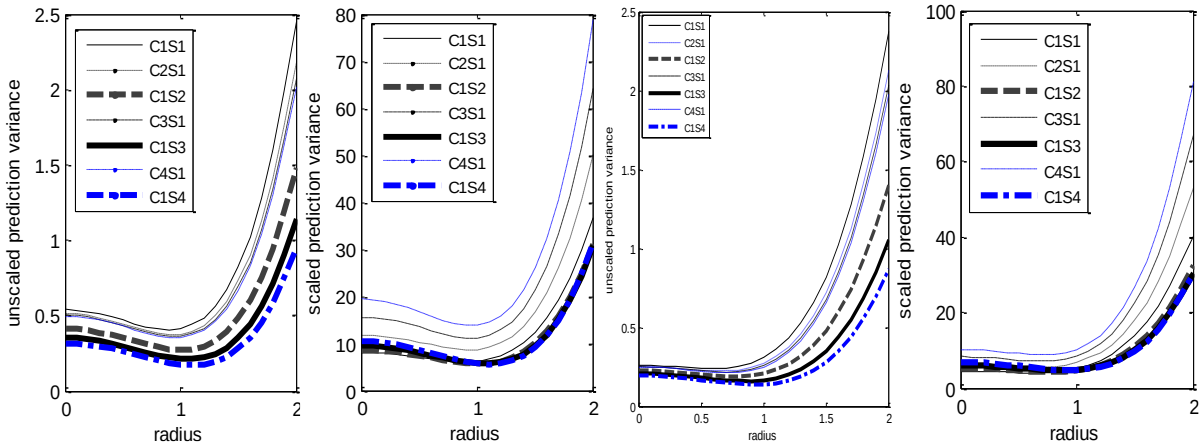


Figure 1: VDG for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, and k = 3

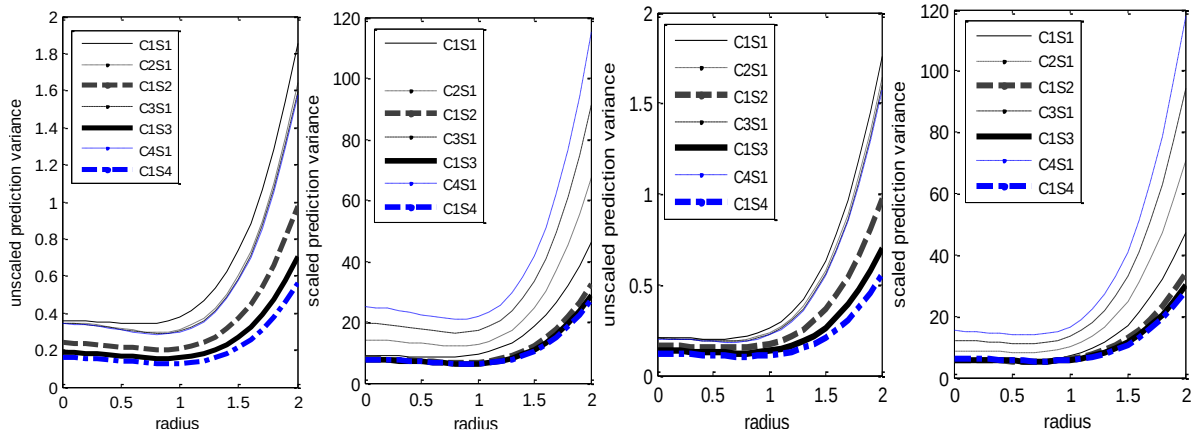


Figure 2: VDG for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$ and, k = 4

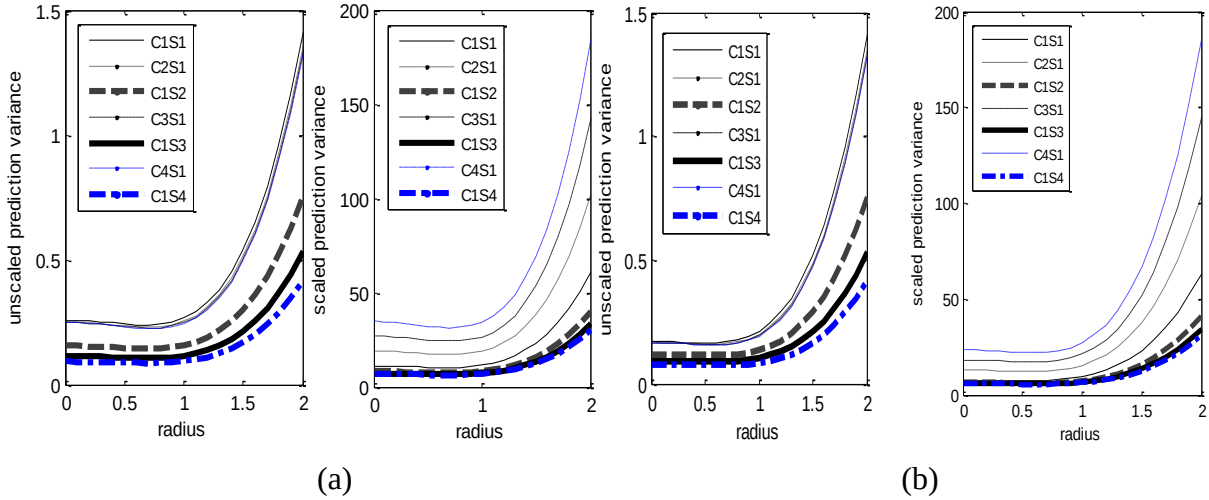


Figure 3: VDG of Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, for k = 5

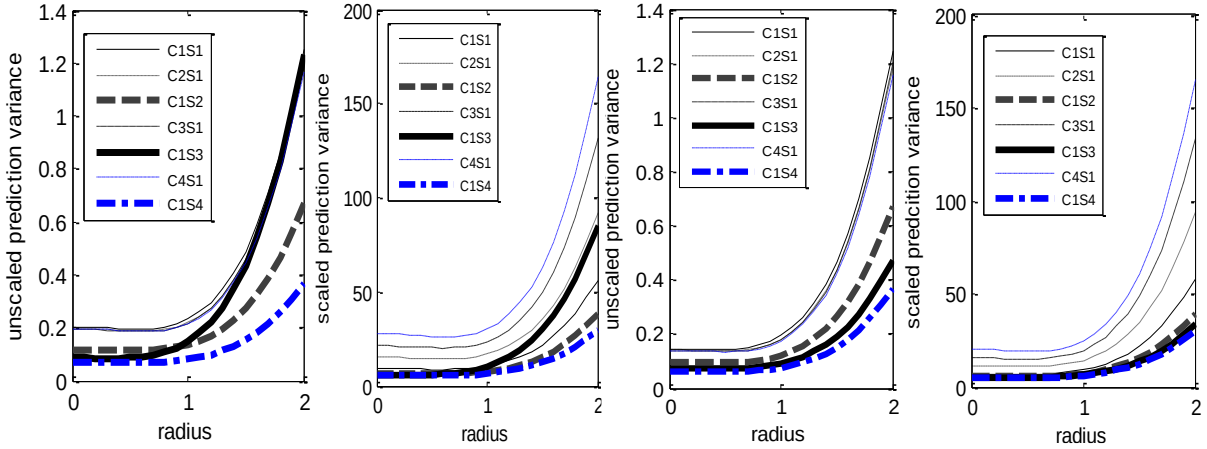


Figure 4: VDG for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, and k = 6

The FDS plots for k = 3 show unique results. The unreplicated CCD option, C_1S_1, have the worst unscaled prediction variance irrespective of the number of centre points. Scaling the prediction variance gives a design undue advantage over the others due to the smaller number of runs compared to the replicated options. Hence, the design has the smallest scaled prediction variance for the entire design space with or without additional centre point. The star-replicated options display the best (smallest) unscaled prediction variance, which improves with additional centre point. For k = 4, 5 and 6, the star-replicated options display the smallest and most stable unscaled and scaled prediction variance throughout the entrie design space. The prediction variance of the star-replicated CCD options get better with scaling and with additional centre point. Hence, the star-replicated designs are recommended for prediction of responses involving practical CCD in the spherical region.

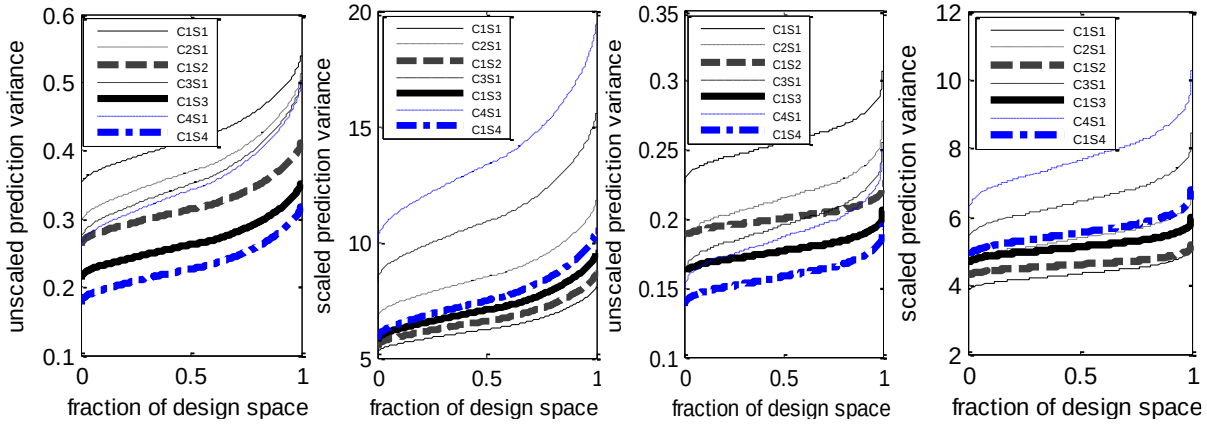
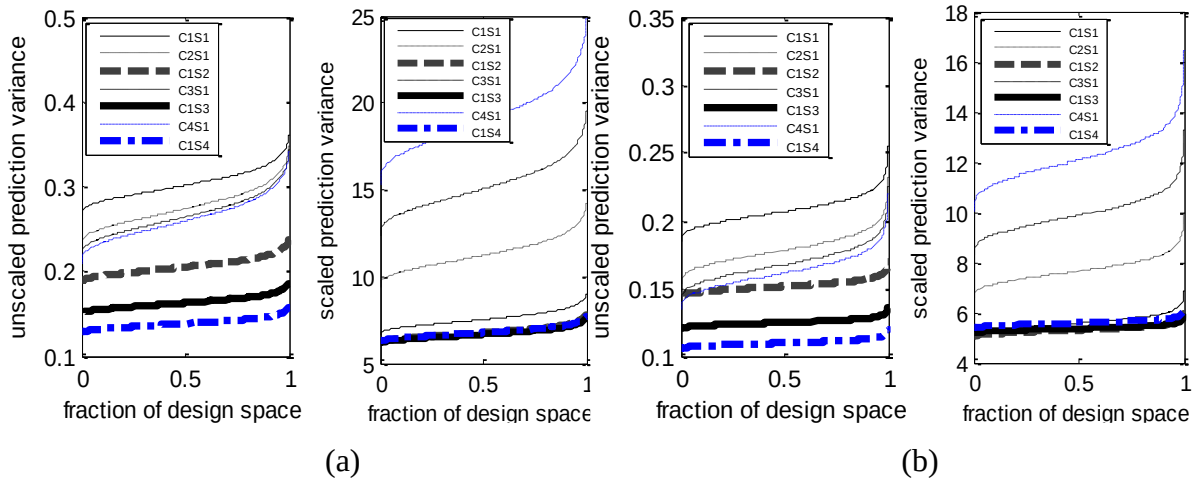
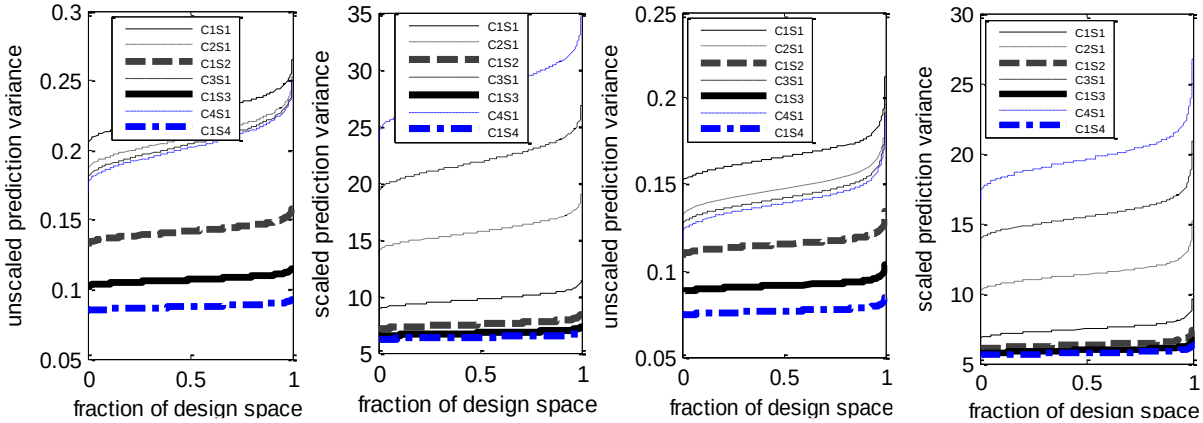


Figure 5: FDS Graphs for the Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, k = 3



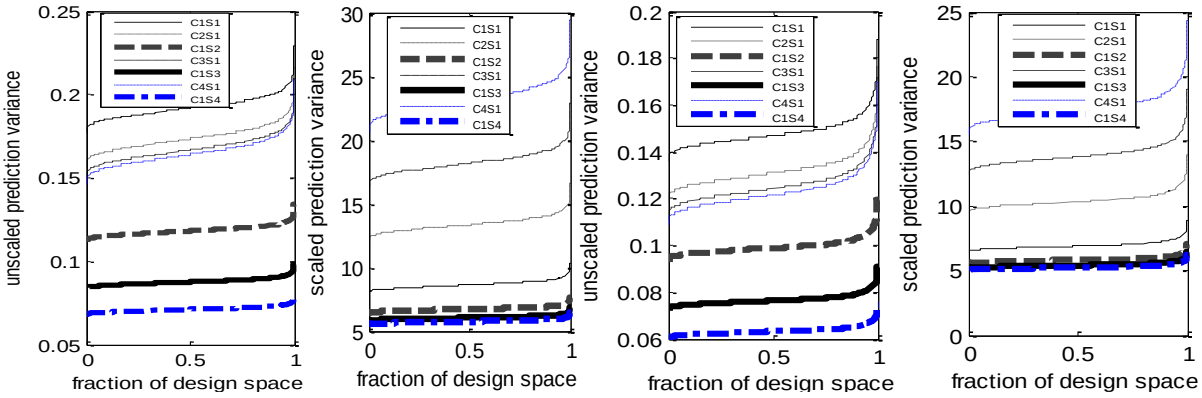
(a) (b)

Figure 6: FDS Graphs for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, $k = 4$



(a) (b)

Figure 7: FDS Graphs for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, $k = 5$



(a) (b)

Figure 8: FDS Graphs for Unscaled and Scaled Prediction Variance for (a) $n_0 = 1$ and (b) $n_0 = 3$, $k = 6$

4.3. Determination of Degrees of Freedom

According to Draper (1982) and Montgomery (2013), one of the reasons for replicating the centre point of a design was to obtain the required error degrees of freedom for the test of

hypothesis. The recommendation from Montgomery (2013) is a minimum of three degrees of freedom for model lack-of-fit and four degrees of freedom for pure error. Fewer, degrees of freedom will lead to a test that may not detect model lack-of-fit. In this section, we show that we can obtain the exact degrees of freedom required for pure error and test of model lack-of-fit by partially replicating the cube and sta portions of the CCD with practical axial distance and with or without replicating the centre point. According to Montgomery (2013), the total error sum of squares is calculated as the aggregate of the pure error sum of squares and the lack-of-fit sum of squares. Furthermore, the residual or total error degrees of freedom, RE_{df}, is calculated as the difference between the total number of runs, *N*, and the number of model parameters, *p*, being tested. That is:

$$RE_{df} = N - p. \tag{1}$$

The pure error degrees of freedom, PE_{df}, is the sum over all the replicates of each run. That is

$$PE_{df} = \sum_{i=1}^{N}\left(N_{reps,i} - 1\right) = \sum_{i=1}^{F}\left(F_{rep,i} - 1\right) + \sum_{i=1}^{H}\left(H_{rep,i} - 1\right) + n_0 - 1, \tag{2}$$

$$= (n_c - 1)f + 2(n_s - 1)k + (n_0 - 1) \tag{3}$$

where $N_{reps,i}$, = number of times the i^{th} run is replicated, $F_{rep,i}$ = the number of times the cube is replicated, $H_{rep,i}$ = the number of times the star is replicated, and $H = 2n_s k$. Therefore, any portion of the CCD that was not replicated does not contribute to the degrees of freedom. Therefore, the lack-of-fit degrees of freedom, LOF_{df}, is the difference between the residual degrees of freedom and the pure error degrees of freedom. That is,

$$LOF_{df} = RE_{df} - PE_{df}. \tag{4}$$

The exact pure error and lack-of-fit degrees of freedom for some variations of the partially replicated practical CCD are presented in Table 2 for *k* = 3 to 6 factors.

Table 2: Exact Pure Error and Lack-of-Fit Degrees of Freedom for *k* = 3 to 6 Factors

k	Design	*f*	n_c	*F*	n_s	*H*	*p*	$n_0 = 1$			$n_0 = 3$		
								N	LOF_{df}	PE_{df}	*N*	LOF_{df}	PE_{df}
3	C_1S_1	8	1	8	1	6	10	15	5	0	17	5	2
	C_2S_1	8	2	16	1	6		23	5	8	25	5	10
	C_1S_2	8	1	8	2	12		21	5	6	23	5	8

	C_3S_1	8	3	24	1	6		31	5	16	33	5	18
	C_1S_3	8	1	8	3	18		27	5	12	29	5	14
	C_4S_1	8	4	32	1	6		39	5	24	41	5	26
	C_1S_4	8	1	8	4	24		33	5	18	35	5	20
4	C_1S_1	16	1	16	1	8	15	25	10	0	27	10	2
	C_2S_1	16	2	32	1	8		41	10	16	43	10	18
	C_1S_2	16	1	16	2	16		33	10	8	35	10	10
	C_3S_1	16	3	48	1	8		57	10	32	59	10	34
	C_1S_3	16	1	16	3	24		41	10	16	43	10	18
	C_4S_1	16	4	64	1	8		73	10	48	75	10	50
	C_1S_4	16	1	16	4	32		49	10	24	51	10	26
5	C_1S_1	32	1	32	1	10	21	43	22	0	45	22	2
	C_2S_1	32	2	64	1	10		75	22	32	77	22	34
	C_1S_2	32	1	32	2	20		53	22	10	55	22	12
	C_3S_1	32	3	96	1	10		107	22	64	109	22	66
	C_1S_3	32	1	32	3	30		63	22	20	65	22	22
	C_4S_1	32	4	128	1	10		139	22	96	141	22	98
	C_1S_4	32	1	32	4	40		73	22	30	75	22	32
6	C_1S_1	64	1	64	1	12	28	77	49	0	79	49	2
	C_2S_1	64	2	128	1	12		141	49	64	143	49	66
	C_1S_2	64	1	64	2	24		89	49	12	91	49	14
	C_3S_1	64	3	192	1	12		205	49	128	207	49	130
	C_1S_3	64	1	64	3	36		101	49	24	103	49	26
	C_4S_1	64	4	256	1	12		269	49	192	271	49	194
	C_1S_4	64	1	64	4	48		113	49	36	115	49	38

5. Summary and Recommendations

Additional centre points improved the *A* and *V* for all the factors considered. The unreplicated CCD, C_1S_1, offers the best values for the *A* and *V* for k = 3 factors. Replicating the cube is beneficial only up to C_3S_1, the replications improved the performances of the designs' alphabetic criteria, beyond which alphabetic criteria begin to deteriorate. For k = 5 factors, the higher the replication of the star, the smaller the *V*, and the better the design; this characteristic is unique to k = 5 factors. Except for k = 3, the best value for *A* was obtained from the star-replicated CCD, C_1S_2, with or without an additional centre point.

In general, the higher the replication of the star portion of the CCD provided uniform distribution of both the unscaled and scaled prediction variances throughout the entire design space. This is true using both the variance dispersion graphs and fraction of design space graphs and for all the factors considered in the study.

References

Anderson, M., & Whitcomb, P. (2005). *RSM simplified: Optimizing processes using response surface methods for design of experiment*. Productivity Press.

Anderson-Cook, C. M., Borror, C. M., & Montgomery, D. C. (2009). Response surface design evaluation and comparison. *Journal of Statistical Planning and Inference*, *139*, 629–641.

Box, G. E. P., & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(1), 1–45.

Draper, N. R. (1982). Centre points in second-order response surface designs. *Technometrics*, *24*(2), 127–133.

Giovannitti-Jensen, A., & Myers, R. H. (1989). Graphical assessment of the prediction capability of response surface designs. *Technometrics*, *31*(2), 159–171.

Li, J., Liang, L., Borror, C. M., Anderson-Cook, C. M., & Montgomery, D. C. (2009). Graphical summaries to compare prediction variance performance for variations of the central composite design for 6 to 10 factors. *Quality Technology and Quantitative Management*, *6*(4), 433–449.

Montgomery, D. C. (2013). *Design and analysis of experiments* (8th ed.). John Wiley & Sons, Inc.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2009). *Response surface methodology: Process and product optimization using designed experiments* (3rd ed.). Wiley.

Ukaegbu, E. C. (2018). *A monograph on partial replications of some central composite designs: Design, evaluation and characterization*. Lambert Academic Publishing.

Zahran, A., Anderson-Cook, C. M., & Myers, R. H. (2003). Fraction of design space to assess prediction capability of response surface designs. *Journal of Quality Technology*, *35*(4), 377–386.

Unraveling the Complex Factors Influencing Insecurity and Crime Rates in Niger State, Nigeria: A Principal Component Analysis (PCA) Approach

Ahmed Sule Shehu, Usman Nuhu Mariga, and Saidu Umaru

Department of Mathematics and Statistics, Niger State Polytechnic, Zungeru

School of Natural and Applied Sciences (SNAS)

∗email: ahmedsule710@gmail.com, Tel: 08036154396

Abstract

Insecurity and crime rates have become pervasive issues in Niger State, Nigeria, threatening economic development, social harmony, and human well-being. This study employs Principal Component Analysis (PCA) to identify the underlying factors contributing to insecurity and crime rates in the region. Using a multivariate approach, we examine the relationships between socio-economic, political, environmental factors, insecurity and crime rates. The study utilizes a dataset consisting of 13 variables. Data obtained based on these factors reveal 77.15% total variation. Our results unveil the most significant predictors of insecurity and crime rates, including; Illiteracy rate, Moderate Household Incomes, Population and Drug Offenders Arrested. The study contributes to the existing literatures by providing a comprehensive understanding of the complex determinants of insecurity and crime rates in Niger State. The findings have important implications for policy makers, stakeholders, and researchers seeking to develop effective strategies for mitigating insecurity and crime rates in the region. This study suggested that, addressing the root causes of crime such as population, drug/arrest, domestic product and sex-ratio will mitigate the insecurity and as well as crime rate in the state.

Keyword: Crime Rate, Eigenvalues, Eigenvectors, Poverty Rate, Principal Components

1. Introduction

Multivariate analysis is a statistical method that simultaneously examines multiple variables and their relationships to identify patterns, correlations, and trends. Multivariate analysis involves analyzing; multiple dependent variables (outcomes), multiple independent variables (predictors) and interrelationships between variables. The goals is to identify patterns and correlations, understand complex relationships, predict outcomes, classify or cluster data and so on (Black and Babin, 2019)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into lower-dimensional data while retaining most of the information. It identifies orthogonal (uncorrelated) components. Explains variance in the data and reduces dimensionality. The goals are to simplify complex data, identify underlying patterns and reduce noise (Abdi and Williams, 2010).

Niger State faces a severe crime crisis, with one of the highest crime rates in the country. Between 2019 and 2022, the state witnessed numerous deadly attacks by bandits, particularly in Shiror, Rafi, Mariga, Munya, and neighboring Local Government Areas (LGAs), resulting in significant loss of life. Moreover, crime rates are disproportionately higher in rural areas compared to urban centers, poverty is identified as the primary driver of most crimes in Niger state. In response, various vigilante groups have formed to combat crime in affected areas, seeking to address the security vacuum and protect local communities.

Some of these assaults happened as domestic violence, while others are inflicted by criminals on guards specifically under the volatile situation (Oshunkeye, 2014).

It was observed that the highest and commonly committed crimes are assault, grievous harm and wounding, theft/stealing, burglary, housebreaking, false presence, unlawful arms possession, and breach of the peace (Femi *et al*, 2015).

Usman *et al* (2012), in their analysis carried out using NCSS and GESS 2007 Software. The outcome of the results shows that three Principal components have been maintained using the Scree plot and Loading plot giving that correlation exist between crimes against persons and crime against properties.

Nigeria record one of the highest crime rates in the world. Killing often accompanies minor burglaries; Influential Nigerians live in high-security apartments. Security men in some states are enforced to “short on sight” (financial times, 2011). No disagreement from both macro and micro side of studies that the rate of crime in Nigeria has reached an unacceptable level.

Atanu (2019), examined Nigeria's 2017 crime data, focusing on eight major crimes reported to the police: robbery, theft, house breakings, grievous harm and wounding, murder, rape, and assault. To uncover correlations between these crimes, the researchers employed principal component analysis (PCA). Their findings revealed a significant link between armed robbery, theft, and grievous harm and wounding. Kebbi state recorded the lowest crime rate. While, Lagos state topped the list with the highest overall crime rate. Crime Prevalence in Lagos state;

theft, house breakings, grievous hurt and wounding, murder, and rape were more common. In Rivers state; armed robbery was prevalent. Abia state; kidnapping and assault were more frequent. The PCA identified two primary components explaining approximately 85.166% of the data's total variability. Notably, the study concluded that some crime data can predict others, with an interesting correlation between kidnapping and murder.

However, from human and sociological effect of crime, there is a significant economic cost to the country in which rate of crime is high. Such economic consequences include decrease in labor market participation. Reduce productivity (Krug *et al,* 2012). The increase in urban crime rate in Nigeria is among the major social problems facing the country in recent time.

Chinyere *et al* (2024), in their study investigated the Nigeria Police Force and insecurity challenges in Rivers State, Nigeria. It explores the relationship between police inefficiency and insecurity, factors contributing to insecurity, and challenges facing the police force. The study employed a descriptive survey research design, selecting 400 respondents from three local government areas in Rivers State (Ogba/Egbema/Ndoni, Emohua, and Gokana) using the Taro Yamane formula. Data analysis utilized frequency, simple percentage, mean, and Spearman Rank Order Correlation. In their Findings.

Khamis and El-Refae (2020), their primary goals of this study were to identify the sources of variation in the dependent variables (DVs) and to determine which governorates exhibited similar or divergent trends for each DV. They employed a time-series and cross-sectional research design to achieve these objectives. Their main hypotheses posited that the independent variables (IVs) would have a collective and separate impact on the DVs. To test these hypotheses, the researchers conducted multivariate and univariate analyses of variance. The study population consisted of 12 governorates in Jordan, and they analyzed 15 years of data (2000-2015).

Ahmed *et al* (2023), carried out a research that employs principal component analysis (PCA) to investigate the causes of insecurity and crime rates in Niger state, Nigeria. The PCA results reveal that; total variation explained is 76.74%. Number of principal components (PCs) extracted is 3 out of 13 original variables. Extracted PCs are Population, Population density and Sex ratio. The findings suggest that population, population density, and sex ratio are key factors contributing to insecurity and crime rates in Niger State, Nigeria. These insights can inform policy decisions and interventions aimed at addressing these issues.

The situation confirmed from the report of the International Crime Victim Survey (ICVS) shows that, the report was conducted on six major continents for the 1989 – 1996 period that more than half of the urban respondents reported being victimized at least once regardless of what part of the world they inhabit (Ackerman and Murray, 2014).

Badiora and Aborisade (2024), this study investigates safety concerns among Lagos residents and their implications for individual and city-level security planning. Researchers randomly selected respondents from 49 wards, representing 20% of Lagos' 245 wards. At least two residents were chosen from each selected ward, resulting in 288 participants. However, only 134 respondents accurately completed the survey, yielding a 58.7% response rate. Results show that 78% of the respondents were concerned about their safety while some 90% took precaution by avoiding certain places and activities.

This study aims to achieve three primary objectives; identify the drivers of insecurity, examine the effects of insecurity and proffer measures to address insecurity.The study adopts the strain theory framework and utilizes secondary data, which were analyzed through content analysis. Findings: The study reveals that several factors have contributed to the increasing rate of insecurity in Nigeria, including; porous borders, proliferation of arms, bad governance, marginalization, poverty, unemployment, religious fanaticism, inadequate security architecture and weak security system (Odalonu, 2022)

Agbor *et al* (2022): This study compares major crimes in Calabar South and Calabar Municipality, two local government areas in Cross River State, Nigeria. The regions have seen rising crime rates due to rapid population growth, rural-urban migration, and urbanization. To gather data, questionnaires were administered to 120 residents in each locality through random sampling. Oral interviews and direct observations identified crime hotspots, road networks, and neighborhood designs. Twenty crime-prone neighborhoods were randomly selected for further analysis using a parametric scoring method. The study revealed that; higher youth involvement in crime in both areas. Disparities in education, occupational status, monthly income, and household size. Calabar South had higher crime rates than Calabar Municipality, according to the independent t-test.

This study applied Principal Component Analysis (PCA) to examine the spatial patterns of criminal victimization across 11 local government areas (LGAs) in Katsina Senatorial Zone, Katsina State, Nigeria. Data collection involved face-to-face interviews using a stratified multistage random sampling procedure. The dataset comprised eight crime categories;

Burglary, Robbery, Theft, Fraud, Assault, Murder, Thuggery and Rape. Logarithmic transformation normalized the victimization data, facilitating spatial distribution analysis. The study revealed that; Batsari recorded the highest overall average victimizations. Rimi had the lowest average victimizations. LGAs along the Katsina-Zamfara border, particularly around Rugu forest andusufazing reserve exhibited increased violence and property crime victimizations, including: Murder, Robbery, Rape, Burglary, and Livestock theft (Yusuf *et al*, 2014).

Zubair, (2023): In his study that investigated economic and psychological factors driving crime rates in Sahiwal, Pakistan, and examined the role of law enforcement and societal responses in mitigating these issues. Purposive sampling technique was used and case studies of 60 criminals, Structured, questionnaires and interviews was adopted. Data analysis using SPSS and Smart PLS software. Descriptive and inferential statistics (chi-square, gamma). Expected Outcomes: Identification of factors influencing crime rates, Understanding crime's impact on individuals and public life, informing strategies for a crime-free society. This research aims to contribute to evidence-based policy-making, enhancing crime prevention and community safety in Sahiwal, Pakistan.

This study employed Principal Component Analysis (PCA) on Nigeria's crime data to identify key factors influencing state-level safety. Four Principal Components, explaining 75.024% of the total variation, were retained using scree plot and Kaiser's criterion, revealing critical variables for targeted crime prevention strategies (Dikko *et al*, 2013).

Chris *et al* (2019), in their research conducted to identify the major causes and effects of crimes in South-Western Nigeria. It was discovered that the once peaceful region of the country is gradually becoming the crime capital, losing lives and properties worth millions on yearly basis to the recurrent cases of crimes. Data were generated from both primary and secondary sources and tested with Pearson Product Moments Correlation Coefficient at 0.01 level of significance, through the use of SPSS. Findings revealed that there is a positive correlation between unresponsive government and increase crimes in South-Western Nigeria which has led to business shutdown. The study concludes that if adequate measures to curb these crimes are not urgently taken, more lives will be lost.

This study investigates the societal factors that contribute to insecurity in Nigerian society, utilizing the Autoregressive Distributed Lag (ARDL) co-integration technique to analyze data from 1991 to 2022. The ARDL co-integration technique is employed to analyze the

relationships between these societal factors and insecurity in Nigeria over the 31-year period (Omodero, 2024).

Guobadia, (2021): Analyzing Nigeria's crime data from 2009 to 2019 reveals interesting trends. The data, comprising averages of 50 crimes reported to the police, includes offenses like robbery, car theft, breakage, theft, serious accidents, murder, rape, and assault. Correlation analysis and Principal Component Analysis (PCA) helped clarify relationships between crimes and their distribution across Nigeria. The study revealed that; no significant correlation among crimes against persons. Moderate to strong correlations among Crimes against Property. Significant correlations among murder, armed robbery, grievous harm, and wounding. Significant correlations among assault, kidnapping, attempted murder, theft, store breaking, house breaking, and escape from lawful arrest. PCA retained six components explaining 95.24% of data variability.

Umar, (2019), conducted a study in Suleja to examine the level of security consciousness, safety, and coping strategies against crime among residents. The findings revealed that nearly 47% of residents experienced some form of crime in the past year, with burglary being the most prevalent (49.4%) followed by armed robbery (28.8%). Causes of Crime include; Drug Abuse (35%), Unemployment (39.2%), Poverty (23.5%), Others(2.3%).

This theoretical paper aims to demonstrate that insecurity in Nigeria is a complex, multi-causal phenomenon that cannot be attributed solely to economic weakness. Instead, it argues that insecurity stems from systemic deficiencies in various areas, including; Geography, Politics, Socioeconomic, Military, Religion, World technology and External influences. The paper utilizes an analytical framework inspired by interstate warfare studies and relies on documentary sources of information to support its arguments (Esetang and Atai, 2023).

In a study to identify the statistical relationship between crime and socio-economic status in Ottawa and Saskatoon, the PCA was used to replace a set of variable with a smaller number of components, which are made up of inter-correlated variable representing the original data set as possible (Exp, 2008).

Principal component analysis can also be used to examine the overall criminality. When the first eigenvector shows approximately equal loadings on all variables than the first PC measure the overall crime rate. In Printcom (2003) for 1997 US crime data, the overall crime rate was

examined from the first PC, and the same outcome was achieved (Hurdle *et al*, 2007) for the 1985 US crime data.

According to Olufolabo *et al* (2015), in Oyo State's crime data from 1996-2014, focusing on 18 major crimes reported to the police, including robbery, kidnapping, theft, murder, rape, assault, and others. Correlation analysis and Principal Component Analysis (PCA) were used to identify relationships between crimes and their distribution in the state. The results show a strong correlation between Store breaking and Stealing, Positive relationships between Assaults and False Pretence/Breach of Peace, Kidnapping and False Pretence/Breach of Peace. Weak positive relationships between Murder and various crimes, including Armed Robbery and Suicide. Positive relationships between Murder and Rape/Arson, and between Armed Robbery and Kidnapping. Stealing/Theft and Assaults are the most prevalent crimes in Oyo State. The PCA suggests retaining six components that explain approximately 83.79% of the data's total variability.

Despite government efforts to curb criminal activities, insecurity and crime rates continue to rise in Ibadan, causing widespread concern. This study investigates the state of insecurity and crime within the Ibadan community and its surroundings. The research employed a mixed-methods approach, combining: Questionnaires administered to University of Ibadan students and residents within and outside the campus. Data from the Nigeria Prisons Service, Agodi Prison Division. The study utilized: Descriptive and Inferential statistics, Time series analyses, including stationary tests, model identification, estimation, evaluation, and forecasting. The results revealed a significant relationship between the causes and potential solutions to insecurity in Ibadan at a 5% significance level (p= 0.000). The study also established dynamic models to predict crime rates for 2017-2020, indicating an increase in insecurity and crime over time. Factors contributing to this trend may stem from both government and individual actions. (Saheed, 2019).

The main objectives of this study is to identify the critical factors contributing to insecurity and crime rates in Niger state. Also, to examine the relationship between socioeconomic factor, poverty, and crime rate.

Research gap: The researcher identifies some risk factors associated with insecurity and crime rates in Umar (2019), but the study is limited to part of Niger state. While, in this stud we take cognizant of the entire state.

2. Materials and Methods

This study utilizes a dataset consisting of 13 variables, aggregated into four broad categories: demographic, economic, social, and housing characteristics, in addition to police personnel capacity. It is important to acknowledge that cultural and transportation factors, known to impact crime rates, are not included in this analysis. Moreover, some metrics commonly applied at national and state levels are not feasible or relevant for Local Governments in Niger State, Nigeria.

We collected data of poverty rate, unemployment rate, average household size, youth illiteracy rates and percentage moderate income household livelihood from State Bureau of Statistics (Annual Report 2020), population, population density, sex ratio from the National Population Commission (2006 census), GDP(PPP) from Niger State Ministry of Finance, drug arrest and seizure/arrest index(SAI) from National Drug Law Enforcement Agency (Annual report 2019), 2023 General election percentage voters turnout from Independent National Electoral Commission (INEC). Crime record and police strength from Nigeria Police Headquarters, Minna. We used crime record from 2015 because this is the only recent years for which we could find complete data. The dataset totally has 25 entries; each entry represents the information of a particular Local Government in Niger State.

List of Variables

POP => Population, **DENS** => Population Density, **SEX** => Sex Ratio, **AVRG** => Average Household Size, **ILLT** => Illiteracy Rate, **MOD**=> Moderate Household Incomes, **POR** => Poverty Rate, **UNR** => Unemployment Rate, **VTO** => Voters Turnout, **POL** => Police Strength, **GDP** => Growth Domestics Product, **OAR** => Drug Offenders Arrested, **SAI** => Drug Seizure Arrested Index

Principal Component Analysis

The study will utilize Principal Component Analysis (PCA) to perform an orthogonal transformation, thereby converting the initial set of correlated variables into a set of linearly independent components. This dimensionality reduction technique will facilitate the identification of the underlying variance-covariance structure, revealing the most informative linear combinations of the original variables.

Principal Component Analysis (PCA) generates an orthogonal set of variables, known as principal components, which capture the majority of the information present in the original data. These components are ordered by their explanatory power, with the first few components

retaining most of the variation. Notably, PCA differs from Factor Analysis in its deterministic nature, yielding identical solutions from the same data, aside from potential sign reversals.

Let X be a vector of p random variables, the idea of principal component transformation is to look for few variables less than p derived variables that preserved most of the information given by the variance of the p random variables.

Consider a random vector $X' = X_1, X_2, X_{3,} - - -, X_p$ that have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_p \geq 0$. let the linear combinations be;

$$Y_j = \alpha_j' X = \alpha_{j1}X_1 + \alpha_{j2}X_2 + \alpha_{j3}X_3 + \cdots + \alpha_{jp}X_p$$
$$= \sum_{k=1}^{p} \alpha_{jk}X_p \qquad 1$$

Such that $j = 1, 2, \ldots, p$ are the elements of $X \; and \; \alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \ldots, \alpha_{jp}$ are the components of $\alpha_j$vector of p^{th} term.

Then, $Var(Y_j) = \alpha_j' \sum a_j \qquad j = 1, 2, \ldots, p \qquad 2$

$Cov(Y_J . Y_k) = \alpha_j' \sum a_k \qquad j = 1, 2, \ldots, p \qquad 3$

Principal Component Procedure

The principal components are those uncorrelated linear combinations $Y_1, Y_2, Y_3, \ldots, Y_p$ whose variances in (2) are as large as possible. In finding the PCs we concentrate on the variances. The first step is to look for a linear combination $a'X$ with maximum variance, so that

$\alpha_1' X = \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + \cdots + \alpha_{1p}X_p = \sum_{k=1}^{p} \alpha_{1k}X_p \qquad 4$

Then, we look for the linear combination $\alpha_2' X$ uncorrelated with $\alpha_1' X$ having maximum variance and so on, hence at the k^{th} stage a linear combination $\alpha_k' X$ is found that has maximum variance subject to being uncorrelated with $\alpha_1' X$, $\alpha_2' X$, $\alpha_3' X$, …, $\alpha_{k-1}' X$. The k^{th} derived variable $\alpha_k' X$ is the k^{th} principal component. Up to p principal components can be found, but we would hope to stop after the q^{th} stage for ($q \leq p$), i.e. when most of the variation in X would have been accounted for by q^{th} principal components.

Note:

- The variance of a principal component is equal to the eigenvalue corresponding to that principal component,

$Var(Y_j) = \alpha_j' \sum a_j = \lambda_j \qquad j = 1, 2, 3, \ldots, p$

- The total variance in data set is equal to the total variance of principal components

$$\sigma_{11} + \sigma_{22} + \sigma_{33} + \cdots + \sigma_{pp} = \sum_{j=1}^{p} Var(X_j) = \lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p$$

$$= \sum_{j=1}^{p} Var(Y_j)$$

The data would be standardized for the variables to be of similar scale using a common standardization method of transforming all the data to have zero mean and unit standard deviation. For a random vector $X' = [X_1, X_2, X_3, \ldots, X_p]$ the corresponding standardized variables are $z = \left[Z = \frac{(X_j - \mu_j)}{\sqrt{\sigma_j}}\right]$ for $j = 1, 2, 3, \ldots, p$ in matrix notation, $Z = \left(\theta^{1/2}\right)^{-1}(X - \mu)$, Where $\theta^{1/2}$ is the diagonal standard deviation matrix and it's a unit.

Thus, $E(Z) = 0 \; and \; Cov(Z) = \rho$.

The PCs of Z can be obtained from eigenvectors of the correlation matrix ρ of X. All our previous properties for X are applied for the Z, so that the notation Y_j refers to the j^{th} PC and (λ_j, a_j) refers to the eigenvalue – eigenvector pair. However, the quantities derived from Σ are not the same from those derived from ρ (Richard and Dean, 2001).

The j^{th} PC of the standard variables $Z' = [z_1, z_2, z_3, \ldots, z_p]$ with $cov(Z) = \rho$, is given by

$Y_j = \alpha_j' Z = a'\left(\theta^{1/2}\right)^{-1}(X - \mu)$ 5

So that $\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p}(Z) = \rho \qquad for \; j = 1, 2, 3, \ldots, p$

In this case, (λ_1, α_1), (λ_2, α_2), (λ_3, α_3), …, (λ_p, α_p) are the eigenvalue - eigenvector pairs for ρ with $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \lambda_p \geq 0$.

The loading or the eigenvector $\alpha_j = \alpha_1, \alpha_2, \alpha_3, \ldots \alpha_p$ is the measure of the importance of a measured variable for a given PC. When all elements of α_1 are positive, the first component is a weighted average of the variables and is sometimes referred to as measure of overall crime rate. Likewise, the positive and negative coefficients in subsequent components may be regarded as type of crime components. The plot of the first two or three loadings against each other enhances visual interpretation.

The score is a measure of the importance of a PC for an observation. The new PC observations Y_{ij} are obtained simply by substituting the original variables X_{ij} into the set of the first q PCs. This gives

$$Y_{ij} = \alpha'_{j1}X_{i1} + \alpha'_{j2}X_{i2} + \alpha'_{j3}X_{i3} + \cdots + \alpha'_{jp}X_{ip} \qquad i = 1, 2, 3, \ldots, p \qquad j = 1, 2, 3, \ldots, p$$

The plot of the first two or three PCs against each other enhances visual interpretation.
The proportion of variance will be used to tell us the PC that best explained the original variables and the measure of how well the first q PCs of Z explain the variation is given by:

$$\phi_q = \frac{\sum_{j=1}^{q} \lambda_j}{p} = \frac{\sum_{j=1}^{q} V(Z_j)}{p} \qquad 6$$

A cumulative proportion of explained variance is a useful criterion for determining the number of components to be retained in the analysis. A Scree plot provides a good graphical representation of the ability of the PCs to explain the variation in the data.

3. Results and Discussion

The analysis focuses on two primary crime categories: Crimes against Persons; Murder, Grievous Harm and Wounding (GHW), Assault, Kidnapping and Rape. And Crime against Property; Armed Robbery, Theft and House Breaking.

Table 1: Correlation Matrix

	POP	DENS	SEX	AVRG	ILLT	MOD	POR	UNR	VTO	POL	GDP	OAR	SAI
POP	1												
DENS	0.2402	1											
SEX	0.2251	0.0797	1										
AVRG	0.1184	0.1105	-0.1469	1									
ILLT	-0.0182	-0.6758	-0.1524	0.0645	1								
MOD	0.2468	0.1954	0.2056	-0.3901	-0.1337	1							
POR	-0.1130	-0.0683	-0.0985	0.3681	-0.0244	-0.7358	1						
UNR	0.6169	0.1810	0.0284	0.1119	-0.0114	-0.0479	-0.0661	1					
VTO	-0.7870	-0.1623	-0.0826	-0.0318	-0.0830	-0.0508	0.0402	-0.5655	1				
POL	1.0000	0.2394	0.2242	0.1198	-0.0172	0.2468	-0.1125	0.6187	-0.7866	1			
GDP	1.0000	0.2402	0.2252	0.1184	-0.0183	0.2468	-0.1130	0.6169	-0.7870	1.0000	1		
OAR	0.9956	0.2650	0.2065	0.1329	-0.0448	0.2042	-0.0693	0.6269	-0.7863	0.9954	0.9956	1	
SAI	0.9996	0.2495	0.2272	0.1175	-0.0328	0.2455	-0.1134	0.6276	-0.7855	0.9996	1.0000	0.9960	1

Table 1: The correlation matrix reveals varying degrees of correlation between the variables. Specifically; Weak correlations exist between Population and illiteracy rate/poverty rate. Population density and sex ratio/average household size/poverty rate. Strong correlations also observed between Population and unemployment rate. Population and police strength. Population and growth domestic product (GDP). Population and drug-related variables (offenders arrested, seizure/arrest index). The correlation matrix's determinant is 0.000, indicating no multicollinearity issues. Therefore, no variables need to be eliminated at this stage.

Table 2: Kaiser-Meyer-Olkin (KMO)

POP	DENS	SEX	AVRG	ILLT	MOD	POR	UNR	VTO	POL	GDP	OAR	SAI	
0.71757	0.47066	0.66013	0.54955	0.27992	0.51947	0.37453	0.66034	0.84361	0.85567	0.71797	0.87591	0.91409	0.73154

Table 2: The Kaiser-Meyer-Olkin (KMO) statistic assesses sampling adequacy for Principal Component Analysis (PCA). With values ranging from 0 to 1, Kaiser (1970) recommends values ≥ 0.5. Our KMO value, 0.73154, meets this criterion, indicating sufficient data quality for PCA.

Fig. 1: Scree Plot of the Principal Components

Table 3: Eigenvalues of the Correlation Matrix Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	eValue	% of variance	% of Cumm	eValue	%	Cum %
1	6.297298	48.44%	48.44%	6.297298	48.44%	48.44%
2	2.107213	16.21%	64.65%	2.107213	16.21%	64.65%
3	1.624447	12.50%	77.15%	1.624447	12.50%	77.15%
4	0.959240	7.38%	84.52%			
5	0.726330	5.59%	90.11%			
6	0.552635	4.25%	94.36%			
7	0.336252	2.59%	96.95%			
8	0.227712	1.75%	98.70%			
9	0.164550	1.27%	99.97%			
10	0.003897	0.03%	100.00%			
11	0.000389	0.00%	100.00%			
12	2.74E-05	0.00%	100.00%			
13	9.06E-06	0.00%	100.00%			

Table 3: The first and very important step of PCA is to determine the number of PCs. Here, the Kaiser's criterion is used to do the selection which is accurate when there are less than 30 variables and communalities after extraction are greater than 0.7. Therefore based on Kaiser's rule we retain all PCs with eigenvalues greater than 1, which leaves us with first three PCs that explain up to 77.15% of the total variability of the data set as shown in the Table 3. We can also use the scree plot to choose the number of PCs to retain, as shown in Figure 1 above, the

“elbow" appears at the 4th PC, therefore the first three PCs should be retained which account for 77.15% of the total variance, as a valuable reduction in dimensionality.

Therefore, it is a fact that the first three PCs accounted for 77.15% of total variance of the original variables and simultaneously reduce the data dimension from 13 to 3.

Table 4: Factor Matrix (rotated varimax)

	1	2	3
POP	0.9888	0.0738	-0.0368
DENS	0.1858	0.2939	-0.8338
SEX	0.1200	0.3466	-0.0910
AVRG	0.1019	-0.6282	-0.2884
ILLT	0.0164	-0.3370	0.8603
MOD	0.1832	0.8764	0.1353
POR	-0.0486	-0.7954	-0.3026
UNR	0.5970	-0.1150	-0.0629
VTO	-0.8141	0.0617	-0.0512
POL	0.9887	0.0727	-0.0362
GDP	0.9888	0.0738	-0.0368
OAR	0.9889	0.0395	-0.0775
SAI	0.9876	0.0768	-0.0493
	6.0009	2.1560	1.6557

The next step is to look at the content of variables that load onto the same factor to try identifying common themes. In Table 4, we visualize the coefficients of the variables in each PC and combined with the exact values of these coefficients (also known as loading Matrix of PCA). It can be seen clearly which variables are dominant in each PC. For example in PC2, Moderate Household Income have relative large absolute value, hence PC2 denotes as more economical factor. Similarly, view PC3 as Illiteracy rate. However, it is kind of difficult to define PC1 in which Population, Police, GDP, Offender Arrested and Seizure Arrested all have relative large absolute value of greater than 0.7 level, as such we assume the PC with the highest value (ie Offender Arrested).

4. Conclusion

Principal Component Analysis (PCA) successfully applied to our dataset, resulting in a significant dimensionality reduction from thirteen original variables to three principal components (PCs). These PCs capture a substantial portion of the variance in the original dataset, ensuring minimal information loss. The PCs can be labeled based on highly correlated

original variables: PC2 strongly associated with Moderate Household Income, indicating socioeconomic factors. Illiteracy rate in PC3 and PC1 contribute from various original variables with Offender Arrest recording highest significant impact, suggesting complex interactions. This study suggested that, addressing the root causes of crime such as population, drug/arrest, domestic product and Illiteracy will mitigate the insecurity and as well as crime rate in the state.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459.

Ackerman, W. V., & Murray, A. T. (2014). Assessing spatial patterns of crime in Lima, Ohio. *Cities, 21*(5), 432–437.

Agbor, E. A., & Eyo, E. I. (2022). Comparative analysis of major crimes in Calabar Municipality and Calabar South Local Government Areas, Cross River State, Nigeria. *International Journal of Research and Advanced Studies, 9*(2), 102–111.

Ahmed, S. S., Usman, N. M., Jibrin, M., Saidu, U., & Ahmed, R. T. (2023). Use of principal component analysis for evaluation of causes of insecurity and crime rate investigation in Niger State, Nigeria. *Journal of Applied Sciences and Environmental Management, 27*(5), 1003–1008.

Atanu, E. Y. (2019). Analysis of Nigeria's crime data: A principal component approach using correlation matrix. *International Journal of Scientific Research Publications, 9*, 1–8.

Badiora, A. I., & Aborisade, R. A. (2024). Insecurity and terrorists' threats of attacks on selected Nigerian cities: Implications on Lagos residents' concerns for safety and the city security planning. *UCC Law Journal, 3*(2), 112–154.

Black, W., & Babin, B. J. (2019). Multivariate data analysis: Its approach, evolution, and impact. In J. F. Hair Jr. (Ed.), *The great facilitator: Reflections on the contributions of Joseph F. Hair, Jr. to marketing and business research* (pp. 121–130). Springer International Publishing.

Chinyere, I. E., & Terna, J. H. (2024). Nigeria Police Force and the challenges of insecurity in Nigeria: A study of Rivers State, Nigeria. *Bureaucracy Journal: Indonesia Journal of Law and Social-Political Governance, 4*(1), 70–92.

Chris, N. I., & Etim, E. E. (2019). Identifying major causes, effects and approaches to curbing crimes in South-Western Nigeria. *International Journal of Scientific Research and Engineering Development, 2*(3), 883–895.

Dikko, H., Osi, A. A., & Bello, Y. (2013). Detecting factors impacting crime rates in Nigeria using principal component analysis. *IOSR Journal of Mathematics, 9*(3), 8–17.

Esetang, A., & Atai, A. J. (2023). Nigeria and insecurity: Theoretical analysis of internal characteristics and causal factors. *International Journal of Culture and Society, 1*(2), 1–24.

Exp. (2008). *Canada's justice research and statistics*. Department of Justice Canada. http://www.justice.gc.ca/eng/pi/rs/rep-rap/2006/rr06/6/p5.html

Femi, J. A., Adeyemi, M. A., & Jabaru, S. O. (2015). On the estimation of crime rate in the southwest of Nigeria: Principal component analysis approach. *Unpublished manuscript*.

Financial Times. (2011, April 30). *London edition*.

Guobadia, K. E. (2021). Econometric modeling of Nigeria's GDP: A variable selection approach. *Scholar Journal of Economics and Business Management, 1*, 14–25.

Hurdle, W., & Zdeněk, H. (2007). *Multivariate statistics: Exercises and solutions*. Springer-Verlag.

Khamis, F., & El-Refae, G. (2020). Applying multivariate and univariate analysis of variance on socioeconomic, health, and security variables in Jordan. *Statistics, Optimization & Information Computing, 8*(2), 386–402.

Krug, E., Dahlberg, L. L., & Mercy, J. A. (2012). *Access Continuing Education*. https://www.accesscontinuingeducation.com

Odalonu, B. H. (2022). Worsening insecurity in Nigeria and its implications on governance and national development. *African Journal of Humanities and Contemporary Education Research, 5*(1), 31–49.

Olufolabo, O. O., Adejoh, A. J., & Ekum, M. I. (2015). Analyzing the distribution of crimes in Oyo State (Nigeria) using principal component analysis. *IOSR Journal of Mathematics Ver. I, 11*(3), 22–28.

Omodero, C. O. (2024). Social factors associated with insecurity in Nigerian society. *Societies, 14*(6), 74.

Oshunkeye, S. (2014, August 25). Africa journalism in culture of brown envelopes [Paper presentation]. Media Project's Regional Conference, Accra, Ghana. Retrieved from J. vs brown envelops.

Printcom. (2003). *SAS 9.1 online documentation*. http://support.sas.com/onlinedoc/912/getDoc

Richard, A. J., & Dean, W. W. (2001). *Applied multivariate statistical analysis* (3rd ed.). Prentice-Hall.

Saheed, A. A. (2019). Statistical analysis of insecurity and criminal cases in Ibadan, Oyo State, Nigeria. *CAL, 2*(2), 0–5.

Umar, A. (2019). *Assessment of crime proliferation and prevention strategies for sustainable development of Suleja, Niger State, Nigeria* (Doctoral dissertation). [Institution not specified].

Usman, U., Yakubu, M., & Bello, A. Z. (2012). An investigation on the rate of crime in Sokoto State using principal component analysis. *Nigeria Journal of Basic and Applied Sciences, 20*(2), 152–160.

Yusuf, B., Batsari, Y. U., & Charanchi, A. S. (2014). Principal component analysis of crime victimizations in Katsina Senatorial Zone. *International Journal of Science and Technology, 3*(4), 192–202.

Zubair, H. (2023). Factors of crime boosting the crime ratio in Sahiwal (Pakistan). *Web of Scientists and Scholars: Journal of Multidisciplinary Research, 1*(1), 29–47.

The truncated Schröter recursive algorithm for computation of aggregate claim amounts

Friday I. Agu

Mathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 81743 Bratislava, Slovakia
&
Department of Statistics, University of Calabar, Calabar, Nigeria

Corresponding Author's email: agu1@uniba.sk

Abstract

The profitability of insurance companies relies on accurately pricing premiums and aggregate claim amounts. However, the computation of aggregate claim amount becomes challenging when the number of claims and claim severities is large, especially when there are claim events. Traditional convolution methods, while theoretically sound, are complex and computationally intensive. This study introduces and evaluates a truncated Schröter recursive algorithm for computing aggregate claim amounts in the insurance sector. The algorithm addresses limitations in existing methods by incorporating truncation at 1, which is crucial for accurately modeling insurance claims where events leading to a claim are pivotal. Using the AutoCollision dataset, the study compares the truncated Schröter algorithm with Panjer and Schröter recursion algorithms, focusing on computational efficiency and accuracy. Furthermore, the descriptive statistics of the data reveal a wide range of claim amounts, with a mean of 279.44 and a high variance of 58374.38, indicating substantial variability. The dataset exhibits positive kurtosis and skewness, signifying the presence of outliers and a right-skewed distribution. Analysis of average claim severity by vehicle use shows that business-use vehicles have the highest average claim severity (395.21) despite a lower claim count, indicating a higher financial impact per claim than others. Results demonstrate that the truncated Schröter algorithm substantially reduces execution time while maintaining high accuracy, making it a superior tool for aggregate claim amount computation, risk management, and premium setting.

Keywords: Insurance claim amounts, aggregate claim distribution, recursive algorithm, insurance risk management, computational efficiency

1. Introduction

In the insurance domain, profitability hinges on the delicate balance between premium payments received and the claim amounts disbursed to policyholders. Unlike other market sectors, such as manufacturing, setting the appropriate premium for insurance portfolios is a complex task. This process necessitates accounting for future uncertainties to ensure sufficient

and continuous investment. Insurance companies rely on models that accurately compute aggregate claim amounts within a collective risk framework, aiming to predict whether the total claims will stay within a specified threshold. Typically, insurers begin by estimating an expected cost to determine a reliable premium, adding margins to cover uncertainties, company profits, and potential claim payouts (Yartey, 2020).

The distribution of aggregate claim amounts, which results from the convolution of the number of claims and claim severities, plays a pivotal role in pricing insurance portfolios. This distribution is crucial for determining the probability that the total claims will exceed a specific limit, influencing both premium setting and risk management strategies. Consequently, insurers strive for precise calculations of aggregate claim amounts to make informed decisions regarding competitive pricing and to ensure adequate margins for future uncertainties. However, one of the challenges in actuarial mathematics lies in modeling the distribution of aggregate claims, especially when dealing with non-negative integers and discrete random variables representing the number of claims and claim severities.

Historically, actuaries relied on approximation methods without a strong theoretical foundation to estimate aggregate claim amounts. As computational tools have evolved, so have the methodologies for modeling these distributions. Traditional approaches often involve selecting appropriate counting distributions for the number of claims and fitting them separately to claim severities, which are typically modeled using continuous distributions. Despite the progress made in this area, as highlighted by various studies (e.g., Hogg & Stuart, 2009; Gray & Pitts, 2012; Packova & Brebera, 2015), the conventional methods do not always adequately capture the convolution of claim number and severity distributions—an essential aspect of aggregate claim distribution.

The computation of aggregate claim amounts has traditionally relied on convolution, which, while theoretically sound, is computationally intensive and complex. As a result, alternative methods have been developed, including the normal power and fast Fourier transform techniques (Beard et al., 1977; Cooley & Tukey, 1965; Heckman & Meyers, 1983). These methods, though useful, have limitations in accuracy and become increasingly complex as the volume of claims and severity of distributions grow. This complexity has driven researchers to explore more robust methods for computing aggregate claims, leading to the development of the recursive method, or the "exact method," which computes aggregate claims recursively with less computational time (Sundt & Vernic, 2009).

A substantial advancement in this area was introduced by Panjer (1981), who developed the Panjer recursive family of discrete distributions, providing a recursion formula that simplified the computation of aggregate claims. This formula has been widely studied and applied (Sundt & Vernic, 2009; Yartey, 2020; Dickson, 2016; Ghinawan et al., 2021). However, its application is limited to certain counting distributions with fixed positive probabilities at zero. To address these limitations, Schröter (1990) introduced a recursive formula that extends the range of counting distributions and offers a more accurate representation of aggregate claims dynamics. However, this method still involves complex convolution components, and its computational demands increase substantially with the number of claims and claim severity.

The accurate calculation of aggregate claim amounts is crucial not only for risk management but also for the strategic pricing of insurance coverage. Insurers are driven by the need to minimize claims while maximizing premium income, ensuring they can withstand future uncertainties and potentially catastrophic losses. The competitive nature of the insurance industry, coupled with the unpredictability of claim events, further underscores the importance of precise and efficient modeling techniques.

Despite the Schröter recursive formula's advantages, it does not fully account for situations where claim amounts are truncated at 1—a common practice in insurance when the focus is on the number of events leading to a claim rather than the claim severity itself. Truncated distributions are vital in risk theory, particularly for modeling claim severity and the intervals between successive claims. These distributions provide insurers and actuaries with insights into the frequency and severity of potential losses, facilitating more accurate policy pricing and risk management.

Given these considerations, there is a pressing need for a method that can efficiently handle datasets characterized by high variance and low-frequency events, which are typical in insurance claims data. The Panjer recursion, while efficient for certain distributions, lacks flexibility in handling truncated data, and FFT methods can become computationally prohibitive for large datasets. In contrast, a truncated version of the Schröter recursion offers a more tailored approach, simplifying the computational process without sacrificing accuracy. This is particularly beneficial in environments where insurers prioritize events that trigger claims.

The primary contribution of this study is the introduction and evaluation of a truncated version of the Schröter recursive algorithm. This innovation aims to compute aggregate claim amounts

with reduced computational complexity while maintaining high accuracy, considerably improving existing methods. This advancement is crucial for the insurance industry, where handling large, complex datasets efficiently is essential for risk management, premium setting, and overall financial stability. The study leverages the Panjer and Schröter recursive algorithms as foundational elements to develop the truncated Schröter algorithm, focusing on enhancing accuracy and computational efficiency. By demonstrating the practical benefits of this algorithm, particularly in terms of reduced run time, this research offers valuable insights for the insurance industry.

2. The Recursive Formulas

2.1. The Panjer recursive formula

The Panjer (1981) recursive formula is defined as

$$P_k = \left(a + \frac{b}{k}\right) P_{k-1}, k = 1, 2, 3, \ldots \tag{1}$$

where a and b are parameters and, by definition, $P_k = 0$ for $k < 0$. The counting distributions that satisfied (1) were explored in Panjer (1981). Furthermore, Panjer (1981) obtained the corresponding recursion algorithm for (1) defined as:

$$g(s) = \frac{1}{1 - af_0} \sum_{i=1}^{s} \left(a + \frac{bi}{s}\right) f_i g\left(s - i\right), \tag{2}$$

and by definition, $f_0 = P(X = 0) = 0$ and $g(0) = p_0$ (initial probability), where p_0 denotes the counting distribution, say p_n belonging to (1) and in this case, $n = 0$ (probability of no claim).

2.2. The Schröter recursive formula

While the Panjer recursive formula addresses the challenges of the traditional convolution method, it is limited to a few number of distributions. Hence, Schröter (1990) generalized (1) and obtained the recursive formula expressed as:

$$P_k = \left(a + \frac{b}{k}\right) P_{k-1} + \frac{c}{k} P_{k-2}, k = 1, 2, 3, \ldots, \tag{3}$$

where *a*, *b*, and *c* are parameters and $P_k = 0$ for $k < 0$ (by definition). Note that for $c = 0$, (1) becomes a particular case of (3). Additionally, the counting distributions defined by (3) also contain the convolutions of the Poisson distribution and another distribution from (1) (see

Schröter, 1990). Furthermore, Schröter (1990) obtained the corresponding recursion algorithm for (3) defined as:

$$g(s)=\frac{1}{1-af_0}\sum_{i=1}^{s}\left[\left(a+\frac{bi}{s}\right)f_i+\frac{ci}{2s}\,f_i^{2*}\right]g(s-i), \tag{4}$$

where f_i^{2*} has to be evaluated by the convolution formula $f_i^{2*}=\sum_{j=0}^{i} f_j f_{i-j}$ and for $c=0$, (4) becomes (2). The parameter estimation of (3) has been studied in Agu, Mačutek, and Szűcs (2023).

3. The truncated Schröter recursive formula

In this section, we introduce the truncated Schröter recursive formula at 1, a mathematical innovation designed to simplify the convolution process inherent in the standard Schröter recursion algorithm. By focusing on events where claims occur, this truncation reduces the computational complexity, especially in large datasets, by eliminating the need for extensive convolution calculations. This reduction in dimensionality leads to faster execution times while preserving accuracy. The streamlined model also aligns with the practical needs of insurance risk management, where the primary concern is the occurrence of claims rather than their distribution across a broad range of values.

We defined the truncated Schröter recursive formula as:

$$P_k=\left(a+\frac{b}{k}\right)P_{k-1}+\frac{c}{k}P_{k-2},\ k=2,3,4,..., \tag{5}$$

First, using the fact that

$$G(s)=\sum_{k=0}^{\infty}s^k P_k.$$

The probability generating function corresponding to (5) as:

$$G(s)=e^{-\frac{c(s-1)}{a}}\left(\frac{1-a}{1-as}\right)^{\frac{a(a+b)+c}{a^2}}, \tag{6}$$

for $|as|\neq 1$ and $s\in[0,1]$ such that $G(s)\geq 0$.

The derived truncated probability mass function corresponding to (5) is:

$$q_n = \frac{e^{\frac{c}{a}}(1-a)^{\frac{a(a+b)+c}{a^2}} \sum_{i=0}^{n} \binom{\frac{a(a+b)+c}{a^2}+i-1}{i} \frac{\left(-\frac{c}{a}\right)^{n-i} a^i}{(n-i)!}}{1-e^{\frac{c}{a}}(1-a)^{\frac{a(a+b)+c}{a^2}}}, n=1,2,..., 0<a<1, b,c \in \mathbb{R}. \quad (7)$$

Let $r = \frac{a(a+b)+c}{a^2}$, $x = \frac{c}{a}$, and define the generating function for the negative binomial coefficient as:

$$\sum_{k=0}^{\infty} \binom{r+k-1}{k} z^k = (1-z)^{-r}, |z|<1.$$

The goal is to express the finite sum in a form that leverages the generating function.

To relate $\sum_{i=0}^{n} \binom{r+i-1}{i} \frac{(-x)^{n-i} a^i}{(n-i)!}$ to the generating function of the negative binomial above, we take the nth derivative of $(1-z)^{-r}$ with respect to z and evaluated at $z = x-a$ to obtain terms that match the structure of the sum in (7). Hence, we have that

$$\sum_{i=0}^{n} \binom{r+i-1}{i} \frac{(-x)^{n-i} a^i}{(n-i)!} = \frac{\Gamma(r+n)}{n!\Gamma(r)} \left(\frac{a}{a-c+a^2}\right)^{(r+n)}.$$

Hence, (7) can be expressed as

$$q_n = \frac{e^{\frac{c}{a}}(1-a)^r \frac{\Gamma(r+n)}{n!\Gamma(r)} \left(\frac{a}{a-c+a^2}\right)^{(r+n)}}{1-e^{\frac{c}{a}}(1-a)^r}, n=1,2,..., 0<a<1, b,c \in \mathbb{R}.$$

The log-likelihood function corresponding to (7) is given as:

$$\ell(a,b,c \mid n_1,...,n_k) = \frac{ck}{a} + \left[\frac{a(a+b)+c}{a^2}\right] \log(1-a) + \sum_{j=1}^{k} \log \left[\sum_{i=0}^{n_j} \binom{\frac{a(a+b)+c}{a^2}+i-1}{i} \frac{\left(-\frac{c}{a}\right)^{n_j-i} a^i}{(n_j-i)!} \right] - k \log \left[1-e^{\frac{c}{a}}(1-a)^{\frac{a(a+b)+c}{a^2}}\right].$$

3.1. The truncated Schröter algorithm

Let $S_1, S_2, ..., S_k$ be a sequence of stochastically independent and identically distributed random claims over non-negative integers with probability density $f_j = p(S_i = j)$ for $i = 1, 2, ..., k$, $j = 0, 1, 2, ...,$ and f^{k*} denotes the n-fold convolution of f_j. Additionally, let N signifies a discrete random variable representing the number of claims with a discrete probability mass function $P_k = p(N = k)$, such that S_i are stochastically independent of N. For all the severity distributions f^{k*}, we derived the recursive algorithm as:

$$g(s) = \sum_{k=2}^{\infty} P_k f^{k*}(s), \tag{8}$$

The Panjer recursion formula defined in (2) is based on the expression $f^{k*}(s) = \frac{k}{s}\sum_{i=1}^{s} i f_i f_{s-i}^{k-1}$, $s = k = 1, 2, 3$ (see Schröter,1990; page 164). We can write this as: $f(s) = \frac{1}{s}\sum_{i=1}^{s} i f_i$.

Thus,

$$g(s) = \sum_{k=2}^{\infty}\left[\left(a + \frac{b}{k}\right)P_{k-1} + \frac{c}{k}P_{k-2}\right] f^{k*}(s). \tag{9}$$

We have that

$$g(s) = a\sum_{k=0}^{\infty} P_k\left(\sum_{i=0}^{s} f_i f_{s-i}^{k-1}\right) + \sum_{k=0}^{\infty} P_k\left(\sum_{i=1}^{s} \frac{bi}{s} f_i f_{s-i}^{k-1}\right) + \gamma,$$

where $\gamma = \sum_{k=0}^{\infty} P_k\left(\sum_{i=1}^{s} \frac{ci}{s} f_i f_{s-i}^{k-1}\right)$.

Note that $\sum_{k=0}^{\infty} P_k = 1$.

Therefore, it follows that

$$g(s)=\frac{1}{1-af_0}\sum_{i=1}^{s}\left[a+\frac{i}{s}(b+c)\right]f_i g\left(s-i\right), \tag{10}$$

for $s\neq 0$. Additionally, $f_0=P(S=0)=0$ and $g(0)=p_0$ (initial probability), where p_0 denotes the counting distribution, say p_n of the number of claims. If $c=0$ in (10), we obtain (2), and if we define $f^{k*}(s)$ as $f^{k*}(s)=\frac{k}{ts}\sum_{i=1}^{s}if_i^{t*}f_{s-i}^{(k-t)*}$, $i=1,2,...,$ for $t\in\{1,2,...,k\}$ in (9), (4) becomes a special case of (10). To execute (10), we treat f_i as the claim frequencies per number of policies.

Theoretically, unlike (4), the recursion algorithm defined in (10) eliminates the need for any form of convolution.

This study considers the Negative binomial distribution as the count distribution for the number of claims (see Section 4, Table 4).

The probability mass function for the Negative binomial distribution is defined as

$$h(S=s)=\binom{s+r-1}{s}(1-p)^s\,p^r,\ s=0,1,2...,\ r>0,\ p\in[0,1].$$

We have that

$$h(S=0)=\binom{r-1}{0}p^r.$$

From (10), we define

$$g\left(0\right)=p_0=h(S=0)=p^r. \tag{11}$$

4. Numerical Evaluation

In this section, we examine the run time computational efficiency of the introduced truncated Schröter algorithm using the Automobile UK Collision Claims (AutoCollision) data obtained from https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html.

First, we explored the descriptive statistics of the data and examined average claim severity, average claim counts across different age groups, and vehicle categories to identify patterns and how often each group files claims. Furthermore, we look for combinations of age groups and vehicle use categories with high claim severity and frequency. These combinations represent higher risk factors for insurers and may require insurance coverage adjustments. Additionally, we fit the Negative binomial and Generalized Poisson distributions to the data to determine the distribution of the claim count data. These distributions are selected based on their overdispersion property for counting datasets. The selection of the claim count distribution is based on the distribution that provides a better fit to the data using the least Akaike information criteria (AIC) and Bayesian information criteria (BIC) values. Furthermore, to utilize the truncated algorithm in (10), we employ the truncated probability mass function defined in (7) to obtain the parameter estimate of $\hat{a}, \hat{b},$ and $\hat{c}$ as $\hat{a} = 0.99070,$ $\hat{b} = 1.29297,$ and $\hat{c} = 0.29330$ numerically using the maximum likelihood estimation method. Finally, we present the computation of aggregate claims for (2), (4), and (10) in Section 4.1.

In this study, all analyses and computational run times of the recursion algorithms were performed using RStudio on a Lenovo PC with processor 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42GHz and 8.00 GB RAM.

Table 1: Descriptive Statistics

Min.	Max.	Mean	Variance	Kurtosis	Skewness
5.00	970.00	279.44	58374.38	4.08	1.25

The descriptive statistics provide a comprehensive overview of the dataset's distribution and central tendency. The minimum and maximum values indicate the range of the data, while the mean offers a measure of central tendency. The high variance indicates substantial variability (overdispersion) in the data, while the positive kurtosis and skewness values indicate the presence of outliers and a right-skewed distribution, respectively.

Table 2: Analysis of Average Claim Severity by Vehicle Use

Vehicle Use	Average Claim Severity	Claim Count
Business	395.21	1075
DriveLong	265.26	2710
DriveShort	231.74	3888

Pleasure	213.20	1269

Table 2 shows that vehicles used for business purposes have the highest average claim severity. Despite the relatively lower claim count compared to other categories, the financial impact of each claim is substantial, indicating that business use poses a higher risk of costly claims.

Long drives have a moderate average claim severity, which is substantially lower than that of business use but higher than that of short drives and pleasure use. The claim count is relatively high, indicating that long-distance driving is associated with frequent claims. However, the severity per claim is not as high as for business use.

Short drives have the highest claim count but a lower average claim severity. This indicates that while short trips result in more frequent claims, the financial impact of each claim is comparatively lower. The high frequency indicates a notable number of incidents but with less severe outcomes per incident.

Pleasure use has the lowest average claim severity and a relatively low claim count. This implies that leisure driving results in fewer and fewer claims, making it the lowest risk category in terms of both frequency and financial impact for the Automobile UK Collision Claims data.

Table 3: Analysis of Average Claim Severity by Age

Age	Average Claim Severity
17–20	391.80
21–24	293.17
25–29	284.84
30–34	279.73
35–39	212.43
40–49	249.99
50–59	251.11
60+	247.68

Table 3 shows that young drivers aged 17–20 have the highest average claim severity. This indicates that accidents involving younger drivers tend to result in higher costs, indicating a substantial financial risk associated with this age group.

Drivers aged 21–24 show a substantial decrease in average claim severity compared to the 17–20 age group. While still relatively high, this implies that slightly older young drivers pose a lower financial risk than the youngest group. The average claim severity decreases for drivers

aged 25–29, indicating an ongoing trend of reduced financial impact from claims as drivers gain more experience and maturity. The average claim severity for drivers aged 30–34 slightly reduces compared to the 25–29 age group, continuing the trend of lower severity with increasing age. This age group (35–39) experiences a notable drop in average claim severity, indicating that drivers in this category are involved in less severe incidents, making them a lower financial risk. There is an increase in average claim severity for the 40–49 age group compared to the 35–39 group. However, it remains lower than the severity for drivers under 30, indicating a moderate financial risk. The average claim severity for drivers aged 50–59 is similar to the 40–49 age group, implying a consistent level of financial risk for middle-aged drivers. Older drivers aged 60+ have a slightly lower average claim severity compared to the 50–59 age group. This indicates a stable, moderate financial risk, slightly lower than the middle-aged groups but higher than the 35–39 age group.

Table 4: The fitting of Negative binomial and generalized Poisson distributions

	Negative binomial	**Generalized Poisson**
Parameters Estimate	$\hat{p} = 0.00453$, $\hat{r} = 1.25042$	$\hat{\theta} = 12.78279$, $\hat{\lambda} = 0.95426$
N-Loglikelihood	-211.9633	-216.1883
AIC	427.9267	436.3767
BIC	430.8581	439.3081

From Table 4, the Negative binomial distribution has a less negative loglikelihood and the least AIC and BIC values, indicating a better fit to the data than the Generalized Poisson distribution, making it the more suitable choice for modeling claim count data in this scenario. Hence, the distribution of the AutoCollision. The AIC and BIC consider model complexity and penalize for overfitting, particularly relevant in this context.

Using (11), $g(0) = 0.00117$

4.1. Computation of Aggregate Claim

In this section, we used the estimate of $\hat{a}, \hat{b},$ and $\hat{c}$ for (2), (4), and (10) to compute aggregate claim amounts for each recursive algorithm and their computational run time using Claim count data from the AutoCollision data.

Table 5: Aggregate claim for the truncated Schröter, Panjer, and Schröter algorithms

s	g(s) The Truncated Schröter	g(s) The Panjer	g(s) The Schröter
1	711000×10^{-3}	711000×10^{-3}	711000×10^{-3}
2	13488×10^{-5}	11953×10^{-5}	11958×10^{-5}
3	78790×10^{-6}	69718×10^{-6}	69763×10^{-6}
4	18245×10^{-6}	16054×10^{-6}	16132×10^{-6}
5	21313×10^{-5}	18881×10^{-5}	18908×10^{-5}
6	58042×10^{-5}	51406×10^{-5}	51450×10^{-5}
7	32214×10^{-5}	28454×10^{-5}	28493×10^{-5}
8	15979×10^{-5}	14070×10^{-5}	14113×10^{-5}
9	48176×10^{-5}	42617×10^{-5}	42727×10^{-5}
10	11819×10^{-4}	10454×10^{-4}	10474×10^{-4}
11	11254×10^{-4}	99323×10^{-5}	99514×10^{-5}
12	49170×10^{-5}	43135×10^{-5}	43306×10^{-5}
13	46463×10^{-5}	40787×10^{-5}	41095×10^{-5}
14	15944×10^{-4}	14065×10^{-4}	14119×10^{-4}
15	13782×10^{-4}	12090×10^{-4}	12149×10^{-4}
16	74915×10^{-5}	65005×10^{-5}	65502×10^{-5}
17	65619×10^{-5}	57003×10^{-5}	57681×10^{-5}
18	18092×10^{-4}	15882×10^{-4}	15999×10^{-4}
19	16315×10^{-4}	14191×10^{-4}	14324×10^{-4}
20	95516×10^{-5}	81572×10^{-5}	82706×10^{-5}
21	11549×10^{-4}	99748×10^{-5}	10108×10^{-4}
22	36393×10^{-4}	31961×10^{-4}	32173×10^{-4}
23	30416×10^{-4}	26473×10^{-4}	26719×10^{-4}
24	17428×10^{-4}	14879×10^{-4}	15094×10^{-4}
25	15023×10^{-4}	12823×10^{-4}	13062×10^{-4}
26	35570×10^{-4}	30989×10^{-4}	31358×10^{-4}
27	27630×10^{-4}	23653×10^{-4}	24074×10^{-4}
28	17519×10^{-4}	14558×10^{-4}	14920×10^{-4}
29	18952×10^{-4}	15950×10^{-4}	16353×10^{-4}
30	30447×10^{-4}	26076×10^{-4}	26676×10^{-4}
31	27511×10^{-4}	23012×10^{-4}	23706×10^{-4}
32	22584×10^{-4}	18434×10^{-4}	19048×10^{-4}
Sum of Probabilities	0.005483	0.004887	0.004535

Execution time in seconds(s)	0.051093	0.060898	0.173438

Based on Table 5, the truncated Schröter recursion algorithm demonstrates the fastest execution time at 0.051093 s. This can be attributed to its simplified recursive structure that omits the convolution components in the traditional Schröter formula. This simplification results from the truncation mechanism, which reduces the number of calculations required at each step. The algorithm efficiently captures the essential characteristics of the claim data, focusing on the occurrence of claims rather than the entire distribution, thereby offering a more computationally efficient and practical solution for aggregate claim calculations. (see **Fig. 1**). This result implies a probability of approximately 0.55% that the total claim amount will not exceed 32 units. In contrast, the probability that the total claim amount will exceed 32 units is 99.45%. The Panjer recursion algorithm, while slightly slower with a run time of 0.060898 s, computes an aggregate claim sum of 0.004887. This signifies relative efficiency but is potentially more conservative or less sensitive to data variations (see **Fig. 2**). This result indicates a probability of approximately 0.49% that the total claim amount will not exceed 32 units, with a 99.51% probability that it will exceed 32 units. Lastly, the Schröter recursion algorithm, which has the slowest run time of 0.173438 s, computes an aggregate claim sum of 0.004930. This implies a balance between sensitivity and comprehensiveness, though, with higher computational complexity due to the additional convolution component f_i^{2*} (see **Fig. 3**). This result indicates a probability of 0.45% that the total claim amount will not exceed 32 units. In contrast, the probability that it will exceed 32 units is approximately 99.55%.

The general interpretation of these results is that the likelihood of the total claim amount being less than or equal to 32 units is Low, with probabilities ranging from approximately 0.45% to 0.55%. Consequently, the probability that the total claim amount will exceed 32 units is extremely high, ranging from 99.45% to 99.55% for the AutoCollision data. This indicates that, under the different recursion algorithms evaluated, it is almost certain that the total claim amount will surpass 32 units, reflecting the high-risk nature of the claims being modeled.

These results can substantially aid in effective risk management and premium setting in the insurance sector by providing insights into the probability of large claim amounts. The high likelihood that total claims will exceed 32 units could signify that insurers need to prepare for substantial payouts. This information can help insurers understand the distribution and frequency of high claims, allowing for more accurate risk assessment and appropriate premium

setting to ensure financial stability. Additionally, it enables insurers to allocate sufficient reserves to handle the expected high claims, reducing the risk of insolvency. Insurers can also design policies with deductibles, limits, and exclusions that reflect the likelihood of large claims, balancing affordability for customers with profitability for the insurer. Furthermore, these insights facilitate the development of reinsurance strategies to transfer portions of high-risk exposures, minimizing the financial impact of large claims on the insurer.

The differences in computational run times and aggregate claim sums can be attributed to the complexity and nature of the recursion algorithms. The truncated Schröter algorithm efficiently balances the complexity introduced by its three parameters, leading to fast computations and higher aggregate claims. The Panjer recursion algorithm, with its two parameters, simplifies computations but might not capture as many nuances. The Schröter recursion algorithm adds complexity with the additional term, resulting in longer computation times but providing a different perspective on the aggregate claims.

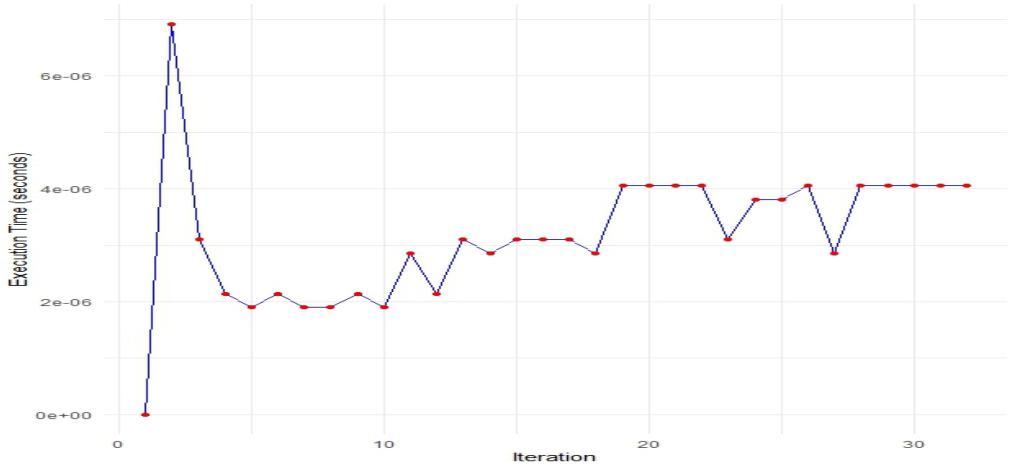


Fig. 1: The execution time plot of the truncated Schröter recursion algorithm for each iteration

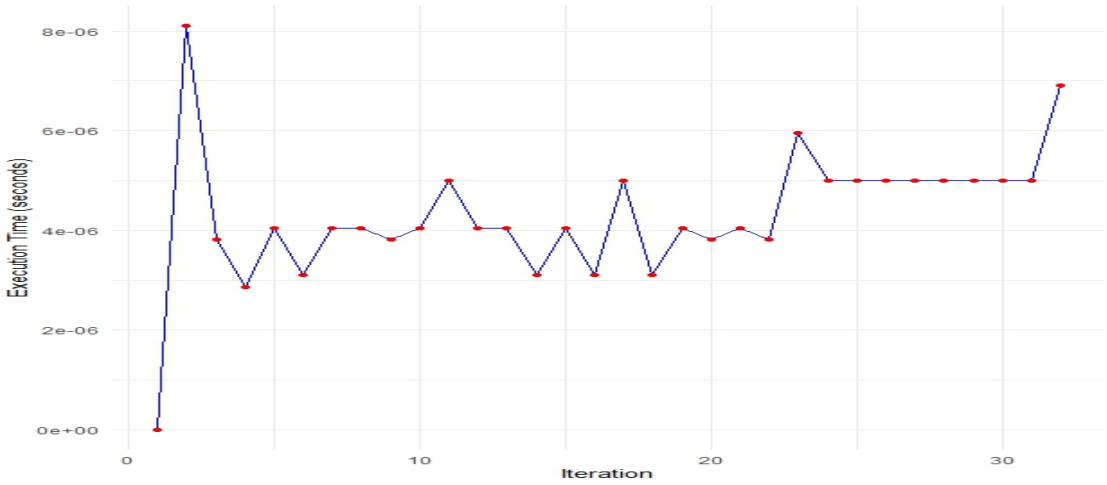


Fig. 2: The execution time plot of the Panjer recursion algorithm for each iteration

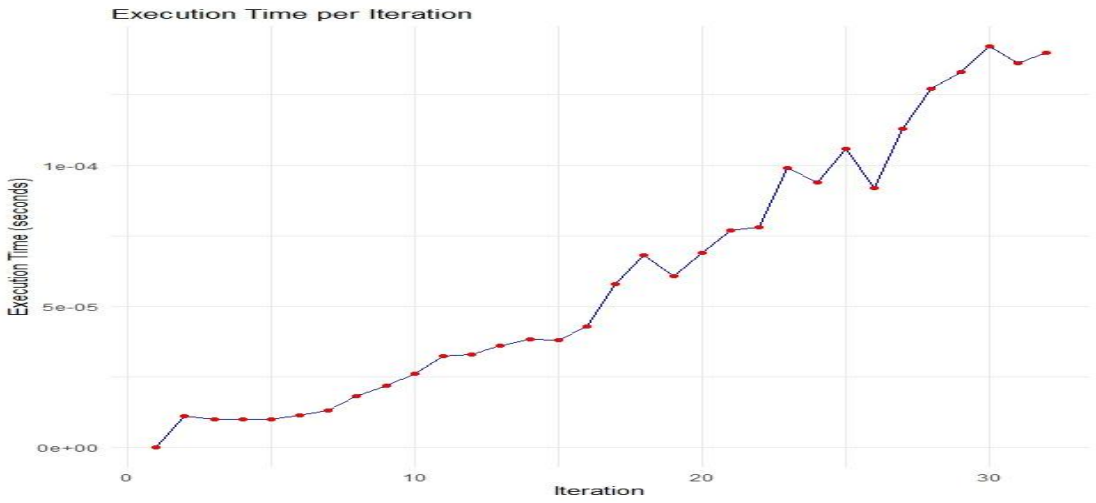


Fig. 3: The execution time plot of the Schröter recursion algorithm for each iteration

4.2. Simulation Study

Here, we generate random claim amounts data from the Negative binomial distribution by setting $r = 100$ and $p = 0.05$. We varied the sample size to examine the aggregate claim computational efficiency and run time of the truncated Schröter, Panjer, and Schröter recursion algorithms. Initially, we generated 5000 random numbers from the Negative binomial distribution and fit (7) to the data to obtain the estimate of the parameters $\hat{a}, \hat{b},$ and $\hat{c}$ as $\hat{a} = 0.9905,\ \hat{b} = 18.3096,\ \hat{c} = -1.2840$, and compute $g(0) = 0.00117$ to implement the algorithms. Tables 6, 7, and 8 present the sample sizes, aggregate claim amounts, and the execution time in seconds for each recursion algorithm.

Table 6: Efficiency of the truncated Schröter algorithm on simulated claim data

Recursion Algorithm	Sample (n)	sum of g(s) Execution time (s)
The truncated Schröter algorithm		
20	2.7080348	0.0006919
50	1.5211150	0.0036724
100	1.1730166	0.0188396
150	1.1747598	0.0335643
200	1.1296751	0.0586591
300	0.9208012	0.1158113
600	0.8504156	0.3374069
1500	0.6127028	0.8211629
2000	0.5397614	1.0198436
3000	0.4756758	1.2040498
4000	0.4180231	1.4369745
5000	0.3979199	1.5046701

Table 7: Efficiency of the Panjer algorithm on simulated claim data

Recursion Algorithm	Sample (n)	sum of g(s) Execution time (s)
The Panjer algorithm		
20	0.0075441	0.0014127
50	0.0074026	0.0047266

100	0.0074002	0.0212255
150	0.0074619	0.0378468
200	0.0074672	0.0692441
300	0.0072761	0.1391730
600	0.0073219	0.4021811
1500	0.0073130	1.0212069
2000	0.0072711	1.0606146
3000	0.0072833	1.3638492
4000	0.0072209	1.6406126
5000	0.0072351	1.7650454

Table 8: Efficiency of the Schröter algorithm on simulated claim data

Recursion Algorithm	**Sample (n)**	**sum of g(s)** **Execution time (s)**
The Schröter algorithm		
20	0.0062785	0.0028598
50	0.0063757	0.0195415
100	0.0064331	0.0765483
150	0.0064909	0.1735694
200	0.0065073	0.2585254
300	0.0063946	0.6078202
600	0.0064442	1.8534706
1500	0.0064868	4.6577935
2000	0.0064729	5.4729755
3000	0.0065035	6.8906786
4000	0.0064749	8.0277340
5000	0.00649490	8.7507973

Fig. 4 illustrates the execution times of the truncated Schröter recursion, Panjer recursion, and Schröter recursion algorithms for varying *n*, highlighting significant differences in their computational efficiency as *n* increases. The truncated Schröter recursion algorithm consistently demonstrates the lowest execution times across all *n*, starting at 0.0006919 s for n = 20 and increasing to 1.5046701 s for n = 5000 (see **Fig. 4**). This indicates that it is highly efficient and scalable, handling larger datasets relatively easily. The Panjer recursion

algorithm, while also showing an increase in execution time with larger n, has moderately higher times compared to the truncated Schröter algorithm, beginning at 0.0014127 s for n = 20 and reaching 1.7650454 s for n = 5000, demonstrating reasonable efficiency but less scalability compared to the truncated Schröter algorithm (see **Fig. 4**). In contrast, the Schröter recursion algorithm, which includes the convolution component, f_i^{2*}, shows substantially higher execution times, starting at 0.0028598 s for n = 20 and rising steeply to 8.7507973 s for n = 5000 (see **Fig. 4**). This steep increase indicates poor scalability and inefficiency, especially for larger n, making it the least optimal algorithm among the three. Overall, the truncated Schröter recursion algorithm proves to be the most efficient and scalable, followed by the Panjer recursion algorithm. In contrast, the Schröter recursion algorithm with the convolution component is the least efficient, particularly for larger datasets. We conducted five runs for each sample size to assess computational time variation. The purpose of these repeated runs is to measure and evaluate the variability in the time it takes to perform the computation. Notably, the variation in computational time was negligible (i.e., it showed minimal variation, indicating that the time it took to complete the tasks was fairly consistent across multiple runs). However, the current operations of the PC could also influence it.

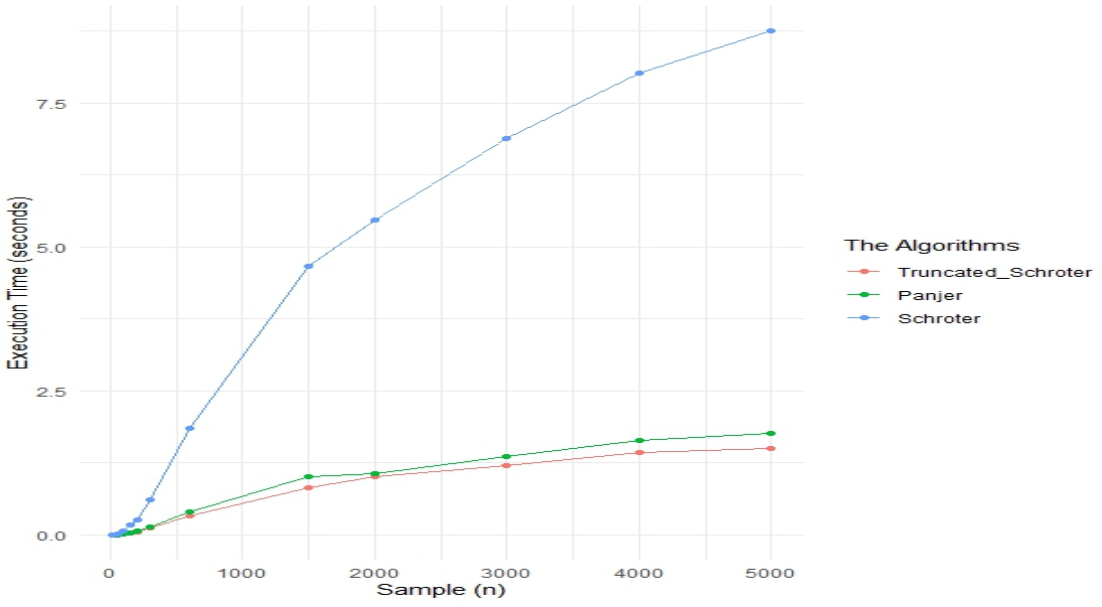


Fig. 4. Visual representation of the execution time of the truncated Schröter recursion, Panjer recursion, and Schröter recursion algorithms.

5. Conclusion

This study explored the computation of aggregate claim amounts using various recursive algorithms, focusing on the newly introduced truncated Schröter recursive algorithm. The primary goal was to improve accuracy and computational efficiency in determining aggregate claim amounts, which is crucial for effective risk management and premium setting in the insurance industry.

The truncated Schröter recursion algorithm demonstrated superior performance in numerical evaluations and comparisons, largely due to its innovative truncation mechanism. This feature significantly reduces computational complexity by simplifying the recursive steps and eliminating the convolution operations required in other algorithms. The AutoCollision dataset results underscore the algorithm's ability to achieve faster execution times while maintaining high accuracy, proving its practical value in the insurance industry. The Negative Binomial distribution was preferred over the Poisson distribution for modeling claim count data due to its ability to handle overdispersion, validated by AIC and BIC criteria.

Simulation studies confirmed the efficiency of the truncated Schröter recursion algorithm across varying sample sizes. The algorithm consistently outperformed other algorithms in execution time and maintained scalability. The algorithm effectively captured and incorporated various data variations, leading to thorough and nuanced analyses. The study also highlighted the importance of selecting appropriate counting distributions and recursive algorithms for computing aggregate claim amounts.

In conclusion, the truncated Schröter algorithm emerges as a robust tool for calculating aggregate claim amounts, balancing computational efficiency and accuracy. Its implementation can enhance risk management and premium decision-making, benefiting insurers and policyholders. Future research could explore enhancements to the truncated Schröter algorithm, such as integrating machine learning techniques to dynamically adjust parameters based on evolving claim data patterns and the algorithm's predictive power. Additionally, applying the algorithm to various types of insurance data beyond automobile claims could validate its versatility and identify any domain-specific adjustments required.

Funding

This study was supported by the Slovak Academy of Sciences Doktogrant. Grant Number: APP0515 (Friday I. Agu).

Dedication

This manuscript is dedicated to my esteemed supervisor, Jan Macutek, whose guidance, support, attention, and encouragement have been instrumental in exploring this study area. Your unwavering belief in my abilities and invaluable insights have greatly contributed to my academic and professional growth. Thank you for being an inspiring mentor.

Conflict of Interest

There is no conflict of interest for this study.

References

Agu, F. I., Mačutek, J., & Szűcs, G. (2023). A simple estimation of parameters for discrete distributions from the Schröter family. *Statistika: Statistics & Economy Journal, 103*(2).

Beard, R. E., Pentikäinen, T., & Pesonen, E. (1977). *Risk theory* (2nd ed.). Chapman and Hall.

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation, 19*(90), 297–301.

Dickson, D. C. (2016). *Insurance risk and ruin*. Cambridge University Press.

Dzidzornu, S. B., & Minkah, R. (2021). Assessing the performance of the discrete generalised Pareto distribution in modelling non-life insurance claims. *Journal of Probability and Statistics, 2021*, Article 5518583. https://doi.org/10.1155/2021/5518583

Ghinawan, F., Nurrohmah, S., & Fithriani, I. (2021). Recursive and moment-based approximation of aggregate loss distribution. In *Journal of Physics: Conference Series* (Vol. 1725, No. 1, p. 012101). IOP Publishing. https://doi.org/10.1088/1742-6596/1725/1/012101

Gray, R. J., & Pitts, S. M. (2012). *Risk modeling in general insurance: From principles to practice*. Cambridge University Press.

Heckman, P. E., & Meyers, G. G. (1983). The calculation of aggregate loss distributions from claim severity and claim count distributions. In *Proceedings of the Casualty Actuarial Society* (Vol. 70, No. 133–134, pp. 49–66). Casualty Actuarial Society.

Hogg, R. V., & Klugman, S. A. (2009). *Loss distributions*. John Wiley & Sons.

Jindrová, P., & Pacáková, V. (2016). Modeling of extreme losses in natural disasters. *International Journal of Mathematical Models and Methods in Applied Sciences, 10*.

Pacáková, V., & Gogola, J. (2013). Pareto distribution in insurance and reinsurance. In *Proceedings of the 9th International Scientific Conference Financial Management of Firms and Financial Institutions* (pp. 298–306). VŠB Ostrava.

Packová, V., & Brebera, D. (2015). Loss distributions in insurance risk management. In *Recent Advances on Economics and Business Administration* (pp. 17–22).

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin: The Journal of the IAA, 12*(1), 22–26.

Schröter, K. J. (1990). On a family of counting distributions and recursions for related compound distributions. *Scandinavian Actuarial Journal, 1990*(2–3), 161–175.

Sundt, B., & Vernic, R. (2009). *Recursions for convolutions and compound distributions with insurance applications*. Springer Science & Business Media.

Yartey, E. (2020). *The (a, b, r) class of discrete distributions with applications* (Doctoral dissertation, Laurentian University of Sudbury).

Modified canonical correlation analysis based on LASSO technique

E. A, Iguodala [1], J. E, Osemwenkhae [2], F.O, Oyegue [2], N, Ikoba [3], A, Iduseri [2] and A, Edokpayi [1].

[1] Statistics Division, Nigerian Institute for Oil Palm Research (NIFOR), P.M.B 1030, Benin City, Edo State, Nigeria.
[2] Department of Statistics, Faculty of Physical Sciences, University of Benin, Benin City, Nigeria.

[3] Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Kware State, Nigeria.

**Corresponding author's e-mail: eddyrandley2k @ yahoo.com.*

Abstract

The act of describing the association of multiple dependent variables and multiple independent variables using Canonical correlation analysis (CCA) has attracted a great deal of attention. Over the years, various efforts had been made to formulate the conventional CCA as a modelbased method that is useful in modeling and prediction using various formulations of the conventional CCA as optimization problem with objective functions and constraints. The modified optim code in R package for the least squares algorithm was used as a numerical approximation for the analytically intensive solution. The resulting model produced canonical results that are not only equivalent to that of the conventional model-based CCA for low dimensional data, but also produced better interpretable results in situations where the number of variables greatly exceeds the number of subjects. The resulting model were able to simplify the interpretation of the CCA components, as well as enable us find out the components among

the independent variables that contributed the most to the dependent variable on the same subject.

1. Introduction

Canonical Correlation Analysis (CCA) introduced by Hotelling (1935; 1936) is a family of multivariate statistical methods for the analysis of paired sets of variables. CCA looks for a linear combination of all the variables in one data set that correlates maximally with a linear combination of all the variables in the other data set (Thompson, 1984).

This study, in particular, looked at the relationship between CCA and least squares formulations that are equivalent to multiple regression model for CCA problems. Under a mild condition which tends to hold for multicollinearity, overfitting and high-dimensional data setting, the least squared formulations often fail to select pairwise variables that will mutually maximize the correlation between the common subspaces and estimate coefficients simultaneously, thereby resulting to increase in model complexity with unstable parameter estimates which often makes interpretability difficulty.

Over the years, the CCA technique and it's equivalent model based formulations had received increased attentions, and efficient computational procedures have been offered. But not without some limitations, such as the issue of multicollinearity, matrix inversion and high-dimensionality which often lead to difficulties in interpretability. Hence, the modified CCA technique based on LASSO adopt various constraints as an additional restriction on the loglikelihood function to maximize the correlation between the two CCA components, with the hope of removing non-informative variables in order to produce meaningful results that are interpretable. This study, describe the use of the newly developed CCA in estimating the association between plant characteristics variables and yield components variables, by employing the LASSO techniques (Tibshirani 1996), to simplify the interpretation of the CCA components. We also, analyzed a data regarding a nutrition study in mouse (Nutrimouse dataset) as an illustration of our method. Based on this equivalence relationship, hence propose CCA extension including penalized CCA using the 1-norm regularization. A modification of CCA that will estimate the association between multiple dependent variables and multiple independent variables, by employing the Least Absolute Shrinkage and Selection Operator (LASSO) as an additional restriction on the loglikelihood function in order to maximize the correlation between the two CCA components or canonical correlation is proposed. The

modification of CCA evaluated the theoretical results using a collection of two data sets. Results also show that, even when the assumption hold, This results achieve very interpretable performance, which also demonstrate the effectiveness of the proposed CCA extensions.

2. Literature Review

Canonical Correlation Analysis (CCA) originated from Hotelling (1935) who applied the technique to a data set in which one set of variables consisted of mental tests and the other involved physical measurements. With the objective to finding pairs of some basic vectors that maximize the correlation of a set of paired variables. This was achieved by considering two set of multivariate random variables $X \in R^p$ and $Y \in R^q$, where $X = (x_1, x_2 \dots, x_p)$ and $Y = (y_1, y_2 \dots, y_q)$ are column vectors.

The CCA procedure is to choose directions $a_x \in R^p$ and $b_y \in R^q$ onto vectors X and Y respectively, such that the projections are maximize. Let $u_x \in R^n$ and $v_y \in R^n$ be projections of $X \in R^{n \times p}$ and $Y \in R^{n \times q}$, where p, qn, nN

Since CCA is based on linear transformations, we have: $u_x = a_X^T X$

and

$C_{xx} \in R^{p \times p}$ is the covariance matrix of set X

$C_{yy} \in R^{q \times q}$ is the covariance matrix of set Y

$C_{xy} = C_{yx}^T \in R^{p \times q}$ is the cross-covariance matrice between set X and Y.

The correlation coefficient ρ of u_x and v_y is given as;

$$\rho(u, v) = \frac{\text{cov}(u, v)}{\sqrt{\text{var}(u)\text{var}(v)}} = \frac{a_x^T C_{xy} b_y}{\sqrt{((a_x^T C_{xx} a_x)(b_y^T C_{yy} b_y)}} \quad (2)$$

The joint covariance matrix is then

$$\begin{pmatrix} C_{XX} & C_{XY} \\ C_{XY} & C_{YY} \end{pmatrix}$$

Now the eigenvalues and eigenvectors of the characteristic equation are computed

$$(C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy} - \rho^2 1) b_y = 0 \quad (3)$$

In linear CCA, the canonical correlations equal to the square roots of the eigenvalues. however, others researchers works on its limitations. According to Muller (1982) proposed an alternative approach which presented CCA as a least squares problem. This multivariate multiple regression representation of CCA amounts to finding an estimate of β, α and $D(p_k)$ in the following model equation:

$$Y\beta = X\alpha D(p_k) + \varepsilon \tag{4}$$

The matrices β, α, and $D(p_k)$ correspond in the sense that the k^{th} columns of β and α provide the linear combinations that are correlated ρ_k, which is the (k, k) element of $D(p_k)$. However, the multivariate formulation introduced some greater complication as a result of alternating the vectors with matrices. For instance, the equivalence of β, α, and $D(p_k)$ in the standard statement of CCA are vectors (not matrices). Hence, the standard statement of canonical correlation has more in common with the univariate statement than with the multivariate that was developed in the study. To address the limitation of nonlinear associations between sets of variables Lai and Fyfe (2002) proposed the method of Kernel Canonical Correlation Analysis (KCCA) to solve the issue of linearity.

The main idea of KCCA is that one first virtually maps data X and Y into a high dimensional feature space H_x and H_y via mappings Φ_x and Φ_y such that, data in the feature space become

$$K_x(x_i, x_j) = \left(\phi_x(x_1), \phi_x(x_2), \dots, \phi_x(x_n)\right)_{H_x} \in R^{Nx \times n} \tag{5}$$

$$K_y(y_i, y_j) = \left(\phi_y(y_1), \phi_y(y_2), \dots, \phi_y(y_n)\right)_{H_y} \in R^{Ny \times n} \tag{5}$$

where $N_x = n, N_y = n$ are the dimensions of ϕ_x, ϕ_y feature map. As derived in [Bach and Jordan 2005], the original data matrices $X \in R^{n \times p}$ and $Y \in R^{n \times q}$ can be substituted by the Gram matrices $K_x \in R^{n \times n}$ and $K_y \in R^{n \times n}$. Let α and β denote the positions in the kernel space R^n such that, we have $U = K_x\alpha$ and $V = K_x\beta$

KCCA solve the following optimization problem formulated as

$$\begin{aligned} \max_{\alpha_x, \beta_y \in \Re^n} \quad & \alpha_x' K_x' k_y \beta_y \\ \text{s.t} \quad & \alpha_x' K_x^2 \alpha_x = 1 \\ \text{and} \quad & \beta_y' K_y^2 \beta_y = 1 \end{aligned} \tag{6}$$

However, the dimensions of feature spaces H_x and H_y can be very high. Overall, this novel approach of linearity and nonlinearity; however, did not take care high dimensionality problem which introduced some complication. Hence, Waaijenborg et al. (2008) proposed a least squares approach with elastic net to take care of the size and discretization of the search grid in order to generalized the penalized CCA. The ridge parameter of the elastic net is set to be large, which tends to ignore the dependency structure within each set of variables. A simulation study of the method showed that it was capable of selecting relevant variables. However, the probability of identifying the relevant variables depends on the size of the multiple correlation of the data set in the pathway. Hence, the process needs to be repeated until the residual matrices contain no more information or until the decision is made that further analysis is no longer useful which is also very expensive to compute.

Furthermore, Parkhomenko et al., (2009) introduced an alternative technique to overcome the problems of computational cost associated with Waaijenborg et al. (2008) known as the sparse CCA techniques. The sparse CCA imposes extra penalties on the traditional CCA to reduce the canonical loading so that the sparse solution would have explicit understanding of a complex situation.

The sparse CCA optimization problem is formulated as;

$$\begin{aligned} &\max_{a_x, b_y} a_x^T X^T Y b_y \\ \text{s.t } &a_x^T X^T X a_x = 1, b_y^T Y^T Y b_y = 1 \\ &\|a_x\| \leq c_1, \|b_y\| \leq c_2 \end{aligned} \tag{7}$$

where $c_1 > 0$ and $c_2 > 0$ control the level of sparsity. Setting small values for c_1 and c_2 increase the penalty and forces more canonical correlation coefficient (CCC) to zero to select relevant variables. However the setback of SCCA is that it does not have a direct control over sparsity of the solution. In other words, sparsity is the canonical vectors is only guaranteed if $\hat{C}_{xx}$ and $\hat{C}_{yy}$ are replaced by their corresponding diagonal matrices. Hence the use of Bayesian Information Criterion (BIC) to further filter out unimportant features. While, Witten et al.(2009) improve on the work of Parkhomenko et al., (2009) by simply penalizing the canonical vectors to reduce many of their coefficients to zero with the assumption is that one can replace the matrices C_{xx} and C_{yy} by identity matrices. The optimization problem is formulated as

$$\max_{a,b} a_x^T C_{xy} b_y$$

$$\text{s.t } a_x^T C_{xx} a_x \leq 1, b_y^T C_{yy} b_y \leq 1 \quad (8)$$
$$P_1(a_x) \leq c_1, P_2(b_y) \leq c_2$$

where the penalties P_1 and P_2 can be defined as the L_1 norm penalty regularization. The parameters c_1 and c_2 can be selected by cross-validation to control the sparsity of a_x and b_y. The penalty function acts on individual variables so that some coefficients of the solution are reduced to zero. However, the simple substitution for C_{xx} and C_{yy} can be unrealistic for high-dimensional setting, which leads to some inaccurate and non-sparse solution. To overcome the challenge of inaccurate and non-sparse solution for high dimensional data. Chen et al. (2012) proposed the group structure sparse CCA by exploring the association between a model composed of explanatory and response variables where the group structural knowledge on response variable is of more interest. To this end, the first view penalty $P_1(a_x) = \|b_y\|_1$ is fixed, and the second view $P_2(b_y)$ is integrated with group structure. Consequently, the group sparsity only work on one view of data, although it achieves promising performance on particular data type, but lack generalization, because it only focuses on one set of variable. Lin et al. (2013) modified the work of Chen et al., (2012) in order to addressed the issue of lack of generality by penalizing the low rank approximation of the matrix. Clearly, the model was a more general extension for formulation. However, the model was essentially seeking for the sparse representation instead of the original projection of CCA, which could result to inaccurate estimation. Although their regularization methods and algorithms differ from each other, but they all shared one thing in common by putting the penalty assumption on the coefficient of canonical variate. The new formulation for canonical correlation analysis as an equivalent multiple regression model for two Gaussian random vectors by Ayodeji and Obilade (2016) presented a probabilistic formulation for CCA using a novel combination of the techniques of multiple regression of two Gaussian random vectors in order to overcome the problem of inaccurate estimation associated with the work of Lin et al., (2013).

They considered n observations of two sets of standardized variables with $X = \{X_{i,j}\}_{i=1,\ldots,n,j=1,\ldots,p}$ and $Y = \{y_{i,j}\}_{i=1,\ldots,N,j=1,\ldots,q}$, and obtained the estimates of α and β from the regression model

$$Y_\beta = X_\alpha + \epsilon, \epsilon \sim N(0,1) \quad (9)$$

$$\max_{\alpha\beta} \alpha' R_{XY} \beta = \max_{\alpha\beta} (Y\beta - X\alpha)'(Y\beta - X\alpha), \quad (10)$$

subject to $(X\alpha)'(X\alpha) = 1$ and $(Y\beta)'(Y\beta) = 1$

Equation (10) may be equivalently expressed as

$$\max_{\alpha\beta} \alpha' R_{XY} \beta = \max_{\alpha\beta} -\frac{N}{2} \log\ 2\pi - \frac{1}{2}(Y\beta - X\alpha)'(Y\beta - X\alpha) \quad (11)$$

to be the likelihood of Equation (10)

By applying the lagrange multipliers to Equation (11); yield

$$Q = -\frac{N}{2} \log\ 2\pi - \frac{1}{2}(Y\beta - X\alpha)'(Y\beta - X\alpha) - \frac{k}{2}\{(X\alpha)'(X\alpha) - 1\} - \frac{L}{2}\{(Y\beta)'(Y\beta) - 1\} (12)$$

Equation (12) yields two normal equations which may then be solved conventionally for α, β The corresponding eigenvalue problem providing the solution to this optimization criterion is given by

$$\left(R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} - \rho^2 1\right)\alpha = 0 \quad (13)$$

and

$$\left(R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} - \rho^2 1\right)\beta = 0 \quad (14)$$

This method determines the linear combinations for all set of variables in the original space leading to the solution with higher computation time. Also, when the number of predictor variables become large, the outcome is difficult to interpret, In addition, the method does not perform variable selection and hence unimodal feature selection is performed for each modality separately.

3. Methodology

In this section, the proposed modification of the Canonical Correlation Analysis (CCA) that will estimate the association between multiple dependent variables and multiple independent variables, by employing the Least Absolute Shrinkage and Selection Operator (LASSO) as an additional restriction on the loglikelihood function in order to maximize the correlation

between the two CCA components or canonical correlation is presented. An equivalent way of deriving LASSO estimates is to maximize the residual sum of squares with the addition of a penalty function. For some multiplier λ, and for any given value of t in the first LASSO formulation there is a value of λ in the second formulation that gives equivalent results. Thus, the equivalent loglikelihood function for the proposed method is given as:

$$L = -\frac{n}{2}\log\ 2\pi - \frac{1}{2}(Y\beta - X\alpha)'(Y\beta - X\alpha) + \lambda_1 \sum_{i=1}^{p} |\alpha_i| + \lambda_2 \sum_{j=1}^{q} |\beta_j| \qquad (15)$$

A suitable choices of t been some tuning parameter for this constraint has an interesting property that is forces some of the coefficients in the equation to zero, while the LASSO rightfully enjoys popularity in the statistical community due to its good predictive properties, and depending on the signal-to-noise ratio in the data.

Thus, we maximize the penalized log likelihood *L* as;

$$\begin{aligned} \max_{\alpha}, \beta = -\frac{n}{2}\log\ 2\pi - \frac{1}{2}(Y\beta - X\alpha)^T(Y\beta - X\alpha) - \\ \frac{n}{2}\lambda_1 \sum_{i=1}^{p} |\alpha_i| - \frac{n}{2}\lambda_2 \sum_{j=1}^{q} |\beta_j| \\ \text{subject to } (X\alpha)^T(X\alpha) = n \text{ and } (Y\beta)^T(Y\beta) = n \end{aligned} \qquad (16)$$

Note:

$\alpha^T C_{xx}\alpha = 1$ and $\beta^T C_{yy}\beta = 1$: $\alpha\frac{1}{n}X^T X\alpha = 1$ and $\beta\frac{1}{n}Y^T Y\beta = 1$: $\alpha X^T X\alpha = n$ and $\beta Y^T Y\beta = n$.

In line with the conditions of CCA to find a pair of vectors α and β which yields set of composite variables $X\alpha$ and $Y\beta$ such that $Y\beta$ is most predictable from the variables of X . In other words, it seek a pair of vectors (α, β) that maximizes the corresponding lagrange function for equation (16) given as;

$$\begin{aligned} Q = -\frac{n}{2}\log\ 2\pi - \frac{1}{2}(Y\beta - X\alpha)'(Y\beta - X\alpha) - \\ -\frac{K}{2}(\alpha' X' X\alpha - n) - \frac{L}{2}(\beta' Y' Y\beta - n) \end{aligned}$$

where K and L denote the lagrange multipliers, λ_1, λ_2 that control the shrinkage that is applied to parameters α, β respectively.

The next step is to look for the first order condition, by taking derivatives with respect to α and β, and equating to zero. Then the first order conditions in equation(17) are;

$$\frac{\partial Q}{\partial \alpha} = X'Y\beta - (1+K)X'X\alpha - \frac{n}{2}\lambda_1 \text{sign}(\alpha) = 0 \tag{18}$$

and

$$\frac{\partial Q}{\partial \beta} = Y'X\alpha - (1+L)Y'Y\beta - \frac{n}{2}\lambda_2 \text{sign}(\beta) = 0 \tag{19}$$

Dividing equations (18) and (19) by n and multiplying by α' and β' it obtain the equations respectively as;

$$C_{xy}\beta - AC_{xx}\alpha - \frac{\lambda_1}{2}\text{sign}(\alpha) = 0 \tag{20}$$

and

$$C_{yx}\alpha - BC_{yy}\beta - \frac{\lambda_2}{2}\text{sign}(\beta) = 0 \tag{21}$$

where $A = 1 + K$ and $B = 1 + L$, shows that $A = B = \alpha^T C_{XY}\beta = \rho$. Consequently, yields

$$C_{xy}\beta - \rho C_{xx}\alpha - \frac{\lambda_1}{2}\text{sign}(\alpha) = 0 \tag{22}$$

and

$$C_{yx}\alpha - \rho C_{yy}\beta - \frac{\lambda_2}{2}\text{sign}(\beta) = 0 \tag{23}$$

The solution of the stationary equations (22) and (23) will yield the canonical correlation ρ and canonical variates α, β respectively. Thus, from equation (22) and (22) if let

$$MZ = W \tag{24}$$

where

$$M = \begin{pmatrix} -\rho^* C_{xx} & C_{xy} \\ C_{yx} & -\rho^* C_{yy} \end{pmatrix}, Z = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \text{ and } W = \begin{pmatrix} \frac{\lambda_1}{2}\text{sign}(\alpha) \\ \frac{\lambda_2}{2}\text{sign}(\beta) \end{pmatrix} \quad (25)$$

Thus seek ρ^* by the determinant property of block inverse matrices so that given a block matrix $P = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ where $A_{a\times a}$ and $D_{d\times d}$ square matrices are invertible and if $d < a$, it follows that (with $p \equiv a$) and $q \equiv d$

$$P^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix} = V \quad (26)$$

Hence, in case where $M = \begin{pmatrix} -\rho^* C_{xx} & C_{xy} \\ C_{yx} & -\rho^* C_{yy} \end{pmatrix}$

The matrix $V = M^{-1}$ can be further simplified by considering its blocks / submattrix as;

$$V_{11} = -\frac{1}{\rho^*}C_{xx}^{-1} - \frac{1}{\rho^*}C_{xx}^{-1}C_{xy}\left(-\rho^* C_{yy} + \frac{1}{\rho^*}C_{yx}C_{xx}^{-1}C_{xy}\right)^{-1} C_{yx}\frac{-1}{\rho^*}C_{xx}^{-1}$$
$$= -\frac{1}{\rho^*}C_{xx}^{-1}\left[1 - C_{xy}\left(\frac{1}{\rho^*}C_{yx}C_{xx}^{-1}C_{xy} - \rho C_{yy}\right)^{-1} C_{yx}C_{xx}^{-1}\right]$$

$$V_{12} = \frac{1}{\rho^*}C_{xx}^{-1}C_{xy}\left(\frac{1}{\rho^*}C_{yx}C_{xx}^{-1}C_{xy} - \rho C_{yy}\right)^{-1}$$

$$V_{21} = \frac{1}{\rho^*}\left(\frac{1}{\rho^*}C_{yx}C_{xx}^{-1}C_{xy} - \rho C_{yy}\right)^{-1} C_{yx}C_{xx}^{-1}$$

$$V_{22} = \left(\frac{1}{\rho^*}C_{yx}C_{xx}^{-1}C_{xy} - \rho C_{yy}\right)^{-1}$$

Therefore,

$$V = M^{-1} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$
$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = M^{-1}W = \begin{pmatrix} V_{11}\frac{\lambda_1}{2}\text{sign}(\alpha) & +V_{12}\frac{\lambda_2}{2}\text{sign}(\beta) \\ V_{21}\frac{\lambda_1}{2}\text{sign}(\alpha) & +V_{22}\frac{\lambda_2}{2}\text{sign}(\beta) \end{pmatrix}$$

$$\binom{\alpha}{\beta} = \begin{pmatrix} V_{11}\frac{\lambda_1}{2}\text{sign}(\alpha) + V_{12}\frac{\lambda_2}{2}\text{sign}(\beta) \\ V_{21}\frac{\lambda_1}{2}\text{sign}(\alpha) + V_{22}\frac{\lambda_2}{2}\text{sign}(\beta) \end{pmatrix} \quad (27)$$

The first p solution of $M^{-1}W$ are the estimates of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ while the remaining q solution of $M^{-1}W$ are the estimates of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$. Thus, we proposed a modified model-based CCA

4. Results and Discussion

In order to illustrate the effectiveness of our proposed CCA extension, we estimated the association between plant characteristics variables and yield components variables. We also analyzed a data regarding a nutrition study in mouse popularly known as Nutrimouse data set in order to estimate the relationship between the observed variables. A new technique, a modified CCA based on LASSO regression, has been introduced for discovering and interpreting the major sources of variability in a dataset. This was illustrated its usefulness in numerical optimization to estimate parameters and its ability in stabilizing results, removing non- informative and producing meaningful results that are interpretable. Results shown, through real life data, simulated data and independent data set that the developed method is capable of recovering certain types of underlying structure. Consequently, the solution of this modified maximization problem was implemented using Optim R statistical package. The codes require as input the correlation matrix R , the value of the tuning parameter and the number of components to be retained.

The results, also, show that when p (and q) $<$ n, (low-dimensional data set), the classical CCA (Hotelling,1935) the equivalent CCA regression formulation (Ayodeji and Obilade (2016)); and the modified CCA based on LASSO achieve the same classification performance. Furthermore, the developed model performs the best for the data set when p (and q) $>$ n (high-dimensional data set) even if multicollinearity is among variables.

This new formulation for canonical correlation analysis is based on LASSO technique. When applied to a low dimensional data which involves two sets of variables such as: plant characteristics been independent variables or X set and yield components been dependent variables or Yset obtained from the Nigerian Institute for Oil Palm Research (NIFOR), Benin City, Edo State, Nigeria, the modified CCA based on LASSO produced equivalent results with

the classical CCA and some notable equivalent CCA model formulation. In addition, the formulation was able to identify optimal combinations of the key variables on oil palm characteristics, thereby providing accurate prediction of its vegetative growth and development.

Summary Results for Data Set 1

Table 1. Canonical correlation of the various models on Oil Palm Data set

Model	Variable	Technique	Rho value
CCA (1935)	$(210, \mathrm{p} = 4, \mathrm{q} = 3)$	$\lambda_1 = 0, \lambda_2 = 0$	0.904
Ayodeji and Obilade (2016).	$(210, \mathrm{p} = 4, \mathrm{q} = 3)$	$\lambda_1 = 0, \lambda_2 = 0$	0.904
2023	$(210, \mathrm{p} = 4, \mathrm{q} = 3)$		

The modified CCA was also applied to a high dimensional dataset which served as validation data. This data involves two sets of variables measured on 40 mice: the gene been the independent variables or set X comprising expressions of 120 genes potentially involved in nutritional problems and Lipid been the dependent variables or set Y comprising concentrations of 21 hepatic acids. This was again compared to the newly developed model with the classical CCA and a notable equivalent CCA model formulation. Our proposed method was able to identify optimal combinations of the key variables, as well as produced results that are interpretable. However, the classical CCA and the notable equivalent CCA model formulation (Ayodeji and Obilade 2016) could not produce correlational results that are interpretable. Since our proposed method was able to produce interpretable results even with situations where the number of variables greatly exceeds the number of subjects, this means that our proposed method performs better for both low and high dimension data sets.

Summary Results for Data Set 2

Table 2. Canonical correlation of the various models on Nutrimouse Data Set.

Model	Variable	Technique	Rho
CCA (1935)	$(n = 40, p = 120, q = 21)$	$\lambda_1 = 0, \lambda_2 = 0$	-
Ayodeji and Obilade (2016).	$(n = 40, p = 120, q = 21)$	$\lambda_1 = 0, \lambda_2 = 0$	-
Witten et al., (2009). PMA	$(n = 40, p = 120, q = 21)$	$\lambda_1 = 0.3, \lambda_2 = 0.3$	0.880
2023	$(n = 40, p = 120, q = 21)$	$\lambda_1 = 0.8, \lambda_2 = 2.9$	0.997

5. Conclusion

This study has presented a new modification for the conventional model-based canonical correlation analysis (CCA) that will estimate the association between multiple dependent variables and multiple independent variables by employing the least absolute shrinkage and selector operator (LASSO) as an additional constraint in order to maximize the correlation between two canonical variate is developed. All so produced an Optim code in R package for the least square algorithm which was also modified. This will enable us estimate the weight vectors with interactive process that begins by first calculating an initial pair of CCA components based on an initial starting point and there after obtain a new set of the estimated weights and canonical correlation.

Finally, developed a criterion for the determination of the relationship between plant characteristics and yield components in oil palm population is developed. This will help in identifying the plant characteristics meaningful for developing a general and accurate prediction model for exploring the growth and development of oil palm species in Edo State,

Nigeria. Hence, this study presents a new formulation for canonical correlation analysis based on LASSO which produced equivalent results for low dimensional when compared to classical CCA. The approach deals with situations where the number of variables greatly exceeds the number of subjects. Hence, it is applicable to both the low and high dimension data. For the low dimension data, our proposed method was used in the determination of the relationship between plant characteristics been independent variables X set; [Plant Height (PH), Plant Diameter (PD),Number of Plant Leaf (NPL), Average Number of Plant Leaflets (ANPL)] and yield components been dependent variables Y set; [Leaf Area (LA), Fresh Fruit Bunch Weight (FFBW), Fresh Fruit Average Weight (FFAW)] of oil palm plants population collected from the Nigerian Institute For Oil Palm Research (NIFOR), Edo State, Nigeria. The formulation was able to identify optimal combinations of the key variables on oil palm characteristics for accurate prediction of its vegetative growth and development.

Furthermore, for the high dimensional data, the proposed method was used in a data regarding nutrition study in mouse (Nutrimouse Data Set). This data involves two sets of variables measured on 40 mice (see Witten et al.,(2009)): the gene been the independent variables of set X comprised expressions of 120 genes potentially involved in nutritional problems and the Lipid been the dependent variable set Y comprises concentrations of 21 hepatic acids. Hence, we compared the efficacy of our proposed method with that of the classical CCA from the perspective of formulation and optimisation and a notable equivalent CCA formulation, using two real data sets. Three CCA models were obtained for each data.

The first and second were built using the classical CCA and notable equivalent CCA model formulation, the third was built using our proposed method. All the analysis were done using the Optim R statistical package. Finally, the model was used to identify individual variables with optimal contributions so as to ascertain the plant characteristics that have relationship with yield components. This means identifying the plant characteristics meaningful for developing a general and accurate prediction model for exploring the growth and development of oil palm species in Edo State, Nigeria.

The results obtained for the two data sets were not only equivalent to that of the conventional model-based CCA, but were more interpretable for the second data set for which the number of variables greatly exceeds the number of subjects

REFRENCES

Ayodeji, I. O., & Obilade, T. O. (2016). A predictive model for canonical correlation analysis with implications for the simple and multiple correlations.

Bach, F. R., & Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis (Tech. Rep. No. 688). Department of Statistics, University of California, Berkeley.

Chen, X., Han, L., & Carbonell, J. (2012). Structured sparse canonical correlation analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (AISTATS) (pp. 199–207).

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology, 26*(2), 139.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*(3–4), 321–377.

Lai, P. L., & Fyfe, C. (2002). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems, 10*(5), 365–377.

Lin, D., Zhang, J., Li, V. D., Calhoun, H., Deng, W., & Wang, Y. P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics, 14*(1), 245.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2009). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings, 1*(Suppl 1), S7.

Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation*. Sage Publications.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., & Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology, 7*(1), Article 7

Witten, D. M., Tibshirani, R. J., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics, 10*(3), 515–534.

On the spectral density of the modified-ARFIMA model

A. Bello[1], M. Tasi'u[2], H.G. Dikko[2], and B.B. Alhaji[3]

[1]Department of Mathematical Sciences, Gombe State University, Gombe, Gombe State, Nigeria

[2]Department of Statistics, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

[3]Department of Mathematics, Nigerian Defence Academy, Kaduna, Kaduna State, Nigeria

Corresponding Author's email: bellobajoqaqsu@gmail.com

Abstract

This study develops Modified-ARFIMA Model and its spectral density for recursive sequence differencing operator that can handle large data in time series that have long memory characteristics.

Keywords: Spectral density, Modified-ARFIMA, and Long Memory

1. Introduction

Time series analysis uses spectral density analysis for a number of purposes, including data reduction, description, estimation, and hypothesis testing. Estimating spectral density functions and other spectral qualities associated with stationary multiple time series is crucial in any field where the features of the phenomenon under study can be described in terms of how it behaves in the frequency domain (Parzen, 1967). Nevertheless, Spectral analysis methods provide further tools for assessing how well different models fit the data (goodness-of-fit can be determined using sample spectra from the model's residuals) and for recommending potential models to fit (sample spectra often point out explanatory "variables" or "mechanisms" to incorporate into a time series). (Krampe & Paparoditis. 2022). The study modified the existing ARFIMA model of (Granger & Joyeux 1980: Hoskig 1981) using recursive sequence differencing operator that can handle large data set without truncation as observed by researchers such (Tanaka 1999, Rahman and Jibrin 2019). The modified model is presented in Section 2. The significance of this research lies in the fact that the new model will forecast and estimate the parameters more easily than the existing model. Additionally, the primary goal of this paper is to derive the spectral density of the modified model. The spectral density of the modified model has been derived, and the summary results are presented in Section 5.

2. Literature review

Granger and Joyeux (1980) investigate the infinite filter's fractional differencing terms, which correspond to the expansion of $(1\text{-}L)^d$. When the filter is applied to white noise, it creates a class of time series with different features, especially at low frequencies, which could be beneficial for long-memory forecasting. The aggregation of independent components is demonstrated to be a possible source of such models. The generation and estimation of these models are discussed, as well as their applicability to both simulated and real life data. The use of infinite binomial series expansion has a limitation in that it truncates huge data into little data and does not capture the integer difference parameter $d = 1$.

Hosking (1981) studies the generalization of ARIMA processes, commonly used in time series analysis, by considering fractional difference degrees. Introduce a fractional differencing operator that is based on an infinite binomial series expansion using the backward-shift operator. This allowed for the modeling of both long-term persistence and anti-persistence in time series. Long-term persistent processes are useful in fields such as economics and hydrology and the new family of models offers more flexibility in modeling both short and long memory behavior of a time series at the same time. However, the limitations are the same as those previously established by Granger and Joyeux in 1980.

Beran *et al.* (1999) explore SEMIFARMA models through a kernel-based approach to estimate a nonparametric deterministic trend. They employ an approximate maximum likelihood method to estimate the fractional and difference parameters, as well as the autoregressive coefficients. Using the iterative plug-in concept for determining bandwidth in nonparametric regression with extended memory, a data-driven approach is proposed for estimating the entire model. However, the SEMIFARMA model has three key drawbacks: it does not account for parametric trends, relies on binomial series expansion to simplify large datasets, and fails to capture the integer differencing parameter when d =1

Meerschaert, *et al.* (2014) present a new time series model for turbulent flow velocity data. To extend Kolmogorov's basic 5/3 spectral model, the new model uses tempered fractional calculus. To demonstrate the model's practical utility, it is applied to wind speed and water velocity in a huge lake. One limitation is that using binomial series expansion reduces large

amounts of data to small amounts of data; second, the integer differencing parameter d = 1 is not captured; and third, the seasonal differencing parameter is not captured.

Pumi *et al*. (2019) developed a beta ARFIMA model designed for continuous random variables that take values within the range (0,1). The model incorporates explanatory variables and captures long-range dependence in the time series data. The authors also demonstrate the consistency and asymptotic normality of the estimator under specific conditions. It supports hypothesis testing, diagnostic analysis, and forecasting. To evaluate the performance of the partial likelihood estimators in finite samples, they conducted a Monte Carlo simulation. However, the model has limitations, such as the non-discreteness of the random sample and its similarity to the work of Granger and Joyeux (1980).

In their study, Monge and Infante (2022) utilized autoregressive fractionally integrated moving average (ARFIMA) models to examine the historical data on crude oil prices and determine whether shocks to the series result in lasting or short-term effects. An ARFIMA (2,d,2) model was found to be optimal, with the fractional differencing parameter estimated at 0.4. However, the large confidence interval prevents a clear rejection of either the I(0) or I(1) hypotheses. The observed uncertainty is likely a result of structural breaks or nonlinear trends present in the dataset.

Elwasify *et al.* (2023) examined whether long-term memory exists in the daily movements of the EGX30 stock index. Their analysis used an ARFIMA model with a mix of parametric and semi-parametric techniques. To identify long memory, they used various statistical tools such as the rescaled range (R/S) statistic, aggregated variance, and absolute moments. For estimation, they relied on semi-parametric methods like the Geweke and Porter-Hudak (GPH) method, Reisen's (SPR) method, and the local Whittle (LW) estimator, while employing parametric techniques like MLE, exact maximum likelihood (EML), modified profile likelihood (MPL), and conditional sum of squares (CSS).

In 2024, Furuoka *et al*. introduced a novel approach for testing fractional integration, known as the autoregressive neural network–fractional integration (ARNN–FI) test. This method is based on a multilayer perceptron neural network framework, as outlined by Yaya et al. (Oxford Bulletin of Economics and Statistics, 83(4):960–981, 2021). The study discusses the asymptotic theory and properties of the new test. Monte Carlo simulations show that as the

sample size increases, the test's size and power distortions reduce. An empirical application of the test suggests that unemployment rates in three European countries do not show stationarity or mean reversion, supporting the hysteresis hypothesis.

Devianto *et al*. (2024) examined large-scale time series data for agricultural commodities by using an autocorrelation model that incorporates long-term trends, seasonal effects, and the impact of external variables. They chose chili prices as a case study due to their demonstration of long-term memory, seasonal variations, and sensitivity to external factors influencing production. These factors included specific periods, such as the month before the New Year and the week preceding the Eid al-Fitr celebration in Indonesia. To analyze price variations, the study applied the Seasonal Autoregressive Fractionally Integrated Moving Average (SARFIMA) model, which accounts for both seasonality and long-term memory. This model was enhanced by adding exogenous variables, resulting in the SARFIMAX model (SARFIMA with exogenous variables). A comparative analysis revealed that SARFIMAX outperformed SARFIMA in terms of accuracy, emphasizing the significance of incorporating both seasonal effects and historical price data along with external factors. The enhanced SARFIMAX model provided a deeper understanding of price dynamics, offering valuable insights for sustaining chili supply chains in the era of big data.

Swarnalatha *et al.* (2024) investigated the presence of long memory in time series, which is characterized by an autocorrelation function that decays slowly or follows a hyperbolic pattern. To accurately capture this behavior, they applied the autoregressive fractionally integrated moving average (ARFIMA) model, well-suited for analyzing historical financial data. The study focused on assessing the ARFIMA model's effectiveness in modeling long memory processes, using the Geweke and Porter-Hudak (GPH) method for parameter estimation. They analyzed monthly groundnut prices in Andhra Pradesh from January 2002 to December 2023. The ARFIMA (1, 0.43, 1) model provided the best fit, demonstrating strong short-term forecasting performance. It closely aligned with actual prices and outperformed the SARIMA(1,1,3)(0,1,2)12 model in terms of AIC, MSE, and RMSE. The findings indicated that the ARFIMA model offered more accurate price predictions than the SARIMA model.

3. Methodology

This research intended to introduce a Modified-ARFIMA model that could be used to study the large financial and economic time series data. Given a time series Y_1, ...,Y_m, the observations are assumed to be trendy, non-stationary and long memory. Also, they are assumed to have positive autocorrelation and long memory denoted as d. The d is assumed to be in the range $1 < d < 1.5$. Given a series Y_m, $m \geq 0$ that is assumed or exhibit trendy (increasing and decreasing) and long memory then. The fractional filter can be defined as follows: equation (1) is the differencing operator,

$$\nabla Y_m = Y_m - Y_{m-1} \tag{1}$$

therefore,

$$d(\nabla Y_{m-1}) = d(Y_{m-1} - Y_{m-2}) \tag{2}$$

minus equation (2) from equation (1) become

$Q_m = \nabla Ym - d(\nabla Y_{m-1})$

where Q_m is the filter, ∇ is the differencing operator and d is a sequence fractional differencing operator.

Note,

$\nabla Y_{m-1} = Y_{m-1} - Y_{m-2}$

$= Y_m - Y_{m-1} - d(Y_{m-1} - Y_{m-2})$

$$Q_m = Y_m - dY_{m-1} - Y_{m-1} + dY_{m-2} \tag{3}$$

Using lag operators, the fractional filter in equation (3) can be represented as follows:

$= L^0Ym - dL^1Ym - L^1Ym + dL^2Ym$

$= \{1(1 - L) - dL(1 - L)\}Y_m$

$$Q_m = \{(1 - L)(1 - dL)\}Y_m \tag{4}$$

The method for obtaining the fractional filters to induce nonstationarity is as follows:

$$Q_m = \{(1 - L)(1 - dL)\}Y_m = Y_m - Y_{m-1} - d(Y_{m-1} - Y_{m-2}) \tag{5}$$

$$Q_{m-1} = \{(1 - L)(1 - dL)\}Y_{m-1} = Y_{m-1} - Y_{m-2} - d(Y_{m-2} - Y_{m-3}) \tag{6}$$

$$Q_{m-n} = \{(1 - L)(1 - dL)\}Y_{m-n} = Y_{m-n} - Y_{m-n-1} - d(Y_{m-n-1} - Y_{m-n-2}) \tag{7}$$

$= Y_{m-n} - Y_{m-n-1} - d(Y_{m-n-1} - Y_{m-n-2})$

$$= Y_{m-n} - Y_{m-n-1} - dY_{m-n-1} + dY_{m-n-2} \quad (8)$$

then, by applying the lag operator in equation (8) (see Rahman and Jibrin, 2019), we can have

$$= L^n Y_m - dL^{n+1} Y_m - L^{n+1} Y_m + dL^{n+2} Y_m \quad (9)$$

factor Y_m that is common in equation (9) we can have

$$=(L^n - dL^{n+1} - L^{n+1} + dL^{n+2})Y_m \quad (10)$$

Finally, the proposed fractional filter has the following general form:

$$Q_m = \sum_{n=0}^{N} (L^n - dL^{n+1} - L^{n+1} + dL^{n+2}) Y_m \quad (11)$$

The ARMA model of Whittle (1951) $\varphi(L)Y_m = \theta(L)\varepsilon_m$ which became known to researches in time series is given as follows in the book of Box and Jenkins's (1970).

$$\phi(L)Y_m = \theta(L)\varepsilon_m \quad (12)$$

$$\varphi(L)(Q_m)Y_m = \theta(L)\varepsilon_m \quad (13)$$

where $\phi(L)$ and $\theta(L)$ are characteristic polynomials of AR and MA process, Q_m is the sequence fractional filter, Y_m is the series, L is the lag operator, and $\varepsilon_m \sim WN(0, \sigma^2)$. The lag representation of the proposed Modified-ARFIMA model shown below in equation (14)

$$\varphi(L)\{(1 - L)(1 - dL)\}Y_m = \theta(L)\varepsilon_m \quad (14)$$

Where $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ and $\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q$ are characteristic polynomials of AR and MA process, d is the fractional differencing filter, L is the backward shift operator, and $\varepsilon_m \sim WN(0, \sigma^2)$ (Rahman and Jibrin, 2019).

4. The Spectral Density of the Modified-ARFIMA Model

The spectral density is a frequency domain representation of a time series that is directly related to the autocovariance time domain representation. In addition, it can be used to define a function of frequency instead of time, do regression, in which we regress, and split a time series into periodic components (Shumway and Stoffer, 2000).

In order to calculate or derive the Modified-ARFIMA model spectral density function can be recall from equation (14)

$$\varphi(L)\{(1-L)(1-dL)\}Y_m = \theta(L)\varepsilon_m \quad (15)$$

the spectral density for the above equation (15) (see Cryer and Kung, 2015), can be represented as this

$$S(\lambda) = \frac{\sigma^2}{2\pi}\left|\frac{\theta(e^{-i\lambda})}{\varphi(e^{-i\lambda})(e^{-i\lambda})d(e^{-i\lambda})}\right|^2 \quad (16)$$

$$= \frac{\sigma^2\left|\theta(e^{-i\lambda})\right|^2}{2\pi\left|\varphi(e^{-i\lambda})\right|^2\left|(e^{-i\lambda})\right|^2\left|d(e^{-i\lambda})\right|^2} \quad (17)$$

then, let consider the numerator of the above equation (17) and prove the spectral.

$$\left|\theta(e^{-i\lambda})\right|^2 = \left|(1-\theta e^{-i\lambda})\right|^2 = (1-\theta e^{-i\lambda})(1-\theta e^{i\lambda}) \quad (18)$$

$= (1 - \theta e^{-i\lambda})(1 - \theta e^{i\lambda})$

$= (1 - \theta^{i\lambda} - \theta e^{-i\lambda} + \theta 2e^{-i\lambda} \times e^{i\lambda})$

$= 1 - {}_{\theta}e^{i\lambda} - {}_{\theta}e^{-i\lambda} + {}_{\theta}2e^{-i\lambda+i\lambda}$

$= 1 - \theta e^{i\lambda} - \theta e^{-i\lambda} + \theta 2e^{0}$

$= 1 - \theta[e^{i\lambda} + e^{-i\lambda}] + \theta 2 \times 1$

$= 1 + \theta 2 - \theta[e^{i\lambda} + e^{-i\lambda}]$

note that, $e^{i\lambda}=\cos(i_\lambda) + i\sin(i_\lambda)$, also $e^{-i\lambda} = \cos(-i_\lambda) + i\sin(-i_\lambda)$

$= 1 + \theta^2 - \theta[\cos(i\lambda) + i\sin(i\lambda) + \cos(-i\lambda) + i\sin(-i\lambda)]$

note that, $\cos(-i\lambda)=\cos(i\lambda)$ but $i\sin(-i\lambda)=-i\sin(i\lambda)$

$= 1 + \theta^2 - \theta[\cos(i\lambda) + i\sin(i\lambda) + \cos(i\lambda) - i\sin(i\lambda)]$

$= 1 + \theta^2 - \theta[\cos(i\lambda) + \cos(i\lambda)]$

$= 1 + \theta^2 - \theta[2\cos(i\lambda)]$

$= 1 + \theta^2 - 2\theta\cos(i\lambda)$

secondly let consider the denominator (19)

$$\left|\varphi(e^{-i\lambda})(e^{-i\lambda})d(e^{-i\lambda})\right|^2 = \left|\varphi(e^{-i\lambda})\right|^2\left|(e^{-i\lambda})\right|^2\left|d(e^{-i\lambda})\right|^2 \quad (20)$$

$$\left|\varphi(e^{-i\lambda})\right|^2 = \left|(1-\varphi e^{-i\lambda})\right|^2 = (1-\varphi e^{-i\lambda})(1-\varphi e^{i\lambda})$$

$(1 - \phi e^{-i\lambda})(1 - \phi e^{i\lambda}) = 1 - \phi e^{i\lambda} - \phi e^{-i\lambda} + \phi 2e^{-i\lambda} \times \phi e^{i\lambda}$

$= 1 - \phi e^{i\lambda} - \phi e^{-i\lambda} + \phi 2e^{-i\lambda+i\lambda}$

$= 1 - \phi e^{i\lambda} - \phi e^{-i\lambda} + \phi 2e^{0}$

$= 1 - \phi e^{i\lambda} - \phi e^{-i\lambda} + \phi 2 \times 1$

$= 1 + \phi 2 - \phi[e^{i\lambda} + e^{-i\lambda}]$

$= 1 + \phi^2 - \phi[\cos(i\lambda) + i\sin(i\lambda) + \cos(-i\lambda) + i\sin(-i\lambda)]$

$= 1 + \phi^2 - \phi[\cos(i\lambda) + i\sin(i\lambda) + \cos(i\lambda) - i\sin(i\lambda)]$

$= 1 + \phi^2 - \phi[\cos(i\lambda) + \cos(i\lambda)]$

$= 1 + \phi^2 - \phi[2\cos(i\lambda)]$

$= 1 + \phi^2 - 2\phi\cos(i\lambda)$ (21)

the second prove for spectral denominator

$$\left|e^{-i\lambda}\right|^2 = \left|1 - e^{-i\lambda}\right|^2 = (1 - e^{-i\lambda})(1 - e^{i\lambda}) \quad (22)$$

$(1 - e^{-i\lambda})(1 - e^{i\lambda}) = (1 - e^{-i\lambda} - e^{i\lambda} + e^{-i\lambda} \times e^{i\lambda})$

$= 1 - e^{-i\lambda} - e^{i\lambda} + {}_{e}{}^{-i\lambda+i\lambda}$

$= 1 - e^{-i\lambda} - e^{i\lambda} + e^{0}$

$= 1 - e^{-i\lambda} - e^{i\lambda} + 1$

$= 2 - [e^{i\lambda} + e^{-i\lambda}]$

$= 2 - [\cos(i\lambda) + i\sin(i\lambda) + \cos(-i\lambda) + i\sin(-i\lambda)]$

$= 2 - [\cos(i\lambda) + i\sin(i\lambda) + \cos(i\lambda) - i\sin(i\lambda)]$

$= 2 - [\cos(i\lambda) + \cos(i\lambda)]$

$= 2 - 2\cos(i\lambda)$

thirdly prove of spectral denominator (23)

$$\left|de^{-i\lambda}\right|^2 = \left|(1 - de^{-i\lambda})\right|^2 = (1 - de^{-i\lambda})(1 - de^{i\lambda}) \quad (24)$$

$= (1 - de^{-i\lambda})(1 - de^{i\lambda)} = (1 - de^{-i\lambda} - de^{i\lambda} + d2e^{-i\lambda} \times e^{i\lambda})$

$= 1 - de^{-i\lambda} - de^{i\lambda} + d2e^{-i\lambda+i\lambda}$

$= 1 - de^{-i\lambda} - de^{i\lambda} + d2e^{0}$

$= 1 - de^{-i\lambda} - de^{i\lambda} + d^2 \times 1$

$= 1 + d2 - d[e^{i\lambda} + e^{-i\lambda}]$

$= 1 + d^2 - d[\cos(i_\lambda) + i\sin(i_\lambda) + \cos(-i_\lambda) + i\sin(-i_\lambda)]$

$= 1 + d^2 - d[\cos(i\lambda) + i\sin(i\lambda) + \cos(i\lambda) - i\sin(i\lambda)]$

$= 1 + d^2 - d[\cos(i\lambda) + \cos(i\lambda)]$

$= 1 + d^2 - d[2\cos(i_\lambda)]$

$= 1 + d^2 - 2d\cos(i_\lambda)$ (25)

Therefore, the spectral density for the Modified-ARFIMA model is presented by the combination of the following equations (19), (21), (23) and (25) (Cryer and Kung, 2015) below

$$S(\lambda) = \frac{\sigma^2[1+\theta^2 - 2\theta\cos(i\lambda)]}{2\pi[1+\varphi^2 - 2\varphi\cos(i\lambda)][2-2\cos(i\lambda)][1+d^2 - 2d\cos(i\lambda)]} \quad (26)$$

5. Summary

The spectral density for Modified Autoregressive Fractional Integrated Moving Average MARFIMA (p, d, q) model has three components. These components are Moving Average component denoted by θ_j, Autoregressive component denoted by ϕ_j, and Fractional filter (1-L)(1-dL) component. The spectral density component for modified ARFIMA model as derived in section 3 are summarized and for crystal clear are given follows.

i. MA(q) = $1 + \theta^2 - 2\theta\cos(i\lambda)$

ii. AR(p) = $1 + \phi^2 - 2\phi\cos(i\lambda)$

iii. Fractional Filter = $[2 - 2\cos(i\lambda)]\,[1 + d^2 - 2d\cos(_{ij})]$

References

Agu, F. I., Mačutek, J., & Szűcs, G. (2023). A simple estimation of parameters for discrete distributions from the Schröter family. *Statistika: Statistics & Economy Journal, 103*(2).

Beard, R. E., Pentikäinen, T., & Pesonen, E. (1977). *Risk theory* (2nd ed.). Chapman and Hall.

Beran, J. (1999). SEMIFAR models: A semiparametric fractional framework for modelling trends, long-range dependence and nonstationarity [Preprint]. University of Konstanz.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation, 19*(90), 297–301.

Devianto, D., Wahyuni, E., Maiyastri, M., & Yollanda, M. (2024). The seasonal model of chili price movement with the effect of long memory and exogenous variables for improving time series model accuracy. *Frontiers in Applied Mathematics and Statistics, 10,* 1408381.

Dickson, D. C. (2016). *Insurance risk and ruin*. Cambridge University Press.

Dzidzornu, S. B., & Minkah, R. (2021). Assessing the performance of the discrete Generalised Pareto distribution in modelling non-life insurance claims. *Journal of Probability and Statistics, 2021,* 5518583.

Elwasify, A., & Isa, Z. (2023). Estimation of the difference parameter in ARFIMA models using parametric and semiparametric methods. *International Journal of Science, Mathematics and Technology Learning, 31*(2).

Furuoka, F., Gil-Alana, L. A., Yaya, O. S., Aruchunan, E., & Ogbonna, A. E. (2024). A new fractional integration approach based on neural network nonlinearity with an application to testing unemployment hysteresis. *Empirical Economics, 66*(6), 2471–2499.

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis, 1*(1), 15–29.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika, 68*(1), 165–176.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology, 26*(2), 139–142.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika, 28*(3–4), 321–377.

Jonathan D. Cryer & Kung-Sick Chen. (2015). *Time series analysis with applications in R* (2nd ed.). Springer

Krampe, J., & Paparoditis, E. (2022). Frequency domain statistical inference for high-dimensional time series. *arXiv*.

Lai, P. L., & Fyfe, C. (2002). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems, 10*(5), 365–377.

Lin, D., Zhang, J., Li, V. D., Calhoun, H., Deng, W., & Wang, Y. P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics, 14*(1), 245.

Meerschaert, M. M., Sabzikar, F., Phanikumar, M. S., & Zeleke, A. (2014). Tempered fractional time series model for turbulence in geophysical flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(7), P07002.

Monge, M., & Infante, J. (2023). A fractional ARIMA (ARFIMA) model in the analysis of historical crude oil prices. *Energy Research Letters, 4*(1).

Packová, V., & Brebera, D. (2015). Loss distributions in insurance risk management. In *Recent Advances on Economics and Business Administration* (pp. 17–22).

Panjer, H. H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin: The Journal of the IAA, 12*(1), 22–26.

Parzen, E. (1967). The role of spectral analysis in time series analysis. *Revue de l'Institute International de Statistique, 125*(1), 125–141.

Pumi, G., Valk, M., Bisognin, C., Bayer, F. M., & Prass, T. S. (2019). Beta autoregressive fractionally integrated moving average models. *Journal of Statistical Planning and Inference, 200*, 196–212.

Rahman, R. A., & Jibrin, S. A. (2019). Modeling and forecasting Tapis crude oil price: A long memory approach. In *Proceedings of the International Conference on Mathematical Sciences and Technology* (MathTech2018), Pulau-Penang, Malaysia (AIP Conf. Proc. 2184).

Shumway, R. H., & Stoffer, D. S. (2000). *Time series analysis and its applications: With R examples*. Springer.

Swarnalatha, P., Srinivasa Rao, V., Raghunadha Reddy, G., Rathod, S., Ramesh, D., & Uma Devi, K. (2024). Modelling long memory time series for groundnut prices in Andhra Pradesh with autoregressive fractionally integrated moving average for forecasting. *Journal of Scientific Research and Reports, 30*(7), 289–302.

Tanaka, K. (1999). The nonstationary fractional unit root. *Econometric Theory, 15*(4), 549–582

Whittle, P. (1951). *Hypothesis testing in time series analysis* (Doctoral dissertation). Uppsala University.

Yaya, O. S., Ogbonna, A. E., Furuoka, F., & Gil-Alana, L. A. (2021). A new unit root test for unemployment hysteresis based on the autoregressive neural network. *Oxford Bulletin of Economics and Statistics, 83*(4), 960–981.

Model -Robust Regression 2: A Modification of the Model for Selecting Locally Adaptive Mixing Parameters

***Efosa Edionwe[1*] and Omo Eguasa[2]**

[1]Department of Statistics, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria

[2]Department of Physical Sciences, Benson Idahosa University, Benin City, Edo State, Nigeria

***Corresponding Author. Email:** edionwe.efosa@fupre.edu.ng*

Abstract

Model Robust Regression 2 (MRR2) is a semi-parametric regression model applied in the data modeling phase of studies that involve small sample size such as in response surface methodology (RSM). MRR2 combines estimates from both the ordinary least squares (OLS) and local linear regression (LLR) via a mixing parameter. For data of sample size n, MRR2 performs better when the combination of estimates from both the OLS and LLR is done using locally adaptive mixing parameters, (LAMPs) rather than the fixed or global counterpart. In the application of MRR2 in response surface studies, the only model for selecting LAMPs was developed as a strictly increasing function of the absolute values of the OLS residuals at each data point. Ideally, a model for generating LAMPs ought to return a zero mixing parameter at a data point where the OLS residual is zero since it is needless to shrink or supplement the OLS estimate with portion of the LLR estimate at this data point. The existing model for LAMPs fails in this regard, returning zero mixing parameters only when the OLS residuals are all zero. This scenario is quite unrealistic. Therefore, we propose a reformation of the existing model for selecting LAMPs in order to address this shortcoming. Two problems from the literature are analyzed. The use of the LAMPs from the proposed model guarantees comparatively better goodness-of-fit and optimal solutions.

Keywords: desirability function, local bandwidths, local linear regression, locally adaptive mixing parameters, response surface methodology, semi-parametric regression

1. Introduction

For a single response (output, dependent) variable y, which may represent the observable output of a process or system, the overall goal of Response Surface Methodology (RSM) is finding the optimal setting of the k input or explanatory variables, $x_1, x_2, \ldots, x_k$, that optimizes y modelled as a function of the k explanatory variables using data from a designed experiment

(Box & Wilson, 1951; Castillo, 2007). In the modeling phase, it is assumed that the relationship between the two categories of variables takes the form:

$$y_i = f\left(x_{i1,}\ x_{i2}, \dots, x_{ik}\right) + \varepsilon_i, \quad i = 1,2, \dots, n \tag{1}$$

where the mean function f denotes the true but unknown relationship between the variables $x_{ij,}\ i = 1, \dots, n, j = 1, \dots, k,$ are the values of the j^{th} explanatory variable at the i^{th} data point, $\varepsilon_i,\ i = 1,2, \dots, n,$ are random error terms with the assumption that $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$, and n is the sample size (Montgomery, 2005; Myers et al, 2009; He et al, 2012).

Existing classes of regression models applied in the estimation of the unknown function f in (1) include the parametric regression model, the nonparametric regression model and the semi-parametric regression model (Anderson-Cook & Prewitt, 2005; Mays & Birch, 2002). However, for small-sample setting which is typical of RSM, the semi-parametric Model Robust Regression 2 (MRR2) is currently considered one of the best regression models (Mays et al, 2001; Pickle et al, 2008).

MRR2 is ideal in situations where a researcher has partial knowledge of a low-order polynomial that can be used in estimating f in (1). Because this assumed model is only fairly adequate, it is deficient in capturing local trends and patterns in the entire range of data (Mays et al, 2001). Hence, the researcher looks to supplement the estimates from the parametric model by adding portions of estimates from the LLR.

Mathematically, the MRR2 estimate, $\hat{y}_i^{(MRR2)}$ of y_i, $i = 1,2, \dots, n,$ is given by:

$$\hat{y}_i^{(MRR2)} = \boldsymbol{x_i}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} + \lambda_i \tilde{\mathbf{x}}_\mathbf{i}(\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}\tilde{\boldsymbol{X}}^{\boldsymbol{T}})^{-1}\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}(\boldsymbol{I} - (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})\boldsymbol{y}, \tag{2}$$

$$= \boldsymbol{x_i}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} + \lambda_i \tilde{\mathbf{x}}_\mathbf{i}\left(\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}\boldsymbol{r}^{(\boldsymbol{OLS})}, \tag{3}$$

where λ_i, $0 \leq \lambda_i \leq 1$, is the locally adaptive mixing parameter (LAMP) for combining the OLS and the LLR estimates in the i^{th} data point, $\boldsymbol{x_i}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$ is the OLS component, x_i is the i^{th} row vector of the $n \times p$ OLS model matrix $\boldsymbol{X}$, where p is the number of OLS model parameters, y is $n \times 1$ vector of response, $\tilde{\mathbf{x}}_\mathbf{i}(\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}\tilde{\boldsymbol{X}}^{\boldsymbol{T}})^{-1}\tilde{\boldsymbol{X}}^{\boldsymbol{T}}\boldsymbol{W_i}(\boldsymbol{I} - (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T})\boldsymbol{y}$ is the LLR component, $\tilde{\mathbf{x}}_\mathbf{i}$ is the i^{th} row of the LLR model matrix $\tilde{\boldsymbol{X}}$ given as:

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} 1 & \tilde{x}_{11} & \tilde{x}_{12} & \cdots & \tilde{x}_{1k} \\ 1 & \tilde{x}_{21} & \tilde{x}_{22} & \cdots & \tilde{x}_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{x}_{n1} & \tilde{x}_{n2} & & \tilde{x}_{nk} \end{bmatrix},$$

where $\tilde{x}_{ij}$, $i = 1,2,\dots,n, j = 1,2,\dots,k$, denotes the value of the j^{th} explanatory variable in the i^{th} data point, $\boldsymbol{I}$ is an $n \times n$ identity matrix, $\boldsymbol{W_i}$ is a $n \times n$ diagonal weights matrix for getting the LLR estimate at the i^{th} data point, $\boldsymbol{r^{(OLS)}}$ is the $n \times 1$ vector of the OLS residuals, $\boldsymbol{X^T}$ and $\boldsymbol{\tilde{X}^T}$ are the transposed model matrices of the OLS and LLR, respectively (Wan & Birch, 2011; Edionwe et al., 2018).

The r^{th} entry, w_r of weight matrix, $\boldsymbol{W_i}$ is obtained from the Gaussian product kernel as:

$$w_r^{**} = \prod_{j=1}^{k} K\left(\frac{x_{ij}-x_{rj}}{b_i}\right) / \sum_{i=1}^{n}\prod_{j=1}^{k} K\left(\frac{x_{ij}-x_{rj}}{b_i}\right), \quad i = 1,2,\dots,n, j = 1,2,\dots,k, \qquad (4)$$

where $K\left(\frac{x_{ij}-x_{rj}}{b_i}\right) = e^{-\left(\frac{x_{ij}-x_{rj}}{b_i}\right)^2}$ is the simplified Gaussian kernel and b_i, $0 < b_i \leq 1, i = 1,2,\dots,n$, are the locally adaptive bandwidths (Pickle et al., 2008; Edionwe & Mbegbu, 2014; Edionwe & Eguasa, 2023).

For data emanating from response surface studies, a model for selecting locally adaptive bandwidths presented in (Edionwe et al., 2016) and (Edionwe et al., 2017) is given by:

$$b_i = \frac{b^*\{N(TC-y_i)\}}{T(Cn-1)}, \quad i = 1,\dots,n \qquad (5)$$

where b* (which can be taken as 1 order to reduce computational burden) is the fixed optimal bandwidth, $T = \sum_i^n y_i$, $C \geq 0$, $N > 0$ are data-driven tuning parameters whose optimal values C^*, N^*, respectively generate a vector, Φ of n locally adaptive optimal bandwidths $[b_1^*, \dots, b_n^*]$.

In situations where the mixing parameters $\lambda_i, i = 1,2,\dots,n,$ in (2) and (3) are chosen such that $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$, we say that a global or fixed mixing parameter λ is selected. An existing model for selecting LAMPs proposed in (Edionwe et al., 2017) is reviewed in Section 2.

The MRR2 as presented in equation (2) may be expressed in matrix form as:

$$\boldsymbol{y^{(MRR2)}} = \begin{bmatrix} h_1^{(MRR2)} \\ h_2^{(MRR2)} \\ \vdots \\ h_n^{(MRR2)} \end{bmatrix} y = \boldsymbol{H^{(MRR2)}y}, \qquad (6)$$

where $\boldsymbol{h_i^{(MRR2)}} = \boldsymbol{x_i(X^TX)^{-1}X^T} + \boldsymbol{\lambda_i \tilde{x}_i(\tilde{X}^T W_i^* \tilde{X}^T)^{-1}\tilde{X}^T W_i^*(I - X(X^TX)^{-1}X^T)}$ is the i^{th} row of the $n \times n$ MRR2 Hat matrix, $\boldsymbol{H^{(MRR2)}}$.

Given small-sample settings, Mays et al. (2001) and Mays & Birch (2002) presented a bandwidth selection criterion referred to as Penalized Prediction error Sum of Squares (PRESS**) and given by:

$$PRESS^{**}(\Phi) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(MRR2)}(\Phi)\right)^2}{n - trace\left(H^{(MRR2)}(\Phi)\right) + (n-k-1)\frac{SSE_{max} - SSE_{\Phi}}{SSE_{max}}}, \quad (7)$$

where SSE_{max} is the maximum Sum of Squared Errors obtained as $b_1, b_2, \ldots, b_n$ tend to infinity, SSE_{Φ} is the Sum of Squared Errors (SSE) for a specific vector of bandwidths $\Phi = [b_1, b_2, \ldots, b_n]$, $tr\left(H^{(MRR2)}\right)$ is the trace of the MRR2 hat matrix and $\hat{y}_{i,-i}^{(MRR2)}(\Phi)$ is the leave-one-out cross-validation estimate of y_i with the i^{th} observation left out.

Similarly, a version of the PRESS** for choosing the mixing parameter given by Mays & Birch (2002) and adopted in Edionwe et al. (2017) for the selection of the existing LAMPs is given by:

$$PRESS^{**}([\lambda_1, \lambda_2, \ldots, \lambda_n]) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(MRR2)}(\Phi^*, [\lambda_1, \lambda_2, \ldots, \lambda_n])\right)^2}{n - tr\left(H^{(MRR2)}(\Phi^*, [\lambda_1, \lambda_2, \ldots, \lambda_n])\right) + (n-k-1)\frac{SSE_{max} - SSE_{\Phi^*[\lambda_1, \lambda_2, \ldots, \lambda_n]}}{SSE_{max}}}, \quad (8)$$

where $\Phi^* = [b_1^*, b_2^*, \ldots, b_n^*]$ denotes locally optimal bandwidths that minimizes (7) above, $SSE_{\Phi^*[\lambda_1, \lambda_2, \ldots, \lambda_n]}$ is the SSE for optimal bandwidths associated with a given vector of mixing parameters, $[\lambda_1, \lambda_2, \ldots, \lambda_n]$, $tr\left(H^{(MRR2)}(\Phi^*, [\lambda_1, \lambda_2, \ldots, \lambda_n])\right)$ is the trace of MRR2 hat matrix given Φ^* and $[\lambda_1, \lambda_2, \ldots, \lambda_n]$, $DF = n - tr\left(H^{(MRR2)}(\Phi^*, [\lambda_1, \lambda_2, \ldots, \lambda_n])\right)$, and $\hat{y}_{i,-i}^{(MRR2)}(\Phi^*, [\lambda_1, \lambda_2, \ldots, \lambda_n])$ is the leave-one-out cross-validation of MRR2 estimate of y_i given the optimal bandwidths and a vector of mixing parameters, $[\lambda_1, \lambda_2, \ldots, \lambda_n]$.

For a given data, the vector of mixing parameters that minimizes (8) gives the vector of optimal mixing parameter, $\boldsymbol{\lambda}^* = [\lambda_1^*, \lambda_2^*, \ldots, \lambda_n^*]$.

Once the mean function, f, in (1) has been modelled, the resulting fitted curve is subjected to some optimization routine in order to obtain the values of the explanatory variables that optimize the response based on the prescribed production requirements. This task summarizes the aim of the optimization phase of RSM (He et al., 2012; Yeniay, 2014).

For multiple response studies involving $m > 1$ response variables, $y_1, y_2, \ldots y_m$, we need to obtain the optimal values (setting) of the explanatory variables that simultaneously optimize all the response variables based on their individual production requirements.

An efficient optimization criterion for achieving this task is the desirability function (Derringer & Suich, 1980, Harrington, 1965). Given the production requirement of a response, the desirability function transforms the estimated response, $\hat{y}(\boldsymbol{x})$ into a response-type dependent scalar measure, $d(\hat{y}(\boldsymbol{x}))$.

The three individual desirability functions for converting the estimated response into scalar measures are given in many of the literature for instance in Edionwe et al. (2018) and Wan & Birch (2011).

The overall objective of the desirability criterion is to obtain the values of the explanatory variables that simultaneously maximize the geometric mean, D, of all the individual desirability measures, where D is given as:

$$D = \text{maximize}\{[d_1 \times d_2 \times \ldots \times d_m]^{1/m}\}, \qquad (9)$$

where $d_1, \ldots, d_m$ are the individual scalar measures of the m response variables in the study (Derringer & Suich, 1980; Edionwe et al., 2017).

In this paper, all the optimization tasks in (7), (8) and (9) were done using the Genetic Algorithm (GA) optimization toolbox in Matlab.

The GA is an evolutionary optimization tool applied in solving variety of problems including the ones in which the objective function lacks closed form expression such as the MRR2 model (Holland, 1975; Wan & Birch, 2011).

2. An Overview of Existing Model for Selecting Locally Adaptive Mixing Parameters

Given two data points, $i = g$, and $i = h$. By intuition, the OLS residual, $y_g - \hat{y}_g^{(OLS)} = -e_g$ at a data point g, indicates the same degree of model misspecification in the assumed polynomial as the OLS residual, $y_h - \hat{y}_h^{(OLS)} = +e_h$ at a data point h. Therefore, both data points require the same proportion of LLR estimate to supplement or shrink the OLS estimate in order to correct the inadequacy in the assumed model. This intuition inspired the derivation of a model for selecting LAMPs as a strictly increasing function of OLS residuals in Edionwe et al. (2017), with an understanding that a data point with relatively smaller absolute value of OLS residual, $e_i = \left|y_i - \hat{y}_i^{(OLS)}\right|$, $i = 1,2, \ldots, n$, signifies a data point with relatively higher performance of the OLS model and thus requires a relatively smaller mixing parameter to make correction for the inadequacy of the OLS model, and vice versa. Therefore, in Edionwe et al. (2017), the derived model for generating LAMPs is given by:

$$\lambda_i(e_i) = \frac{N_0(T_0C_0+e_i)}{T_0(C_0n+1)}, \quad i = 1,2,\dots,n, \tag{10}$$

where $T_o = \sum_i^n e_i$, $e_i = |y_i - y_i^{(OLS)}|$, $C_o \geq 0$, $N_o \geq 0$, are data-driven tuning parameters. The optimal values of C_o and N_o are herein designated as C_o^*, N_o^*, respectively. The C_o^* and N_o^*, together with T_o and e_i in (10), generate a vector $\boldsymbol{\lambda}^*$ of n optimal LAMPs, $[\lambda_1^*(e_1), \lambda_2^*(e_2), \dots, \lambda_n^*(e_n)]$ based on the minimization of (8)**.**

Results from the application of (10) show considerable improvement in both goodness-of-fit and desirability function when compared with results from the fixed or global mixing parameter. However, a look at (10) reveals that if there exists a data point where $e_i = 0$ (which implies a data point that requires zero proportion of LLR estimate), then for that data point to be rightly assigned a zero mixing parameter, either N_o or C_o or both must be zero.

We highlight the shortcomings of such choices of values for N_o and C_o in the context that not all the data points have zero residual:

One, $N_o = 0$ in (10) implies that the n mixing parameters, $\lambda_i(e_i)$, $i = 1,2,\dots,n$, will all be zero irrespective of the magnitude of $e_i = |y_i - y_i^{(OLS)}|$ at each data point. This is acceptable only in cases where the OLS residuals are all zero. This is a very unlikely situation in reality.

Two, for $C_o = 0$, we get $\lambda_i(e_i) = \frac{N_0(0+e_i)}{T_0(0+1)} = \frac{N_o e_i}{T_o}$, $\quad i = 1,2,\dots,n$, leading to the deformation in the configuration of the model in (10) with its attendant complications. For instance, if the residual at a data point say t is larger than 1, then N_o must be strictly less than T_o in order for the mixing parameter $\lambda_t(e_t)$ at the data point t to lie in the interval [0,1]. Unfortunately, this restricted value of N_o that would ensure that $\lambda_t(e_t) \in [0,1]$ might not be optimal or even near-optimal for the residuals in the remaining$(n-2)$ data points. This problem could be more acute if the number of data points with OLS residuals less than 1 is more than the number of data points with OLS residuals greater than 1, because this would mean that fewer data points are deciding the value of N_0 meant for the entire data points.

The aim of this paper is to modify the existing model for generating LAMPs so as to get an improved model that can return zero mixing parameters to data points where the OLS residuals are zero without forcing the C_o to zero. This will ensure that a relatively greater number of data points are considered in the choice of N_o, Hence, greater number of data points are assigned optimal or nearer optimal mixing parameters.

3. Reformation of the Model for Selecting Variable Mixing Parameters

We present a methodology for the modification of the existing model for LAMPs in order to motivate its adaptation to situations where one or more of the data points with zero OLS residuals give zero OLS mixing parameters.

First, we factor out e_i by writing (10) as:

$$\lambda_i(e_i) = e_i \left\{ \frac{\left(\frac{N_0 C_0 T_0}{e_i}\right) + N_o}{T_0(C_0 n+1)} \right\}, \ i = 1,2,\dots,n, \tag{11}$$

$$\lambda_i(e_i) = e_i \left\{ \frac{\left(\frac{N_0 C_0 T_0 + N_0 e_i}{e_i}\right)}{T_0(C_0 n+1)} \right\}, \ i = 1,2,\dots,n, \tag{12}$$

$$\lambda_i(e_i) = e_i \left\{ \frac{N_0 C_0 T_0 + N_0 e_i}{T_0 e_i (C_0 n+1)} \right\}, \ i = 1,2,\dots,n, \tag{13}$$

Next, we need to find a way to deal with the denominator in (13) so as to avoid mathematical error at data points where OLS residual is zero.

T_o, n and e_i are all defined as positive real numbers. Therefore, for positive values of N_o and C_o, (13) is an increasing function of e_i.

It is easy to see that if $f(x), x \geq 0$ is a strictly increasing function, so is the function $g(x)$ that comprises the reciprocal of $f(x)$ and given as $g(x) = 1 - p^{\left\{x\left(\frac{1}{f(x)}\right)\right\}}$, provided $f(x) > 0$ and $0 < p < 1$.

Therefore, to ensure that the data points with zero residual, $e_i = 0$, return zero mixing parameters, $\lambda_i(e_i) = 0, i = 1,2,\dots,n,$ (13) is reformulated as:

$$\lambda_i(e_i) = \left[1 - p^{e_i\left\{1/\left(\frac{N_0 C_0 T_0 + N_0 e_i}{T_0 e_i (C_0 n+1)}\right)\right\}} \right], \quad i = 1,2,\dots,n, \ 0 < p < 1, \tag{14}$$

$$\lambda_i(e_i) = \left[1 - p^{\left\{e_i\left(\frac{T_0 e_i (C_0 n+1)}{N_0 C_0 T_0 + N_0 e_i}\right)\right\}} \right], \quad i = 1,2,\dots,n, \ 0 < p < 1, \tag{15}$$

$$\lambda_i(e_i) = \left[1 - p^{\left\{e_i^2\left(\frac{T_o (C_0 n+1)}{N_o C_0 T_0 + N_o e_i}\right)\right\}} \right], \ i = 1,2,\dots,n, \ 0 < p < 1, \tag{16}$$

$$\lambda_i(e_i) = \left[1 - p^{\left\{\frac{T_o e_i^2}{N_o}\left(\frac{(C_0 n+1)}{C_0 T_0 + e_i}\right)\right\}} \right], \ i = 1,2,\dots,n, \ 0 < p < 1, \tag{17}$$

where (17) is the proposed reformed model for generating LAMPs.

It is clearly seen that the proposed model (17) can assign a zero mixing parameter to data points with zero OLS residuals without complications because $e = 0$ automatically assigns a zero mixing parameter(s) only to the data point(s) with a zero OLS residual(s).

4. Application to Real Data

This section presents the analysis of two multi-response problems from the literature. For easy reference, results obtained using fixed mixing parameter, LAMPs from the existing model (10), and LAMPs from the proposed model (17), are designated MRR2(F), MRR2(E), and MRR2(P), respectively. Tables that indicate the goodness-of-fit such as SSE, MSE, the Adjusted Coefficient of Determination (R^2_{Adj}), PRESS** as well as the optimization results based on the desirability function are presented for each problem.

Furthermore, in order to determine the LAMPs that give a better mimicry of the absolute values of OLS residuals, we introduce a measure which we refer to as the sum of squared difference (SSD) between:

(i) the coefficient of range of the absolute OLS residuals, CR($\boldsymbol{y_q(e)}$), where $\boldsymbol{y_q(e)}$ is vector of the absolute values of the OLS residuals from the q^{th} response variable, $q = 1,2,\dots,m$, and the coefficient of range, CR($\boldsymbol{\lambda_q(e)}$) of the LAMPs from MRR2(E) and MRR2(P), where $\boldsymbol{\lambda_q(e)}$, $q = 1,2,\dots,m$, are the LAMPs for the q^{th} response variable generated from (10) and (17), respectively;

(ii) the coefficient of variation of the absolute OLS residuals, CV($\boldsymbol{y_q(e)}$), $q = 1,2,\dots,m$, and the coefficient of variation, CV($\boldsymbol{\lambda_q(e)}$), $q = 1,2,\dots,m$, of the LAMPs from MRR2(E) and MRR2(P).

The necessary calculations are as follows:

(a) $CR\big(\boldsymbol{y_q(e)}\big) = \big[\big(\max\left(\boldsymbol{y_q(e)}\right) - \min\left(\boldsymbol{y_q(e)}\right)\big)/\big(\max\big(\boldsymbol{y_q(e)}\big) + \min\left(\boldsymbol{y_q(e)}\right)\big)\big]$, (18)

(b) $CV\big(\boldsymbol{y_q(e)}\big) = \text{standard deviation}(\boldsymbol{y_q(e)})/\text{mean}(\boldsymbol{y_q(e)})$, (19)

(c) $CR\big(\boldsymbol{\lambda_q(e)}\big) = \big[\big(\max\big(\boldsymbol{\lambda_q(e)}\big) - \min\big(\boldsymbol{\lambda_q(e)}\big)\big)/\big(\max\big(\boldsymbol{\lambda_q(e)}\big) + \min\left(\boldsymbol{\lambda_q(e)}\right)\big]$, (20)

(d) $CV\left(\left(\boldsymbol{\lambda_q(e)}\right)\right) = \text{standard deviation}(\left(\boldsymbol{\lambda_q(e)}\right))/\text{mean}(\left(\boldsymbol{\lambda_q(e)}\right))$, (21)

Finally, the SSD for MRR2(E) or MRR2(P) is obtained by:

$$SSD(CR) = \sum_{q=1}^{m}\left\{CR\left(\boldsymbol{y_q(e)}\right) - CR\left(\boldsymbol{\lambda_q(e)}\right)\right\}^2 \quad (22)$$

$$SSD(CV) = \sum_{q=1}^{m}\left\{CV\left(\boldsymbol{y_q(e)}\right) - CV\left(\boldsymbol{\lambda_q(e)}\right)\right\}^2, \quad (23)$$

Both CR$(\boldsymbol{\lambda_q(e)})$ and CV$(\boldsymbol{\lambda_q(e)})$ statistics measure the comparative spread as well as the variability of the mixing parameters in the prescribed interval [0, 1]. The best goodness of fit as well as the best desirability measures are shown in bold.

4.1 **The Minced Fish Quality Data**

This multi-response problem originated from Shah et al. (2004) and has been analyzed in several research papers including Edionwe & Mbegbu, 2014 and Wan & Birch (2011). The study uses a Central Composite Design (CCD) and consists of three explanatory variables, namely x_1(washing temperatures), x_2(washing time) and x_3(washing ratio of water volume to sample weight), respectively. The four response variables are y_1, y_2, y_3, y_4, representing springiness, thiobarbituric acid number, (TBA), cooking loss, and whiteness index, respectively. The research problem is the determination of the setting of the three explanatory variables that will simultaneously optimize the four response variables according to the production requirement of each given below:

Maximize y_1 with lower bound L=1.700, and target value $\emptyset$ =1.9200.

Minimize $y_2$with target value $\emptyset = 20.1600$, and upper bound U=21.0000.

Minimize y_3 with target value $\emptyset$ =16.8000, and upper bound U =20.000.

Maximize y_4 with lower bound L =45.0000, and target value $\emptyset$ =50.9800.

The observed data are presented in Table 1.

Table 1: The minced fish quality data

i	x_1	x_2	x_3	y_1	y_2	y_3	y_4
1	0.2030	0.2030	0.2030	1.83	29.31	29.50	50.36
2	0.7970	0.2030	0.2030	1.73	39.32	19.40	48.16
3	0.2030	0.7970	0.2030	1.85	25.16	25.70	50.72
4	0.7970	0.7970	0.2030	1.67	40.18	27.10	49.69
5	0.2030	0.2030	0.7970	1.86	29.82	21.40	50.09
6	0.7970	0.2030	0.7970	1.77	32.20	24.00	50.61
7	0.2030	0.7970	0.7970	1.88	22.01	19.60	50.36
8	0.7970	0.7970	0.7970	1.66	40.02	25.10	50.42
9	0.0000	0.5000	0.5000	1.81	33.00	24.20	29.31
10	1.0000	0.5000	0.5000	1.37	51.59	30.60	50.67
11	0.5000	0.0000	0.5000	1.85	20.35	20.90	48.75
12	0.5000	1.0000	0.5000	1.92	20.53	18.90	52.70
13	0.5000	0.5000	0.0000	1.88	23.85	23.00	50.19
14	0.5000	0.5000	1.0000	1.90	20.16	21.20	50.86

15	0.5000	0.5000	0.5000	1.89	21.72	18.50	50.84
16	0.5000	0.5000	0.5000	1.88	21.21	18.60	50.93
17	0.5000	0.5000	0.5000	1.87	21.55	16.80	50.98

In the data modelling phase, the polynomials specified for OLS model for responses $y_1$and y_4 include the intercept, x_1 and x_1^2, and the ones specified model for response variables y_2 and y_3 includes the intercept, x_1, x_2, x_1^2, x_1x_2, and intercept, x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_3^2, respectively.

The optimal values of the tuning parameters, N^*, C^* for the locally adaptive bandwidths in (5) for MRR2(F), MRR2(E) and MRR2(P), N_o^*, C_o^* and N_o^*, C_o^*, p^* for generating LAMPs from the existing model in (10) (MRR2(E)) and those from the proposed model in (17) (MRR2(P)), respectively, are presented in Table 2, the corresponding locally adaptive bandwidths are given in Table 3. The absolute values of OLS residuals and the LAMPs for MRR2(E) and MRR2(P) are all shown in Table 4 and the respective values of the CR$(\boldsymbol{y_q(e)})$ and CV$(\boldsymbol{\lambda_q(e)})$ presented in Table 5. The model goodness of fit for MRR2(F), MRR2(E), MRR2(P) and the fixed mixing parameter λ for MRR2(F) are shown in Table 6 while Table 7 presents the respective optimal solutions of the problem based on the desirability measure in (9).

Table 2: Optimal values of the tuning parameters for the minced fish data

Response	Optimal tuning parameters of bandwidths for MRR2(F), MRR2(E) & MRR2(P)			Optimal tuning parameters of LAMPs for MRR2(E)		Optimal tuning parameters of LAMPs for MRR2(P)		
	$\boldsymbol{b^*}$	$\boldsymbol{N^*}$	$\boldsymbol{C^*}$	$\boldsymbol{N_o^*}$	$\boldsymbol{C_o^*}$	$\boldsymbol{N_o^*}$	$\boldsymbol{C_o^*}$	$\boldsymbol{P^*}$
$\boldsymbol{y_1}$	0.1665	8.1582	1.4176	12.5480	0.2698	0.0089	116.9999	0.5206
$\boldsymbol{y_2}$	0.2567	5.8642	0.1094	11.1157	0.1409	0.9045	399.9999	0.9863
$\boldsymbol{y_3}$	0.3646	10.6022	0.0796	10.2442	0.0284	0.3668	203.9999	0.9522
$\boldsymbol{y_4}$	0.1218	11.0554	2.4078	16.4373	3.9290	0.9374	106.9999	0.9532

Table 3: Optimal bandwidths for the minced fish data

i	y_1	y_2	y_3	y_4
1	0.0798	0.0872	0.0315	0.0792
2	0.0800	0.0516	0.3190	0.0792
3	0.0798	0.1020	0.1396	0.0791
4	0.0802	0.0485	0.0998	0.0792
5	0.0798	0.0854	0.2620	0.0792
6	0.0800	0.0769	0.1880	0.0791
7	0.0798	0.1132	0.3133	0.0792
8	0.0802	0.0491	0.1567	0.0792
9	0.0799	0.0741	0.1823	0.0800
10	0.0807	0.0079	0.0002	0.0791
11	0.0798	0.1191	0.2763	0.0792
12	0.0797	0.1185	0.3332	0.0791
13	0.0798	0.1067	0.2165	0.0792
14	0.0797	0.1198	0.2677	0.0791
15	0.0797	0.1142	0.3446	0.0791
16	0.0798	0.1161	0.3417	0.0791
17	0.0798	0.1148	0.3930	0.0791

Table 4: Absolute values of the OLS residuals and LAMPs for the minced fish data

i	**OLS residuals**				**LAMPs for MRR2(E) ($\lambda_q(e)$)**				**LAMPs for MRR2(P) ($\lambda_q(e)$)**			
	$y_1(e)$	$y_2(e)$	$y_3(e)$	$y_4(e)$	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
1	0.0482	1.6531	1.7524	4.3129	0.8255	0.6203	0.7370	0.9759	0.9448	0.5076	0.9991	1.0000
2	0.0465	5.5977	2.0363	2.7610	0.8178	1.0000	0.8246	0.9675	0.9326	0.9997	0.9999	0.9987
3	0.0282	3.0987	1.5768	4.6729	0.7344	0.7595	0.6828	0.9778	0.6291	0.9170	0.9965	1.0000
4	0.0135	1.7333	2.0881	1.2310	0.6674	0.6280	0.8406	0.9593	0.2029	0.5411	0.9999	0.7322
5	0.0182	2.1631	0.0068	4.0429	0.6889	0.6694	0.1983	0.9744	0.3385	0.7027	0.0001	1.0000
6	0.0865	1.5223	0.5044	0.3110	1.0000	0.6077	0.3519	0.9543	0.9999	0.4517	0.4388	0.0807
7	0.0018	0.0513	1.8176	4.3129	0.6142	0.4661	0.7571	0.9759	0.0040	0.0007	0.9994	1.0000
8	0.0235	1.5733	1.9712	0.5010	0.7130	0.6126	0.8045	0.9553	0.4973	0.4736	0.9999	0.1961
9	0.0555	2.3592	2.6040	8.7892	0.8589	0.6883	0.9998	1.0000	0.9786	0.7638	1.0000	1.0000
10	0.0567	2.6671	1.2897	4.3655	0.8641	0.7179	0.5942	0.9762	0.9817	0.8419	0.9771	1.0000
11	0.0346	2.6723	1.9202	3.1591	0.7635	0.7184	0.7887	0.9697	0.7749	0.8430	0.9998	0.9998
12	0.0354	1.7589	0.0387	0.7909	0.7673	0.6305	0.2082	0.9569	0.7907	0.5516	0.0034	0.4195
13	0.0046	1.1944	1.5591	1.7191	0.6269	0.5762	0.6773	0.9619	0.0259	0.3092	0.9960	0.9234
14	0.0154	2.4956	0.2448	1.0491	0.6762	0.7014	0.2718	0.9583	0.2564	0.8010	0.1272	0.6160
15	0.0054	0.9356	0.4593	1.0691	0.6307	0.5513	0.3379	0.9584	0.0359	0.2031	0.3805	0.6299
16	0.0046	1.4456	0.3593	0.9791	0.6269	0.6003	0.3071	0.9579	0.0259	0.4183	0.2540	0.5655
17	0.0146	1.1056	2.1593	0.9291	0.6724	0.5676	0.8625	0.9576	0.2330	0.2716	1.0000	0.5279

Table 5: CR and CV for OLS residuals and LAMPs for the minced fish data

	OLS residuals				LAMPs for MRR2(E)				LAMPs for MRR2(P)			
	$\boldsymbol{y_1(e)}$	$\boldsymbol{y_2(e)}$	$\boldsymbol{y_3(e)}$	$\boldsymbol{y_4(e)}$	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
CR(.)	0.9593	0.9818	0.9948	0.9316	0.2390	0.3640	0.6689	0.0234	0.9920	0.9986	0.9998	0.8507
CV(.)	0.8081	0.5963	0.6479	0.8433	0.1447	0.1756	0.4370	0.0124	0.7653	0.4909	0.5672	0.4114

Using (22) and (23), we have the following results:

SSD(CR) for:

MRR2(E) = $(0.9593-0.2390)^2+(0.9818-0.3640)^2+(0.9948-0.6689)^2+(0.9316-0.0234)^2$ = 1.8315.

For the MRR2(p):

SSD(CR) = $(0.9593-0.9920)^2+(0.9818-0.9986)^2+(0.9948-0.9998)^2+(0.9316-0.8507)^2$ = 0.0079.

Similarly, the SSD(CV) for MRR2(E) and MRR2(P) are respectively 1.3520 and 0.2060.

Therefore, the relatively smaller SSDs for MRR2(P) signifies that the LAMPs from the proposed model mimics the OLS residuals at each data point more precisely than does that of the MRR2(E).

Table 6: Goodness of fit for the minced fish data

Response	**Model**	$\boldsymbol{\lambda}$	**DF**	**PRESS****	**SSE**	**MSE**	$\boldsymbol{R^2_{Adj}}$
$\boldsymbol{y_1}$	MRR2(F)	1.0000	12.0000	0.0021	0.0123	0.0010	94.3888
	MRR2(E)	-	12.4165	0.0020	0.0115	0.0009	94.9611
	MRR2(P)	-	12.5818	**0.0017**	**0.0100**	**0.0008**	**95.6468**
$\boldsymbol{y_2}$	MRR2(F)	0.7136	9.1456	11.8856	42.0717	4.6002	94.6440
	MRR2(E)	-	9.3180	10.4183	36.1608	3.8808	95.4816
	MRR2(P)	-	9.4228	**9.4565**	**32.8665**	**3.4880**	**95.9389**
$\boldsymbol{y_3}$	MRR2(F)	0.8644	3.0879	7.5470	2.9484	0.9548	94.0802
	MRR2(E)	-	4.8275	6.0441	3.9734	0.8231	94.8969
	MRR2(P)	-	4.0211	**5.2178**	**1.7729**	**0.4409**	**97.2664**
$\boldsymbol{y_4}$	MRR2(F)	1.0000	12.0000	17.4484	12.1387	1.0116	96.2653
	MRR2(E)	-	12.0495	17.4510	12.0963	1.0039	96.2937
	MRR2(P)	-	12.3095	**16.9404**	**9.3645**	**0.7608**	**97.1913**

The results in Table 6 show that MRR2(P) performs better across the four responses in terms of all the goodness-of-fit, implying that the MRR2(P) did a better job in fitting of the data than the MRR2(F) and MRR2(E). In addition, the CR(λ) and CV(λ) for MRR2(P) are the closest to

CR(e) and CV(e) given in Table 5. This is an indication that the spread and variability inherent in the OLS residuals is captured much better by LAMPs from the proposed model.

Table 7: Optimal solution based on desirability measure for the minced fish data

Model	x_1	x_2	x_3	y_1	y_2	y_3	y_4	d_1	d_2	d_3	d_4	D(%)
MRR2(F)	0.3491	1.0000	0.6506	1.8953	18.9247	18.1057	51.4171	0.8876	1.0	0.5920	1.0	85.1387
MRR2(E)	0.2968	1.0000	0.7566	1.8928	19.1847	17.0397	51.2365	0.8763	1.0	0.9254	1.0	94.8963
MRR2(P)	0.3182	0.9995	0.7482	1.9053	19.0002	16.8001	51.3358	0.9330	1.0	1.0000	1.0	**98.2797**

Results from Table 7 show that MRR2(P) gives a setting of the explanatory variables that corresponds to the best desirability measure of 98.2797%. This translates to meeting approximately 98% of the prescribed product's requirements.

4.2 Tire Thread Data

The data set originated from a study carried out in Derringer & Suich (1980) using a CCD. It has been referenced by several authors, including Castillo (2007) and Edionwe et al. (2017). The study was used in the development of a chemical compound for the manufacture of tire thread. The explanatory variables that were found to affect the quality of the tire thread include hydrated silica level (x_1), silane coupling agent (x_2), and sulphur level (x_3). The four responses representing the different aspects of quality of the manufactured tire thread are PICO Abrasive index (y_1), 200% modulus (y_2), Elongation at break (y_3), and Hardness (y_4). The data is presented in Table 8.

The problem involves a search for the setting of the three explanatory variables that simultaneously optimizes the four responses according to the following production requirements:

Maximize y_1, with L= 120 and target value $\emptyset$ =170.

Maximize y_2, with L =1000 and $\emptyset$ =2000.

Get the value of y_3 such that $400 < y_3 < 600$, with $\emptyset = 500$.

Get the value of y_4 such that $60 < y_4 < 75$, with $\emptyset = 67.5$.

Table 8: The tire thread data

i	x_1	x_2	x_3	y_1	y_2	y_3	y_4
1	0.1933	0.1933	0.1933	102	900	470	67.5
2	0.8067	0.1933	0.1933	120	860	410	65.0
3	0.1933	0.8067	0.1933	117	800	570	77.5
4	0.8067	0.8067	0.1933	198	2294	240	74.5
5	0.1933	0.1933	0.8067	103	490	640	62.5
6	0.8067	0.1933	0.8067	132	1289	270	67.0
7	0.1933	0.8067	0.8067	132	1270	410	78.0
8	0.8067	0.8067	0.8067	139	1090	380	70.0
9	0.0000	0.5000	0.5000	102	770	590	76.0
10	1.0000	0.5000	0.5000	154	1690	260	70.0
11	0.5000	0.0000	0.5000	96	700	520	63.0
12	0.5000	1.0000	0.5000	163	1540	380	75.0
13	0.5000	0.5000	0.0000	116	2184	520	65.0
14	0.5000	0.5000	1.0000	153	1784	290	71.0
15	0.5000	0.5000	0.5000	133	1300	380	70.0
16	0.5000	0.5000	0.5000	133	1300	380	68.5
17	0.5000	0.5000	0.5000	140	1145	430	68.0
17	0.5000	0.5000	0.5000	142	1090	430	68.0
19	0.5000	0.5000	0.5000	145	1260	390	69.0
20	0.5000	0.5000	0.5000	142	1344	390	70.0

For the OLS method, a full second-order polynomial model was specified for fitting y_1, y_2, and y_4 and a first-order polynomial model for y_3.

The respective tables, including the ones for goodness-of-fit and optimal solution based on desirability measure in (9), are shown below.

Table 9: Optimal values of the tuning parameters for the tire thread data

Response	Optimal tuning parameters of bandwidths			Optimal tuning parameters of LAMPs for MRR2(E)		Optimal tuning parameters of LAMPs for MRR2(P)		
	b^*	N^*	C^*	N_o^*	C_o^*	N_o^*	C_o^*	P^*
y_1	0.1878	10.0501	0.0945	13.6724	0.1079	0.8846	99.9999	0.9990
y_2	0.1881	11.8306	0.0000	18.7913	0.9139	23.396	1008	0.9999
y_3	0.1968	19.1479	0.0194	8.1462	0.0000	9.9991	320	0.9999
y_4	0.2872	19.1998	0.0388	15.0260	0.1968	0.6840	898	0.9654

Table 10: Adaptive optimal bandwidths for the tire thread data

i	y_1	y_2	y_3	y_4
1	0.1191	0.0798	0.2271	0.2356
2	0.1048	0.0762	0.1829	0.1915
3	0.1072	0.0709	0.3009	0.4120
4	0.0427	0.2034	0.0575	0.3591
5	0.1183	0.0434	0.3525	0.1474
6	0.0952	0.1143	0.0796	0.2268
7	0.0952	0.1126	0.1829	0.4208
8	0.0897	0.0966	0.1608	0.2797
9	0.1191	0.0683	0.3156	0.3855
10	0.0777	0.1498	0.0723	0.2797
11	0.1239	0.0621	0.2640	0.1562
12	0.0706	0.1365	0.1608	0.3679
13	0.1080	0.1936	0.2640	0.1915
14	0.0785	0.1582	0.0944	0.2973
15	0.0945	0.1153	0.1608	0.2797
16	0.0945	0.1153	0.1608	0.2532
17	0.0889	0.1015	0.1976	0.2444
18	0.0873	0.0966	0.1976	0.2444
19	0.0849	0.1117	0.1681	0.2620
20	0.0873	0.1192	0.1681	0.2797

Table 11: Absolute values of the OLS residuals and LAMPs for the tire thread data

	OLS residuals ($y_q(e)$)				**LAMPs for MRR2(E) ($\lambda_q(e)$)**				**LAMPs for MRR2(P) ($\lambda_q(e)$)**			
i	$y_1(e)$	$y_2(e)$	$y_3(e)$	$y_4(e)$	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
1	18.483	192.1434	106.0082	1.8907	0.9999	0.9363	0.8737	0.8278	0.9996	0.9574	0.8943	0.9748
2	2.3002	461.3275	32.7221	0.4770	0.5334	1.0000	0.2697	0.6568	0.1128	1.0000	0.1928	0.2089
3	6.4221	454.2729	57.2980	0.2307	0.6522	0.9983	0.4722	0.6270	0.6066	1.0000	0.4814	0.0533
4	15.686	158.5143	73.9718	3.3140	0.9191	0.9283	0.6097	1.0000	0.9962	0.8833	0.6653	1.0000
5	15.049	413.1095	119.2958	3.2525	0.9006	0.9886	0.9832	0.9925	0.9940	1.0000	0.9419	1.0000
6	7.0658	199.6778	51.9740	0.2922	0.6707	0.9381	0.4284	0.6344	0.6768	0.9669	0.4174	0.0842
7	2.8967	203.4162	50.6379	0.5299	0.5506	0.9390	0.4173	0.6632	0.1729	0.9709	0.4012	0.2511
8	17.793	443.4223	121.3321	1.8206	0.9799	0.9958	1.0000	0.8193	0.9992	1.0000	0.9473	0.9670
9	0.2893	164.0961	10.6713	0.6690	0.4755	0.9297	0.0880	0.6800	0.0019	0.8999	0.0225	0.3693
10	1.3304	214.0069	4.6527	0.7748	0.5055	0.9415	0.0383	0.6928	0.0393	0.9801	0.0043	0.4611
11	4.7313	175.2869	51.0593	0.9778	0.6035	0.9323	0.4208	0.7174	0.3974	0.9277	0.4064	0.6263
12	3.7625	207.8885	14.2647	1.0704	0.5756	0.9400	0.1176	0.7286	0.2741	0.9751	0.0399	0.6926
13	15.306	225.7862	57.5822	2.7293	0.9082	0.9443	0.4746	0.9292	0.9950	0.9872	0.4848	0.9995
14	14.337	157.3892	82.2582	2.6367	0.8803	0.9281	0.6780	0.9180	0.9904	0.8797	0.7416	0.9992
15	6.0994	39.1544	37.3380	1.0908	0.6429	0.9001	0.3077	0.7310	0.5690	0.1228	0.2434	0.7063

16	6.0994	39.1544	37.3380	0.4092	0.6429	0.9001	0.3077	0.6486	0.5690	0.1228	0.2434	0.1584
17	0.9006	115.8456	12.6620	0.9092	0.4931	0.9182	0.1044	0.7091	0.0182	0.6825	0.0316	0.5731
18	2.9006	170.8456	12.6620	0.9092	0.5507	0.9313	0.1044	0.7091	0.1734	0.9175	0.0316	0.5731
19	5.9006	0.8456	27.3380	0.0908	0.6372	0.8910	0.2253	0.6101	0.5451	0.0001	0.1389	0.0085
20	18.483	192.1434	106.0082	1.8907	0.5507	0.9105	0.2253	0.7310	0.1734	0.4463	0.1389	0.7063

Table 12: CR and CV for OLS residuals and LAMPs for the tire thread data

	OLS residuals				**LAMPs for MRR2(E)**				**LAMPs for MRR2(P)**			
	$\boldsymbol{y_1(e)}$	$\boldsymbol{y_2(e)}$	$\boldsymbol{y_3(e)}$	$\boldsymbol{y_4(e)}$	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
CR(.)	0.9692	0.9963	0.9261	0.9467	0.3554	0.0576	0.9261	0.2422	0.9962	0.9999	0.9909	0.9832
CV(.)	0.8148	0.6649	0.7202	0.7987	0.2580	0.0345	0.7202	0.1681	0.7382	0.4215	0.8586	0.6212

Using (22) and (23), the SSD(CR) for MRR2(E) and MRR2(p) are 1.7542 and 0.0063, respectively. Similarly, the SSD(CV) of MRR2(E) and MRR2(P) are respectively 1.1051 and 0.1158.

Clearly, the SSD(CR) and SSD(CV) for MRR2(P) are both relatively smaller than those for MRR2(E). This is an indication that the LAMPs from the proposed model simulate the OLS residuals more precisely than the LAMPs for MRR2(E).

Table 13: Goodness-of-fit for the tire thread data

Response	**Model**	$\boldsymbol{\lambda}$	**DF**	**PRESS****	**SSE**	**MSE**	$\boldsymbol{R^2_{Adj}}$
$\boldsymbol{y_1}$	MRR2(F)	0.7929	6.0355	357.8081	200.3606	33.1970	94.4003
	MRR2(E)	-	6.4085	261.3186	**155.5986**	24.2800	95.9044
	MRR2(P)	-	7.1247	**214.9636**	166.0743	**23.3098**	**96.0681**
$\boldsymbol{y_2}$	MRR2(F)	1.0000	5.0000	318830	49941	9988.2	95.4711
	MRR2(E)	-	5.2444	318220	51218	9766.3	95.5717
	MRR2(P)	-	5.2181	**316410**	**46654**	**8940.7**	**95.9460**
$\boldsymbol{y_3}$	MRR2(F)	0.6755	8.6010	4107.4	**10247**	1191.4	90.0792
	MRR2(E)	-	10.8452	**3146.5**	10656	982.5061	91.8187
	MRR2(P)	-	11.0662	3211.9	10661	**963.3883**	**91.9779**
$\boldsymbol{y_4}$	MRR2(F)	1.0000	5.4459	14.2068	**4.5509**	0.8357	95.8570
	MRR2(E)	-	6.4587	13.9013	5.4897	0.8500	95.7861
	MRR2(P)	-	7.1545	**13.2480**	5.6235	**0.7860**	**96.1031**

Results in Table 13 show that out of a total of sixteen cells (that is four goodness-of-fit across four responses), MRR2(P) provides the best results in twelve including the best PRESS** in y_1, y_2 and y_4. Furthermore, SSD(CR) and SSD(CV) of LAMPs for MRR2(P) are respective smaller than those for MRR2(E), implying that the spread and variability inherent in the OLS residuals is replicated much better by LAMPs from MRR2(P) than those of MRR2(E).

Table 14: Optimization results based on desirability function for the tire thread data

Model	x_1	x_2	x_3	y_1	y_2	y_3	y_4	d_1	d_2	d_3	d_4	D(%)
MRR2(F)	0.5981	0.6032	0.0	132.3	2439.6	471.9	67.7	0.3916	1.0	0.6233	0.8303	64.3261
MRR2(E)	0.5969	0.5945	0.0	138.8	2431.0	498.3	67.5	0.3755	1.0	0.9825	0.9999	77.9313
MRR2(P)	0.5798	0.5966	0.0	148.0	2405.1	499.8	67.5	0.5599	1.0	0.9981	1.0000	86.4613

Results presented in Table 14 show that MRR2(P) gives a setting of the explanatory variables that corresponds to the best desirability measure of 86.4613%, implying an approximately 87% compliance with the prescribed product's quality requirements.

5. Discussion and Conclusion

In the application of MRR2, the mixing of the estimates from both OLS and LLR using LAMPs is more efficient than the mixing done using the fixed counterpart for two reasons: One, the relative performance of the OLS at each of the data points is considered in the choice of the local mixing parameter assigned to that each data point. Hence for optimal mixing, a relatively larger value of mixing parameter is assigned to data point where the absolute value of the OLS estimate is large, and vice versa. Two, in the determination of the setting of the explanatory variables that optimizes the response with respect to its requirements, every data point is a viable alternative, unlike what obtains when a fixed mixing parameter is used (Tables 4 and 11).

In this paper, we modified the existing model for selecting LAMPs as far as both MRR2 and response surface studies are concerned. The LAMPs generated from the proposed model are found to exhibit higher spread (range) and variability in the prescribed interval of [0, 1] and therefore, more sensitive to the individual size of the OLS residual at each data point. Furthermore, the proposed model is designed to be able to assign a zero mixing parameter to a data point with zero OLS residual if ever the need arises.

The superiority of the LAMPs from the proposed model was visibly displayed in the comparisons of the goodness-of-fit (Tables 6 and 13) as well as the optimal solutions based on the desirability function of the multi-response problems (Tables 7 and 14).

References

Anderson-Cook, C. M., & Prewitt, K. (2005). Some guidelines for using nonparametric models for modeling data from response surface designs. *Journal of Modern Applied Statistical Models, 4*, 106–119.

Box, G. E. P., & Wilson, B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Methodological), 13*, 1–45.

Castillo, D. E. (2007). *Process optimization: A statistical method*. Springer International Series in Operations Research and Management Science.

Derringer, G., & Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology, 12*(4), 214–219.

Edionwe, E., & Eguasa, O. (2023). Local quadratic regression: Maximizing performance via a modified PRESS for bandwidth selection. *The Philippine Statistician, 72*(1), 51–69.

Edionwe, E., & Mbegbu, J. I. (2014). Local bandwidths for improving the performance statistics of model-robust regression methods. *Journal of Modern Applied Statistical Methods, 13*(2), 508–527.

Edionwe, E., Mbegbu, J. I., & Chinwe, R. (2016). A new function for generating local bandwidths for semi-parametric MRR2 model in response surface methodology. *Journal of Quality Technology, 48*(4), 388–404.

Edionwe, E., Mbegbu, J. I., Ekhosuehi, N., & Obiora-Ilouno, H. O. (2018). An improved robust regression model for response surface methodology. *Croatian Operational Research Review, 9*(2), 317–330.

Edionwe, E., Mbegbu, J. I., & Iguodala, W. A. (2017). Improving the performance of model-robust regression 2 (MRR2) method using new adaptive mixing parameters and a modified penalized error sum of squares. *Journal of the Nigerian Association of Mathematical Physics, 41*, 229–240.

Harrington, E. C. (1965). The desirability function. *Industrial Quality Control, 21*(10), 494–498.

He, Z., Zhu, P. F., & Park, S. H. (2012). A robust desirability function for multi-response surface optimization. *European Journal of Operational Research, 221*, 241–247.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press.

Mays, J. E., & Birch, J. B. (2002). Smoothing for small samples with model misspecification: Nonparametric and semi-parametric concerns. *Journal of Applied Statistics, 29*(7), 1023–1045.

Mays, J. E., Birch, J. B., & Starnes, B. A. (2001). Model robust regression: Combining parametric, nonparametric, and semi-parametric models. *Journal of Nonparametric Statistics, 13*, 245–277.

Montgomery, D. C. (2005). *Design and analysis of experiments* (6th ed.). Wiley.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2009). *Response surface methodology: Process and product optimization using designed experiments* (3rd ed.). Wiley.

Pickle, S. M., Robinson, T. J., Birch, J. B., & Anderson-Cook, C. M. (2008). A semi-parametric model to robust parameter design. *Journal of Statistical Planning and Inference, 138*, 114–131.

Shah, K. H., Montgomery, D. C., & Carlyle, W. M. (2004). Response surface modelling and optimization in multi-response experiments using seemingly unrelated regressions. *Quality Engineering, 16*, 387–397.

Wan, W., & Birch, J. B. (2011). A semi-parametric technique for multi-response optimization. *Quality and Reliability Engineering International, 27*, 47–59.

Yeniay, O. (2014). Comparative study of algorithms for response surface optimization. *Journal of Mathematical and Computational Applications, 19*(1), 93–104.

An Empirical Investigation of the Effect of Cashless Policy on Inflation in Nigeria

[1]Ebuh G.U, [1]Nwosu A.N, [1]Sanusi Y.M and [1*]Bello D.

[1]Central Bank of Nigeria, Abuja

**Corresponding author: DBELLO2@cbn.gov.ng*

Abstract

This paper evaluates the effect of cashless policy on inflation in Nigeria. Monthly time series data from 2010M1 – 2022M12 on inflation rate, Automated Teller Machine, Point of Sales, Web Pay, Cheque, and Mobile Banking Coverage were used. The study leveraged the Autoregressive Distributed Lag (ARDL) and the Error Correction Models to examine the short and long run effects of cashless policy on inflation in Nigeria. Preliminary and post estimation tests were carried out to establish the attributes and stability of the models. Findings revealed that models were stable and automated teller machine and cheque have negative effects while, mobile transaction, web pay and point of sale have positive effects on inflation in Nigeria. The result further indicates evidence of long run relationship among the variables, which corroborates the findings of Titalessy (2020) and Gbawae and Tonye (2023).

Keywords: *Cashless Policy, Inflation Rate, Web Pay, Automated Teller Machine (ATM), Nigeria*

1.0 Introduction

The core mandate of central banks is macroeconomic management aimed at achieving stable inflation rates. Hence, most central banks, including the Central Bank of Nigeria anchor their monetary policy towards low and stable inflation (Goncalves and Salles, 2008; Lin and Ye, 2009; Walsh, 2009; Cuestas and Harrison, 2010). One of the possible ways of evaluating the efficiency of central banks in managing inflation rate is by analysing the impact of cashless policy on inflation.

Cashless policy is an economic policy aimed at reducing the amount of currency notes and coins circulating in the economy. It involves more electronic-based payments such as direct debits, credit transfers, cheques, and payment cards etc. Similarly, a cashless economy is an environment in which money is spent without being physically carried from one person to the

other (Adu, 2016). It is an economic system in which transactions are minimally executed in exchange for actual cash (Daniel, Swartz and Fermar 2004).

The Central Bank of Nigeria (CBN) introduced cashless policy in 2012 to curb excesses in handling of cash in the economy. It prescribed a cash handling charges on daily withdrawal above five hundred thousand Naira (N500,000.00) for individuals and three million Naira (N3,000,000.00) for corporate bodies (CBN, 2015).

Besides reducing the volume of cash-based transactions, the policy was geared towards ensuring the spread of banking services to everyone, create a platform that will ultimately empower all, and influence how businesses are transacted.

Titalessy (2020) opined that non-cash payment instruments are increasing and being preferred because of its convenience, effectiveness, and efficiency. Zandi e t al. (2016) stated that digital payments are convenient and plays an important role in stimulating economic growth in a country. The availability of cashless payment instruments is attributed to innovations in the banking sector, driven by the public's need for practical payment instruments that can provide convenience in conducting transactions. This was buttressed by the outbreak of Covid-19 disease transmitted through droplets, that necessitated various governments to quicken the promotion of cashless payments, to prevent the virus from spreading through cash transactions.

Inflation on the other hand, is a measure of how fast prices of goods and services are rising. Inflation could be a concern because it makes money saved today less valuable tomorrow. It erodes a consumer's purchasing power and can even interfere with the ability to retire. When the general price level rises, a unit of currency buys fewer goods and services. Thus, inflation is associated to a reduction in the purchasing power of money.

However, the study of the effect of cashless policy on inflation is still budding in economic literature particularly in developing economies, and where they exist, they fail to consider the impact of web pay as proxy for cashless policy. Besides, few empirical studies that exist in Nigeria in addition to not including web pay, used quarterly data in their analysis which could account for loss of information and divergence in research findings; hence, constituting a research gap in Nigeria.

Therefore, in contributing to the body of knowledge and extending the literature on the subject matter, this study attempts to empirically investigate how the cashless policy had impacted on

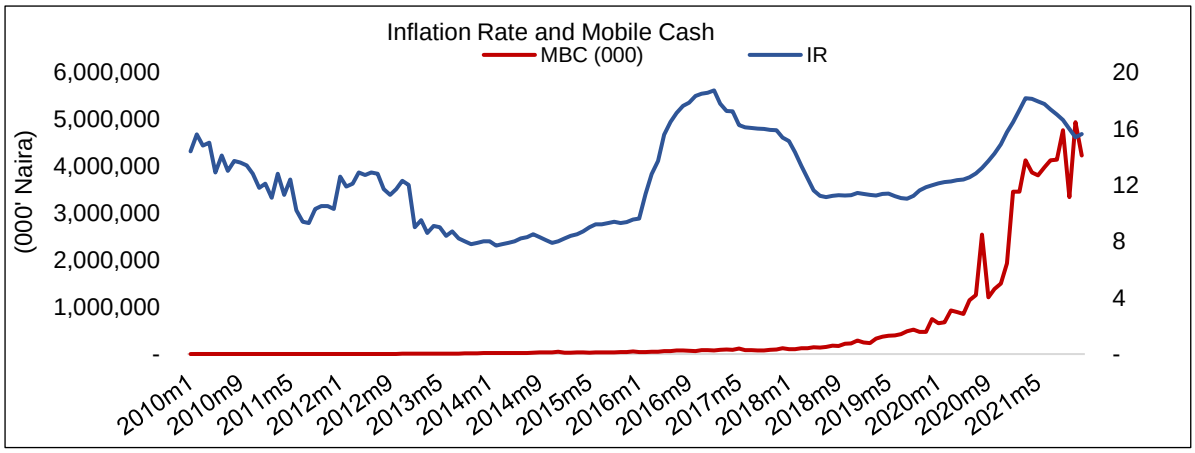
Inflation in Nigeria by incorporating web pay as a proxy for cashless policy using higher frequency data.

Following the introduction, the structure for the study ensues: Section 2 provides the trends between inflation rate and payment channels while section three follows with the theoretical and empirical literature review. Section 4 details the research methodology while section 5 presents and discusses the descriptive and empirical results. Section 6 concludes the study with policy commendations and area for further research.

2.0 Trends of Relationship between Payment Channels and Inflation Rate in Nigeria

From the graphs below (Fig.1 - Fig. 5) inflation has been increasing due to structural and non-structural factors such as the global oil price shocks, recession in Nigeria, exchange rate volatility, and Covid-19 pandemic-related issues. The adoption and implementation of the cashless policy in 2012 as seen in the graph below brought some structural changes in the payments system. The volumes and values of usage of the cashless policy channels increased tremendously.

Fig. 1 Graph showing relationship between Inflation Rate and Mobile Cash



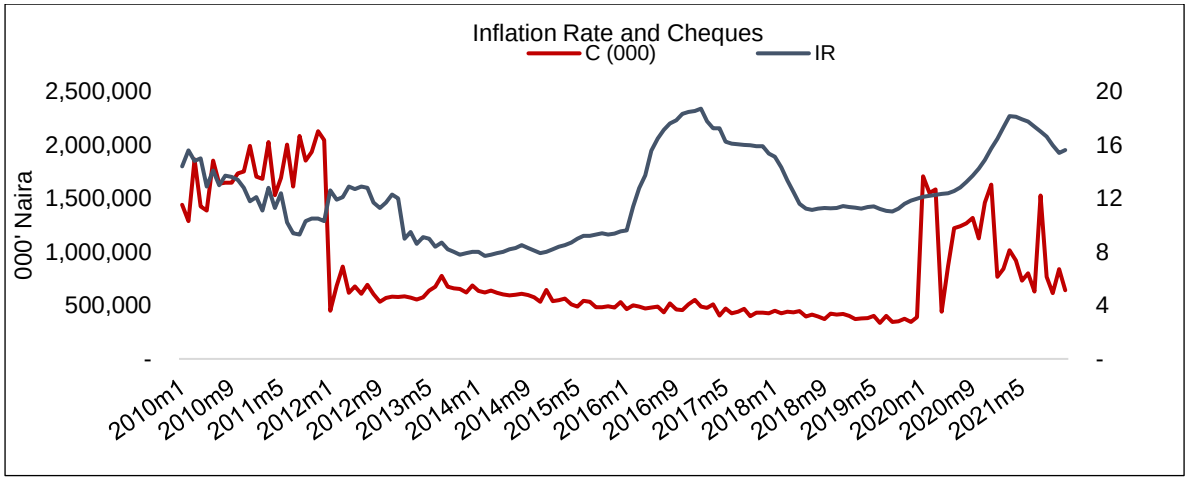


Fig. 2 Graph showing relationship between Inflation Rate and Cheques

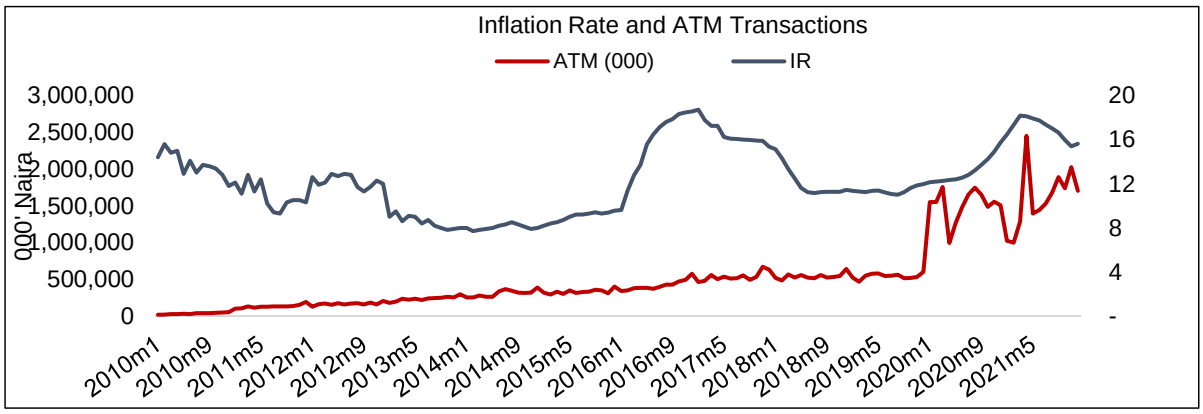


Fig. 3 Graph showing relationship between Inflation Rate and ATM Transactions

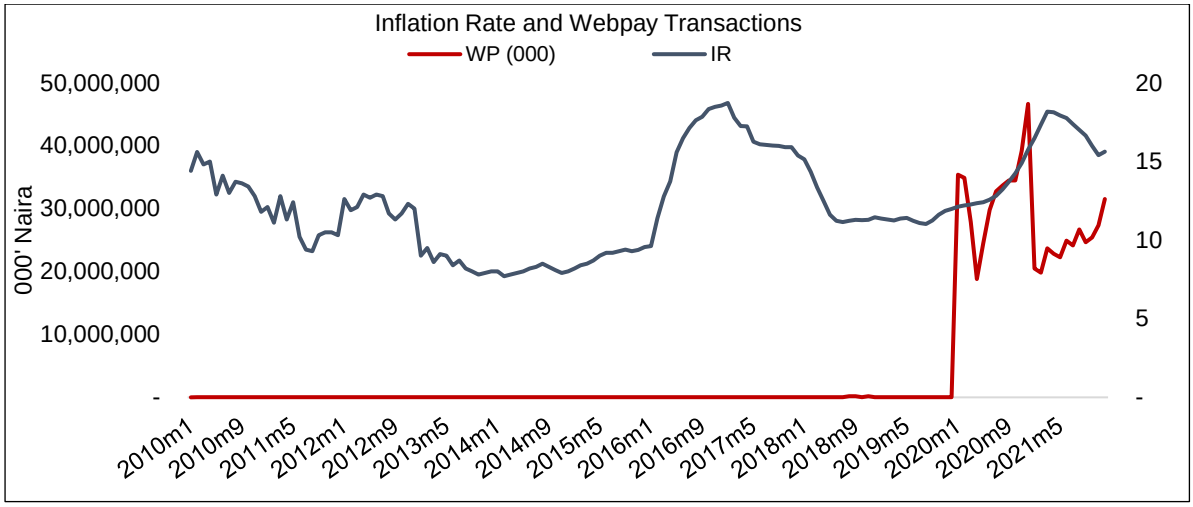


Fig. 4 Graph showing relationship between Inflation Rate and Web Pay

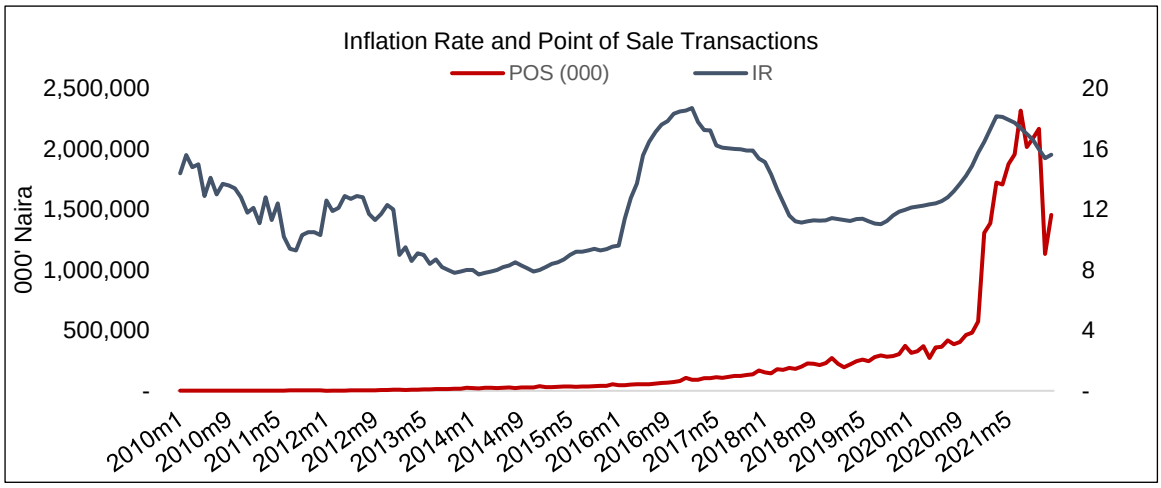


Fig. 5 Graph showing relationship between Inflation Rate and POS Transactions

From Fig. 1, it could be seen that despite the adoption of the cashless policy in 2012, the use of MBC as a channel for payments remained low till 2018. However, an appreciable increase was noticed from 2019 to date. This could be attributed to COVID-19 pandemic. The implication of the increased adoption of this channel shows a positive relationship with inflation such that inflation responded to a dip in the MBC.

Fig. 2 shows that prior to the adoption of the cashless policy, the use of cheques has been very high. It could be seen that from 2011/12, the use of cheques reduced noticeably to the use of alternative medium of payments such as Mobile transfers, use of POS etc. An inverse relationship between the trends of inflation and cheques was observed. Between 2019 and date, it could be seen that the use of cheques also rose, aligning with inflation trends.

Fig. 3 revealed that between 2010 and 2016, inflation was majorly downward slopping while ATM transactions were upward trending between 2010 and 2019. From 2019 to date, ATM transactions fluctuated with inflation.

Fig. 4 shows that despite the adoption of cashless channels for payments, Web Pay (WP) was the least utilised, and static from 2010 to 2019. Within the same period, Inflation was downward slopping although it spiked in 2016. The trend in Web Pay spiked drastically in 2020, which could be attributed to the COVID-19 pandemic. In 2021, the use of Web Pay reduced, which could be associated with the use of alternative channels.

it was observed from Fig. 5 that POS transactions fairly rose. From 2016, the rise steepened and peaked in 2021 due to COVID-19 dynamics. Inflation was majorly downward slopping between 2010 to 2015 but rose and peaked in 2016. From 2020 - date, inflation and POS transactions trended in same direction.

Overall, between 2010 and 2019 from the trend analysis, cashless policy had little or no impact on inflation. However, in 2020 when the global shock that led to the closure of economic activities due to Covid-19 pandemic occurred, a significant impact of cashless policy on inflation was observed. Many Nigerians embraced the cashless mode of payments due to the general restrictions in inter-personal activities. While the use of web pay indicated an inverse relationship with inflation from 2021-2022, the use of cheques depicted a positive relationship.

3.0 Literature Review

3.1 Theoretical Review

The Bank Focused Theory as popularised by Kapoor in 2010 describes how banks choose how to distribute their resources. The theory holds that banks offer their various clients services through non-traditional yet conventional low-cost methods Such methods could be referred to as branchless banking which is anchored on cashless policies use by Banks to provides a wide range of services to its clients, regardless of their location or branch. This theory therefore, offers useful insights into this study by exploring how adjustments to monetary policy decisions and the introduction of cashless system can affect banks' lending behaviour which in turn, have an impact on Nigeria's inflation.

The Technology Acceptance Model (TAM) theory originally proposed by Davis in 1989 is used to forecast the acceptance and adoption of a new technology. The model aims to shed

light on the variables that affect users' intentions to make use of a certain technology and their actual usage behavior. Applying the Technology Acceptance Model to this study would results to a better understanding of how the acceptance and adoption of cashless transactions influence the rate of inflation in Nigeria.

Furthermore, the Quantity Theory of Money which offers a fundamental understanding of the relationship between money supply and inflation could be used to analyze how the adoption of cashless transactions might influence inflation dynamics. As more transactions become digital due to cashless policy, the amount of physical currency in circulation might decrease at the same time increasing the amount of electronic money and digital transactions. Likewise faster circulation of money due to cashless transactions might influence inflationary pressures thereby affecting the overall demand for goods and services.

3.2 Empirical Review

In recent years, the adoption of cashless economies has accelerated globally, with countries embracing more electronic-based payments to boost productivity, openness, and financial inclusion. With the introduction of the cashless policy in Nigeria in 2012, empirical studies both by central banks and academic researchers have examined the potential impact of such policies on key macroeconomic variables, yet Nigeria is just one of the emerging economies with a dearth of empirical research on cashless policy (Olukotun, Golagade and Abubakar, 2023). Owoeye(2023) employed a systematic literature review (SLR) on a sample of 100 research articles to examine the cashless policy and Naira redesign policy impact on the Nigerian macroeconomy. The findings indicated that Nigeria's cashless policy overall has improved the economy's financial performance even though the effect of the policy, as assessed by private sector credit was found to have a negative impact on financial performance. The study suggested that infrastructural and legal framework be enhanced for an effective implementation of the policy, this is in line with the findings of Ewa and Inah (2016). Likewise, Yomere, (2015) in investigating benefits and challenges of Nigeria’s cash-less policy concluded that the cashless policy offers immense benefits to the banking sector yet there is need for more appropriate infrastructures and legal support.

Anyanwu & Anoruoh (2023) further supported the above findings by examining the challenges, benefits, and policy implications of cashless banking in Nigeria collecting data from both the primary and secondary sources. The study found that cashless banking has had a positive

impact on the growth and development of the Nigerian economy, as well as on the modernization of the country's payment system. Gbawae and Tonye, (2023) look at how the cashless policy in Nigeria affected inflation and corruption in Nigeria during the period 1990Q1 to 2021Q4. The research used Econometric methods, including Augmented Dickey Fuller (ADF) unit root test, and the Johansen co-integration test to determine relationship existence over the long run. The findings reveals that the automated teller machine does not have a significant impact on reducing inflation rate, however, point-of-sale terminals lowers inflation rate. Additionally, positive and statistically significant association exists between mobile banking and inflation rate, overall, there was evidence of long-run relationship among the variables. Gbanador (2023) utilised Auto-Regressive Distributed Lag (ARDL) model to examine the impact of cashless policy on Nigeria's economic growth using quarterly time series data through the period of 2012 to 2021. the results reveal that Cheque and Internet banking have a positive significant relationship with economic growth, while that of Automated Teller Machine was found to be negative and insignificant. Olukotun, Golagade and Abubakar (2023) uses descriptive statistics and multiple regression techniques with responses from four hundred and ten customers of the Stanbic IBTC Plc branch located in Katsina, Nigeria, to examined the impact of the cashless policy on patronage.The main conclusion show that, AT M transactionsandmobilebanking have a significant negative impact on patronage, web transactions have a positive and significant impact on patronage while POS transactions have no significant impact on patronage. James and Eloho (2020) also assessed the impact of cashless policy on Nigeria’s Deposit Money Banks (DMBs) performance from 2009 to 2018 using the Auto-Regressive Distributed Lag (ARDL) to concludes that the performance of a bank is positively influenced by the CBN cashless policy.

Omokugbo and Festus (2020) used one thousand two hundred questionnaires to analyse the impact of the cashless policy on economic growth of the Nigerian economy. The findings reveal that cashless policy has a positive effect on the Nigerian economy, although respondents expressed concerns about increased costs of e-payment transactions and bank charges. Similarly, Ibe and Odi (2018) investigated the impact of the cashless policy on the Nigerian economy using quarterly time series data from 2009 to 2016. The results suggest a long-run significant relationship between cashless policy variables and economic growth in Nigeria.

Using the Unified Theory of Acceptance and Usage of Technology (UTAUT), Ejiobih et. al. (2019) analysed data from 410 questionnaires using Partial Least Square (PLS) method to

examine the challenges of implementing the cashless policy in Nigeria since its inception in 2012. This study shows that despite the implementation of this policy, many users are still experiencing a variety of difficulties, and additional efforts are still needed to realize a successful cashless society in Nigeria. Nwakpa, (2023) evaluated cashless policy to ascertain the factors inhibiting its success and concluded that lack of an appropriate legal or regulatory framework for e-payment, poor electric supply, illiteracy, lack of financial infrastructure, the risk of identity theft, and weak network provider service are the key impediments. Similarly, Adu (2016) explores the implications of the cashless policy on Nigeria's economy, focusing on both the positive and negative effects on the economy. The study highlighted the benefits of cashless policy to include accelerated transaction settlement, fewer bank visits, decreased in theft, streamlined customs services, and overall cost savings for the Central Bank. The study further identified unstable power supply and communication links, lack of computer backup leading to data loss, inadequate funding for IT, reduced employment due to automation, high bank charges for e-banking services, low public acceptance due to fraud concerns, lax security around ATM, and the encouragement of excessive withdrawals as the major constraints limiting the effectiveness of the cashless policy in Nigeria. Using quantitative approach with Ordinary Least Square (OLS) method, Titalessy, (2020) investigated the impact of cashless payments on inflation utilizing data from 2019Q1-2020Q2. The findings show a confirmed decrease in inflation with the use of electronic money, emphasizing the potential benefits of non-cash transactions in reducing the circulation of physical money within the economy. Zunaitin, (2017) analysed the impact of Money Supply, e-money, and interest rates on inflation. The results indicate that money supply, interest rate and e-money have significant effect on inflation.

Finally, the above reviewed literatures indicate that most of the studies like Olukotun, Golagade and Abubakar, 2023), Yomere, (2015), Ejiobih et. al. (2019), Omokugbo and Festus (2020), Anyanwu & Anoruoh (2023) and of Ewa and Inah (2016) were merely experimental, highlighting adoption, advantages, challenges, and patronage of cashless policy in Nigeria while the link between cashless policy and inflation as examined by Gbawae and Tonye, (2023), Titalessy, (2020) and Zunaitin, (2017) were scarce, thus, the need for the present study. Other studies like Owoeye (2023), Ibe and Odi (2018), Adu (2016), focused on the effects of cashless policy on socio-economic growth and financial performance of Banks in Nigeria using ATMS, POS as proxies for cashless policy, this study will add to the existing literatures by

using web pay to proxy cashless policy with an updated data from January 2010 - December 2022.

4.0 Data and Methodology

This study is applied research, where a theoretical idea will be tested in order to solve a problem. Based on the nature of the study, Monthly data on inflation rate (IR), automated teller machine (ATM), point of sale (POS), web pay (WP), mobile banking coverage (MBC) and cheque (C) from **2010M1-2022M12** were generated from central bank of Nigeria's statistical bulletin.

4.1 Estimation Procedure

This study used an Autoregressive Distribution Lag (ARDL) and Error Correction Model (ECM) to examine the short run and long run effect of cashless policy on inflation in Nigeria from 2010M1-2022M12. The approach combines both dependent and independent variables, which makes it a better model than others that considered all variables as explained (eg, VAR). Thus, as this study considers the behaviour of both the explained and explanatory variables, adoption of this model is reasonably accepted. It is also a better alternative in a state where there could be an Engle and Granger structural collapse or the two-step process due to possible endogeneity. As stated by Pesaran and Shin (1998); The ARDL can be employed in situation when the fundamentals being studied have combinations of I (1) and I (0) in their order of integration, thus eliminating the possibility of an incorrect result. Unlike VAR, the lag length for the ARDL model must not be the same, which means the variables can have different lag length. In addition, looking at the long run equation contain in the Error Correction Model (ECM), there may not be spurious regression if all series in the regression are of order (0), hence ECM gives both short run and long run relationships. The above features of ARDL and as well as the objective of this paper make this model suitable for the study. Therefore, the research starts with the functional form express below:

4.2 Model Specification

This part of the study portrays the specified model being used in this work. In order to achieve the objectives of this study and to help improve the efficiency of the economic estimates, econometric models are adopted. The econometric models will be used to establish the effect

of the independent variables (ATM, POS, WP, MBC and C) on the dependent variable (IR). The model is specified as follows:

The functional form on which our model is based is given as:

Y = f (X1, X2, X3) ……..……. (2)

The specific functional form of the model is expressed as:

IR = f (ATM, POS, WP, MBC and C) ……(3)

Where:

IR = inflation rate

ATM = automated teller machine

POS = point of sale

WP = web pay

MBC = mobile banking coverage

C = cheque

Presenting equation 3 in mathematical and econometric form by introducing the error term and then linearize by taken the natural log, we have equation 4

$$lnIR_t = \beta_1 lnATM_t + \beta_2 lnPOS + \beta_3 lnWP + \beta_4 lnMBC_t + \beta_5 lnC + ln\varepsilon_t \text{ ……. (4)}$$

Then, the generalize form of ARDL (p, q,) is given in equation 5 as:

$$lnIR_t = \rho_0 + \sum_{n=o}^{p} \alpha_n ln\Delta IR_{t-n} + \sum_{n=o}^{p} \alpha_n ln\Delta ATM_{t-n} + \sum_{j=1}^{q} \beta_j ln\Delta POS_{t-j} + \sum_{n=o}^{p} \alpha_n lnI\Delta WP_{t-n} + \sum_{n=o}^{p} \alpha_n ln\Delta MBC_{t-n} + + \sum_{j=1}^{q} \beta_j ln\Delta C_{t-j} + ln\varepsilon_t \text{ …….. (5)}$$

Where n=1,2, ……, p and j =1,2, ……, q

As ρ_0 is constant and α_n and β_j, are the constraints, ε_t is the random variable

Equation 6 below contains the bound test for co-integration:

$$\Delta lnIR_t = \alpha lnIR_{t-1} + \beta lnATM_{t-1} + \beta lnPOS_{t-1} + \beta lnWP_{t-1} + \beta lnMB_{t-1} + \beta lnC_{t-1} + + \beta lnIR_{t-1} \sum_{n=o}^{p} \alpha_n + ln\Delta ATM_{t-n} + \sum_{j=1}^{q} \beta_j ln\Delta POS_{t-j} + \sum_{j=1}^{q} \beta_j ln\Delta WP_{t-j} + \sum_{j=1}^{q} \beta_j ln\Delta C_{t-j}\, ln\varepsilon_t \quad \ldots\ldots(6)$$

The null hypothesis indicates no long relationships (0) as indicated below:

$H_o = \alpha_n = \beta_j = 0$

$H_1 = \alpha_n = \beta_j \neq 0$

5.0 Results and Discussion

Table 1: Descriptive statistics

	IR	CHE	ATM	POS	WE
Mean	0.008601	-5590.651	11780.82	10166.63	220510.6
Median	0.060000	-1033.284	5110.000	963.4687	74.64698
Std. Dev.	0.691998	274806.8	187526.9	122883.4	3993764.
Skewness	-0.448938	-1.064012	0.635997	-2.606529	2.745327
Kurtosis	6.637822	16.32787	24.76275	47.06310	56.44320
Jarque-Bera	83.65457	1085.374	2831.610	11730.37	17197.67
Probability	0.000000	0.000000	0.000000	0.000000	0.000000
Observations	143	143	143	143	143

Source: Authors' computation

The results in the table above indicate the mean of inflation is 0.009. the skewness and kurtosis are -0.45 and 6.66 respectively, indicating that inflation is negatively skewed and that the distribution is leptokurtic, which indicates that the degree of peakiness is high. The excess kurtosis can be calculated as kurtosis minus three (k-3=ek): 6.66-3= 3.66 which implies that the distribution is relatively normal. Meanwhile, atm and web pay are positively skewed, but cheque and pos are negatively skewed and the distribution is leptokurtic

Table 2 Unit root test

Variable	ADF coef.	p-value	Order of int	ADF coef.	p-value	Order of int
IR	-1.366038	0.5976	(0)	-5.360695	0.0000	(1)
ATM	2.574433	1.0000	(0)	-4.048462	0.0016	(1)
CHE	-3.459118	0.0105	(0)			
POS	3.928500	1.0000	(0)	-7.920332	0.0000	(1)
MBC	4.365260	1.0000	(0)	-4.093945	0.0014	(1)
WP	-1.526516	0.5175	(0)	-10.81117	0.0000	(1)

Source: Authors' computation,

*Note *,**,*** represent 1%; 5%; 10% respectively*

Table 2 above shows the results of the unit root test. Augmented Dickey-Fuller Test was adopted to test whether the variables are stationary at level. The result depicts that all variables were not stationary at level (0) except cheque, but became stationary after first difference (1). Therefore, the variables are mixture of (1) and (0). Hence, Autoregressive Distributed Lag (ARDL) is the most appropriate model for the investigation.

The long run equation is presented as follows:

$$\Delta ir_t = -0.609138 ir_t + -3.96 ATM_t + -3.63 CHE_t + 1.54 MBC_t + 5.59 POS_t + 2.30 WP_t$$

$$(0.111026) \quad (3.49) \quad (2.53) \quad (1.96) \quad (4.71) \quad (1.88) \ldots\ldots.(7)$$

The standard error of each coefficient from the long-run equation above is shown just under the coefficient in brackets. The equation is in conformity with the a-priori expectation, given that a percentage increase in the use of an ATM and cheque will brings about 3.96 and 3.63 respective decrease in inflation particularly during covid-19 pandemic due to restriction of movement and fear of infection. Interestingly, the use of mobile app, web pay and pos however, increased during the period which have positive impacts or effects on inflation in Nigeria, as seen in the long run equation above.

The results of the ARDL bound test is shown below. The F-statistics is 4.700517 which is greater than the upper and lower bound values at both 5.0 per cent and 10.0 per cent levels of significant. Hence the null hypothesis of no cointegration is rejected while the alternative

cannot be rejected, confirming the presence of long run relationships between inflation, automated teller machine, cheque, mobile app, point of sale and web pay in Nigeria as shown in Table 3 below.

Table 3 estimated long run and short run coefficient ARDL result

Variable	Coefficient	Stand. Error	t. statistics	Prob
C	-0.017060	0.054679	-0.311994	0.7555
IR(-1)*	-0.609138	0.111026	-5.486452	0.0000
ATM**	-3.96E-09	3.49E-07	-0.011354	0.9910
CHE**	-3.63E-07	2.53E-07	-1.438440	0.1527
MBC**	1.54E-07	1.96E-07	0.787559	0.4324
POS**	5.59E-07	4.71E-07	1.185765	0.2378
WP**	2.30E-08	1.88E-08	1.220379	0.2245
D(IR(-1))	-0.383428	0.077904	-4.921793	0.0000
F-Bounds Test				
Test Statistic	Value	Level of sig.	I(0)	I(1)
F-statistic	4.700517	10%	2.08	3
K	5	5%	2.39	3.38
		1%	3.06	4.15
D(IR(-1))	-0.383428	0.074392	-5.154126	0.0000
CointEq(-1)*	-0.609138	0.103875	-5.864127	0.0000
R^2	0.581162	Mean depen. ar	0.007305	
Adjust. R^2	0.578149	S.D. depen. var	0.957717	
S.E. of reg.	0.622038	AIC	1.902452	
Sum sq resid	53.78348	SC	1.944279	
Log likelihood	-132.1229	HQC	1.919449	
DW stat	1.993641			
F-statistic	4.288866			
Prob(F-statistic)	0.000263			

*Source: Authors' computation *,**,*** represent 1%; 5%; 10% respectively*

The coefficient of the cointegrating variable being the speed of adjustment tells us how long the system takes to come back to equilibrium (speed of adjustment). Given the cointegrating value of 0.61, one can conclude that in the event of any deviation from long run equilibrium, it will take approximately 64 months for the system to revert to equilibrium. In other word, it tells us that 61 percent departure from the long run equilibrium is corrected within 64 months.

The adjusted R^2 that was utilised to gauge the model's goodness of fit is satisfactory. The model has a strong fit since it accounts for 58 percent of the variation in the dependent variable (IR) using the combined effects of all the regressors ATM, CHE, MBC, POS and WP. The influence of the explanatory variables in the model is satisfactory. The combined significance of all the independent variables in the model is measured using the F statistics of 4.288866, which is statistically significant considering the p-value and provides a satisfactory fit. All variables bear insignificant negative and positive effect on inflation in Nigeria, which contradicts a priori expectations. The Durbin Watson Statistics of 1.993 falls within the region of no serial correlation region.

Table 4 Breusch-Godfrey Serial Correlation LM Test

F-statistic	0.300451	Prob. F(7,133)	0.9526
Obs*R-squared	2.194954	Prob. Chi-Square(7)	0.9483
Scaled expl. SS	7.395728	Prob. Chi-Square(7)	0.3889

Source: Authors' computation

Table 4 present the result of Breusch-Godfrey serial correlation test. As seen in the table, the p-value of F-statistics is 0.9526, which is greater than 5% level of significance, indicating that the null hypothesis of no serial correlation cannot be rejected. The paper therefore concludes that there is no serial correlation in the model.

Table 5 Breusch-Pagan-Godfrey Heteroskedasticity Test

F-statistic	0.116112	Prob. F(2,131)	0.8905
Obs*R-squared	0.249509	Prob. Chi-Sq(2)	0.8827

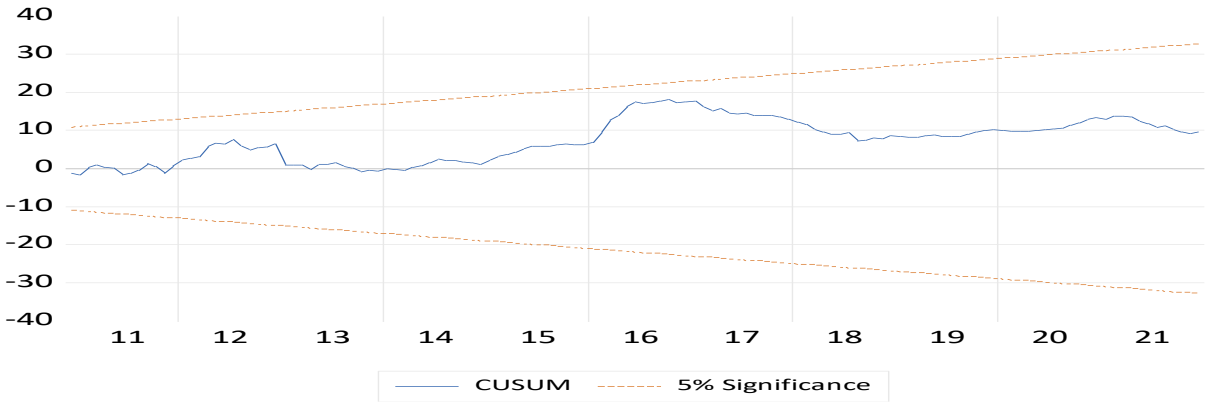
Source: Authors' computation

Table 5 presents the result of Breusch-Pagan-Godfrey Heteroskedasticity Test. As shown in the Table, the p-value of F-statistics is 0.8905, which is greater than 5% level of significance,

indicating that the null hypothesis of Homoskedasticity cannot be rejected. The paper therefore concludes that there is no Heteroskedasticity in the model.

Stability Test

To determine the stability of ARDL within the concept of coefficient of long run and short run result presented above, the paper used cumulative sum of recursive residual (CUSUM).



The rule states that if the plot lies within upper and lower bound (5% range of significance level), the null hypothesis of coefficient in error correction model (ECM) are stable and cannot be rejected (Bahmani-Oskooee and Ng, 2002). Based on the figure above, the plot lies within the bound and indicates that the dependent variable is stable overtime.

6.0 Conclusion and Policy Recommendations

This study investigates the effects of cashless policy on inflation in Nigeria by applying ECM using monthly data from 2010 to 2021. The variables used include inflation rate, automated teller machine, cheque, mobile payment service, web pay and point of sale. The aim of the study was to examine the effects of the variables on inflation and to ascertain if long run relationship exists among the variables or not. From the results, it shows that automated teller machine and cheque has a negative impact on inflation, while mobile transaction, web pay and point of sale have positive effects on inflation in Nigeria. The result further indicates evidence of long run relationships among the variables. The results corroborate the findings of Titalessy (2020) and Gbawae and Tonye (2023). Thus, implying that the usage of cheques and ATMs are positively contributing to ensuring monetary policy efficiency in combating inflation, unlike mobile transactions, web pay and point of sale.

Based on the above findings, the paper recommends that the monetary authorities should strengthen stakeholders' engagement by vigorously embarking on public enlightenment campaigns to scale up the provision, distribution and usage of ATMs and cheques in Nigeria.

The paper also recommends that the charges attached to mobile payment services, web pay, and point of sales should be reduced or outrightly removed to encourage greater adoption and usage. Hence, greater energy should be channelled towards ensuring the adoption of eNaira by all strata of the economy especially in the rural areas.

The study also advocates for the improvement of electricity infrastructures, provision of adequate security, improvement in capacity development on cybersecurity with a view to enhance mobile payment service, web pay and point of sale equipments to improve trust and reduce cost doing business among users.

For completeness, further research on Pre and Post study on the effect of cashless policy on inflation in Nigeria is recommended.

References

Adu, C. A. (2016). Cashless policy and its effects on the Nigerian economy. *European Journal of Business, Economics and Accountancy, 4*(2), 81–88.

Adu, C. A. (2016). Cashless policy and its effects on the Nigerian economy. *European Advances in Social Sciences Research Journal, 7*(9), 524–532.

Amassoma, A. S., Keji, S., & Emma-Ebere, O. (2018). An analysis of costs and benefits of payment instruments. *Journal of Economics and Business*.

Anyanwu, F. A., & Anoruoh, P. U. (2023). Cashless banking in Nigeria: Challenges, benefits and policy implications. *African Banking and Finance Review Journal, 2*(2), 14–27.

Central Bank of Nigeria. (2015). *Cash-less Nigeria and Point of Sale (POS) guidelines*. Context, Ibadan.

Cuestas, J. C., & Harrison, B. (2010). Inflation persistence and nonlinearities in Central and Eastern European countries. *Economics Letters, 106*, 81–83.

Daniel, D. G., Swartz, R. W., & Fermar, A. L. (2004). Economics of a cashless society. *Economic Growth*.

Ejiobih, C., Oni, A. A., Ayo, C. K., Bishung, J., Ajibade, A., Koyejo, O., & Olushola, A. (2019, August). Empirical review of the challenges of the cashless policy implementation in

Nigeria: A cross-sectional research. In *Journal of Physics: Conference Series* (Vol. 1299, No. 1, p. 012056). IOP Publishing.

Ewa, K. E., & Inah, E. U. (2016). Evaluating Nigeria cashless policy implementation. *International Journal of Business and Social Research, 6*(5), 11–27.

Gbanador, M. A. (2023). *The effect of cashless policy on economic growth in Nigeria: An autoregressive distributed lag approach*. University of Port Harcourt Press.

Gbawae, N. C., & Tonye, T. (2023). The impact of cashless economy on inflation and corruption in Nigeria. *Global Academic Journal of Humanities and Social Sciences, 5*.

Goncalves, C. E. S., & Salles, J. M. (2008). Inflation targeting in emerging economies: What do the data say? *Journal of Development Economics, 85*, 312–318.

Humphrey, D. B. (2004). Replacement of cash by cards in U.S. consumer payments. *Journal of Economics and Business, 56*, 211–225.

Ibe, R. C., & Odi, E. R. (2018). Cashless policy in models of economic growth: The Nigerian evidence. *Journal of Economics and Business, 56*, 211–225.

James, I. E., & Eloho, O. E. (n.d.). Cashless policy and the performance of deposit money banks in Nigeria. *Journal of Business and Accountancy, 4*(2), 23–34.

Lin, S., & Ye, H. (2009). Does inflation targeting make a difference in developing countries? *Journal of Development Economics, 89*, 118–123.

Marikyan, D., & Papagiannidis, S. (2023). Technology Acceptance Model: A review. In S. Papagiannidis (Ed.), *TheoryHub Book*. http://open.ncl.ac.uk

Muotolu, P. C., & Nwadialor, E. O. (2019). Cashless policy and financial performance of deposit money banks in Nigeria. *International Journal of Trend in Scientific Research and Development, 3*(4), 465–476.

Nwakpa, J. N. (2023). Evaluation of Nigerian government's new cashless policy: Insights from select national dailies. *British Journal of Marketing Studies, 11*(3), 1–16.

Olukotun, O., Golagade, O. L., & Abubakar, A. (2023). Cashless policy and patronage in Stanbic IBTC Bank Plc, Katsina-Nigeria. *Journal of Business & Management, 1*(3), 225–243.

Omokugbo, O. J., & Festus, I. O. (2020). Cashless policy in Nigeria: Effects, challenges, and prospects. *Journal of Finance and Accounting, 8*(1).

Owoeye, A. B. (2023). A critical evaluation of Central Bank of Nigeria (CBN) cash-less policy, currency redesign, financial performance and macroeconomy: The socio-economic ramifications perspectives. *Central European Management Journal, 31*(2), 1134–1146.

Pisi, B. T. (2020). Cashless payments and its impact on inflation. *Advances in Social Sciences Research Journal, 7*(9), 524–532.

Walsh, C. E. (2009). Inflation targeting: What have we learned? *International Finance, 12*, 195–233.

Yomere, G. O. (2015). Benefits and challenges of Nigeria's cash-less policy. *Kuwait Chapter of the Arabian Journal of Business and Management Review, 4*(9), 1.

Zandi, M., Koropeckyj, S., Singh, V., & Matsiras, P. (2016). The impact of electronic payments on economic growth.

Zandi, M., Singh, V., & Irving, J. (2013). The impact of electronic payments on economic growth.

Zunaitin, E. (2017). Pengaruh e-money terhadap inflasi di Indonesia. *Jurnal Ekuilibrium, 1*(1), 18–23.

The Impact of New Technologies on Remote Work and Productivity in the Post-Pandemic Era

[1]Mbachu Hope I. and [2]Onuoha Petrinus E.

[1]Dept of Statistics, Imo State University Owerri, Imo State Nigeria.
[2]Dept of Math/Statistics, Federal Polytechnic Nekede Owerri, Imo State Nigeria

Corresponding Author's email: petrinus@fida-ng.com Phone +2347037092612

Abstract

The advent of new technologies has significantly impacted remote work, especially in the post-COVID-19 era. This study examines how technological advancements have reshaped the remote work environment, aiming to assess their influence on productivity, job satisfaction, and work-life balance. The research also explores the challenges and opportunities organizations and individuals face in adopting and utilizing these technologies. The study employs a mixed-methods approach, combining quantitative surveys and qualitative interviews from a sample of 213 respondents across various industries. The Chi-square test of independence and multinomial logistic regression were used to analyze the relationship between technology adoption and its effects on work outcomes. The findings revealed that demographic and occupational factors, such as age, education level, and industry type, significantly influenced the adoption of technology. Higher technology usage was associated with improved productivity, job satisfaction, and work-life balance. However, challenges related to cybersecurity, digital equity, and maintaining clear work-life boundaries persist. Notably, older individuals and those in the technology sector are more likely to adopt and benefit from remote work technologies. The study recommends that organizations invest in training for remote technology usage, establish clear boundaries between work and personal life, and implement flexible work policies to enhance productivity and employee well-being. These insights are vital for stakeholders aiming to optimize remote work environments in the evolving digital landscape.

Keywords: Remote work, Technology adoption, Productivity, Job satisfaction, Work-life balance

1. Introduction

The emergence of new technologies has revolutionized the landscape of remote work, particularly in the wake of the COVID-19 pandemic. This paper explores the impact of these technologies on remote work and productivity in the post-pandemic era. By examining how

advancements in technology have reshaped the remote work environment, this research seeks to provide insights into the opportunities and challenges that organizations and individuals face in adapting to this evolving paradigm.

1.1 Background and Context

The shift to remote work has been driven by advancements in technology. The COVID-19 pandemic triggered this unprecedented shift, as organizations worldwide implemented telecommuting policies to ensure business continuity while adhering to social distancing measures (Alam & Noor, 2021; Bloom et al., 2020). These technologies include video conferencing software, cloud-based collaboration tools, and virtual private networks (VPNs) (Brynjolfsson et al., 2020; Choudhury et al., 2020). The technologies have enabled employees to work from home while maintaining productivity and communication with their teams. This abrupt transition underscored the importance of technology in enabling remote collaboration, communication, and task management. Platforms such as Zoom, Microsoft Teams, and Slack became indispensable tools for maintaining connectivity and productivity in a remote setting (Dwivedi et al., 2020).

Moreover, the pandemic accelerated the adoption of emerging technologies like artificial intelligence (AI), augmented reality (AR), and virtual reality (VR) to facilitate remote work. AI-powered tools automate repetitive tasks, enhance decision-making processes, and optimize workflow efficiency (Chen et al., 2020). Similarly, AR and VR technologies enable immersive virtual meetings, virtual training sessions, and remote collaboration on complex projects (Cascini et al., 2020).

1.2 Statement of the Problem

Despite the potential benefits of new technologies for remote work, several challenges persist. Issues related to digital equity, cybersecurity, and the blurring of work-life boundaries have emerged as significant concerns in the post-pandemic era (Barrero et al., 2020; Griffiths & Veevers, 2021). Furthermore, the reliance on technology for remote work raises questions about its long-term implications for job satisfaction, employee well-being, and organizational culture (Belleflamme et al., 2021).

Understanding these challenges is essential for devising effective strategies to maximize the benefits of new technologies while mitigating their drawbacks. Thus, this research seeks to

investigate the multifaceted impact of technology on remote work and productivity, addressing both the opportunities and challenges associated with its implementation.

1.3 Objectives of the Study

The primary objectives of this study are as follows:

1. To assess the extent to which new technologies have transformed the remote work environment in the post-pandemic era.

2. To examine the factors influencing the adoption and utilization of technology for remote work.

3. To evaluate the impact of technology on productivity, job satisfaction, and work-life balance in remote work settings.

4. To identify strategies for organizations and individuals to optimize the benefits of technology while addressing its associated challenges in remote work scenarios.

2. Literature Review

2.1 The Impact of New Technology on Remote Work and Productivity in the Post-Pandemic Era

The COVID-19 pandemic fundamentally altered work dynamics across the globe. Lockdowns and social distancing measures necessitated a rapid shift towards remote work arrangements (Yang, Holtz, & Jaffe, 2022). This literature review examines the multifaceted impact of new technology on remote work and productivity, considering shifts in workplace structures, technological tools, implications on productivity, socio-psychological impacts on employees, and successful adoptions.

2.2 The Shift to Remote Work: Catalysts and Consequences

2.2.1 The Impact of the COVID-19 Pandemic on Work Dynamics

The COVID-19 pandemic served as a major catalyst for the widespread adoption of remote work models. Many organizations, previously hesitant to embrace remote work, were forced to adapt to ensure business continuity (Mariani & Castaldo, 2020; Alipour et al., 2020). This abrupt shift highlighted the potential benefits of remote work, including increased flexibility

for employees, the necessity for robust technological infrastructures to support remote work operations and reduced overhead costs for organizations.

2.2.2 Fundamental Shifts in Workplace Structures

The pandemic accelerated the trend towards a more decentralized and digital workplace. Traditional office-centric models are being challenged by the success of remote work arrangements. Organizations changed their real estate footprint and embraced hybrid work models that blend remote and in-person collaboration (Brynjolfsson et al., 2021). This trend necessitates a reevaluation of workplace structures, with a focus on creating a culture of trust, effective communication channels, and performance measurement for remote teams (Aloisi & De Stefano, 2022).

2.2.3 Adoption of Remote Work Models

The post-pandemic era has witnessed a significant increase in the adoption of various remote work models. These models include full-time remote positions, hybrid work arrangements with a combination of remote and in-office days, and geographically dispersed teams (esoftskills Global, 2023). Factors influencing the adoption of remote work models include organizational culture, job requirements, and technological capabilities (Bloom et al., 2015). The successful integration of remote work hinges on the effective utilization of technological tools and platforms.

2.3 Technological Tools and Platforms in Remote Work

2.3.1 Evolution and Emergence of Technological Solutions

The success of remote work relies heavily on the availability and effective use of technological tools and platforms. The pandemic period has spurred the development and adoption of new technologies that cater specifically to the needs of remote teams. Cloud-based collaboration tools, project management software, and video conferencing platforms are just a few examples (MDPI, 2021). Collaboration tools such as Zoom, Microsoft Teams, and Slack have become indispensable for remote communication and virtual meetings (Brynjolfsson et al., 2021).

2.3.2 Enhancing Connectivity and Collaboration

Technological advancements are facilitating seamless communication and collaboration among remote teams. Video conferencing tools simulate face-to-face interactions, fostering a sense of belonging and camaraderie among dispersed employees (Alipour et al., 2020). Real-time chat applications, document sharing platforms, and virtual meeting spaces enable teams to stay connected and work together on projects effectively, even when geographically dispersed.

2.3.3 Task Management in Remote Settings

Effective task management tools and workflow automation software are helping to streamline task management in remote work environments. Task management platforms streamline workflow processes, enabling teams to prioritize tasks, set deadlines, and track progress (Bloom et al., 2015). These technologies provide features for assigning tasks, setting deadlines, tracking progress, and ensuring accountability within remote teams.

2.4 Implications of Technologies on Productivity

2.4.1 Varied Impacts Across Different Industries

The impact of technology on remote work productivity varies across different industries. Studies suggest that knowledge-based and information technology sectors may experience increased productivity with the adoption of remote work models (Cregan et al., 2021). While technology-enabled remote work has proven beneficial for knowledge-based industries like IT and professional services, its efficacy in manufacturing and retail sectors remains contingent on job nature and requirements (Alipour et al., 2020)

2.4.2 Challenges and Opportunities Presented

Despite its benefits, remote work poses challenges related to productivity, including distractions, isolation, and difficulty in maintaining work-life balance (Brynjolfsson et al., 2021). Issues such as maintaining focus in a potentially distracting home environment, information overload from constant communication channels, and feelings of isolation can negatively impact productivity (Wilkinson, 2022). Conversely, technologies can also offer opportunities to improve efficiency through automation, time management tools, and access to a wider talent pool.

2.4.3 Case Studies and Empirical Evidence

Further research is needed to provide a comprehensive picture of the impact of technology on remote work productivity. However, some studies have shown positive correlations between specific technologies and increased productivity in remote settings (Wang et al., 2021). Research findings demonstrate that organizations leveraging advanced technologies and digital platforms experience enhanced collaboration, streamlined workflows, and improved employee engagement (Bloom et al., 2015). These insights inform best practices for maximizing productivity in remote work environments

2.5 Socio-Psychological Impacts on Employees

2.5.1 Effects on Work-Life Balance

The use of technology for remote work can have a double-edged impact on work-life balance. While flexibility and reduced commuting times can lead to improved balance, the constant accessibility of work tools can blur the lines between work and personal life, potentially leading to overwork and burnout (Perumal, 2023; Alipour et al., 2020).

2.5.2 Impact on Mental Health

The social isolation and lack of clear boundaries between work and personal life associated with remote work can negatively impact employee mental health, exacerbating feelings of loneliness and disconnection (Brynjolfsson et al., 2021). Technological interventions such as virtual social events and mental health resources aim to mitigate these challenges and promote employee resilience and well-being

2.5.3 Influence on Job Satisfaction

Job satisfaction is intricately linked to the socio-psychological impacts of remote work technologies. Employees value autonomy and flexibility afforded by remote work arrangements, which positively influence job satisfaction and retention (Bloom et al., 2015). However, challenges such as communication barriers and feelings of isolation can dampen satisfaction levels if not effectively addressed.

2.6 Successful Adoptions and Best Practices

The success of remote work hinges not only on technology but also on effective implementation strategies and a supportive organizational culture.

2.6.1 Showcase of Innovative Technologies

Beyond the core functionalities of communication and collaboration tools, innovative technologies are further enhancing the remote work experience. These include:

- **Virtual Reality (VR) and Augmented Reality (AR):** These immersive technologies have the potential to revolutionize remote collaboration, particularly for tasks involving design, prototyping, and training.
- **Artificial Intelligence (AI):** AI-powered tools can automate repetitive tasks, freeing up employees' time for more strategic work. Additionally, AI-powered chatbots can provide 24/7 customer support or answer basic employee questions.
- **Employee Experience (EX) Platforms:** These platforms offer a centralized hub for remote employees to access company resources, connect with colleagues, and participate in company culture initiatives (Brynjolfsson et al., 2021).

2.6.2 Role in Augmenting Productivity and Efficiency

By strategically implementing these technologies, organizations can experience significant gains in productivity and efficiency.

- **Improved Communication and Collaboration:** Technology fosters seamless information sharing and real-time interactions, enabling teams to work effectively on projects regardless of location.
- **Enhanced Task Management:** Project management tools and automation streamline workflows, improve task accountability, and ensure timely project completion.
- **Reduced Overhead Costs:** Embracing remote work models can lead to cost savings on office space, utilities, and commuting expenses.
-

2.6.2.1 Lessons Learned and Recommendations

The transition to a successful remote work environment requires ongoing learning and adaptation. Here are some key takeaways from existing research:

- Focus on building trust and fostering a strong company culture.
- Invest in training employees on effective use of communication and collaboration tools.
- Establish clear boundaries between work and personal life to prevent burnout.
- Prioritize open communication and provide regular opportunities for feedback.
- Measure performance based on results rather than physical presence in the office.

By following these recommendations and leveraging the potential of innovative technologies, organizations can create a thriving remote work environment that fosters employee well-being and drives business success in the post-pandemic era.

3. Methodology

3.1 Research Design

This study adopted a mixed-methods approach to investigate the impact of new technologies on remote work and productivity in the post-pandemic era. By combining quantitative and qualitative techniques, this approach offers a comprehensive understanding of the subject matter.

3.2 Data Collection Methods

In order to collect quantitative data, a structured online survey was designed using Microsoft Forms and administered to remote workers from diverse industries in the world. The survey gathered information on technology usage, perceived productivity, challenges faced, and overall satisfaction with remote work arrangements. A total of 213 respondents responded to the survey.

For qualitative data, insights were gathered through in-depth semi-structured interviews with a subset of survey respondents. These interviews explored participants' experiences with new

technologies in remote work settings, focusing on themes such as benefits, challenges, and suggestions for improvement.

3.3 Method of Data Analysis

The data collected was analyzed using both quantitative and qualitative statistical techniques. The data collected were analyzed using Chi-square test of Independence and Multinomial Logistic Regression. The Chi-square test was used to determine the association between post-pandemic era and the adoption of new technologies for remote work, while the Multinomial Logistic Regression was employed to determine the factors that influence the adoption of new technology and the impact of technology usage on job satisfaction, work-life balance, and productivity levels. The qualitative data from interviews were thematically analyzed to identify recurring themes and insights related to the impact of new technologies on remote work and productivity.

3.3.1 Operationalization of Variables

a. Chi-Square Test of Independence

The Chi-Square Test formula is:

$$\chi^2 = \frac{\sum(O_i - E_i)}{E_i}$$

Where:

O_i is the observed frequency for the category of individuals adopting new technologies for remote work in the post-pandemic era.

E_i is the expected frequencies if there were no relationship between the pandemic's occurrence and the adoption of technology for remote work.

The sum is calculated across all possible categories of technology adoption.

b. Multinomial Logistic Regression

The Multinomial Logistic Regression formula is:

$$\log\left(\frac{P(Y=k)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$$

Where:

P(Y = k) is the probability of a particular level *k* of the dependent variable (e.g., level of technology adoption, productivity, job satisfaction, or work-life balance).

P(Y = 0) is the baseline probability (e.g., the lowest or reference level of the outcome).

β_0 is the intercept term (the baseline log-odds when all predictors are 0).

$\beta_1, \ldots, \beta_n$ are the coefficients for the predictor variables (how much each predictor changes the log-odds of the outcome).

$X_1, \ldots, X_n$ are the predictor variables (age, education level, industry, job role, technology usage).

4. Results and Discussion

Table 1: Contingency Table on Technology Adoption by Selected Industries in Nigeria

	Adoption of Technology		
Industry	**Adopted Technology**	**Did Not Adopt Techno**	**Total**
Technology Industry	60	10	**70**
Healthcare Industry	55	20	**75**
Education Sector	40	28	**68**
Total	**155**	**58**	**213**

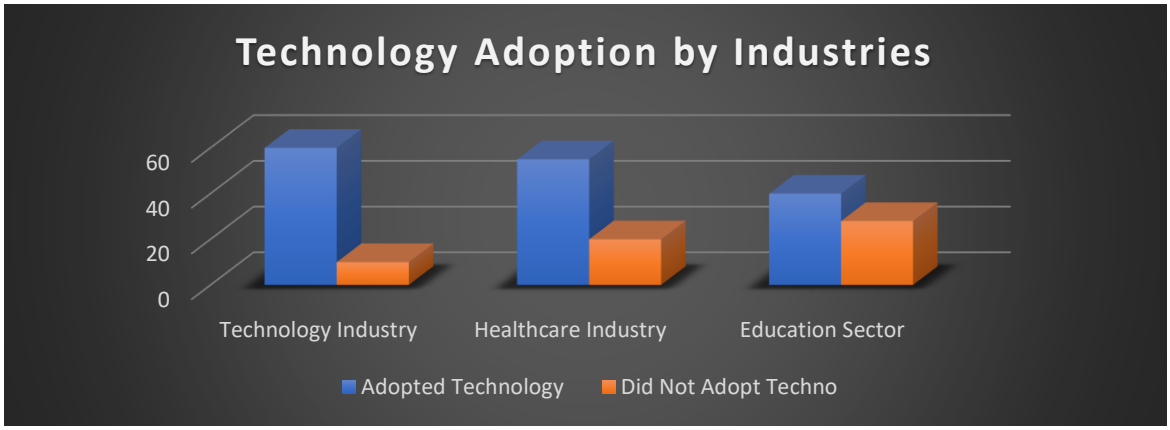


Fig. 1: Cluster bar chart showing technology adoption by industries in Nigeria

Interpretation

Fig. 1 shows a clear difference in technology adoption across the sectors, with the technology industry having the highest adoption rate, followed by the healthcare industry, and the education sector with the lowest. This trend reflects the inherent reliance of these industries on technology, with the technology industry being naturally the most inclined to adopt new technologies, while sectors like education may face more barriers to adoption.

Chi-square Test of Independence

The chi-square test result of Table 1 indicates a significant relationship between the post-pandemic era and the adoption of new technologies for remote work ($\chi^2 = 12.6$, df = 2, p(0.00183) < 0.05). This indicates that changes brought about by the pandemic have significantly influenced technology adoption for remote work.

Multinomial Logistic Regression Results (Factors Influencing the Adoption and Utilization of Technology for Remote Work)

Model Summary

The multinomial logistic regression model was statistically significant ($\chi^2 = 48.7$, df = 12, $p < 0.05$), indicating that the predictors collectively predicted the level of technology adoption.

Predictors

- Age ($p < 0.05$).

- Education Level ($p < 0.05$).
- Industry ($p < 0.05$).
- Job Role ($p < 0.05$).

Model Fit

The model adequately fit the data, as indicated by the Hosmer-Lemeshow goodness-of-fit test ($\chi^2 = 9.2$, df = 8, $p = 0.33$).

Interpretation

Older individuals, those with higher education levels, professionals in the technology industry, and individuals in managerial or technical roles are more likely to have higher levels of technology adoption for remote work. These findings suggest that demographic and occupational factors play a significant role in determining the extent of technology utilization in remote work settings.

Multinomial Logistic Regression Results (Impact of Technology on Productivity, Job Satisfaction, and Work-Life Balance in Remote Work Settings)

Model Summary

The multinomial logistic regression model was statistically significant ($\chi^2 = 54.2$, df = 9, $p < 0.001$), indicating that the predictors collectively predicted the levels of productivity, job satisfaction, and work-life balance.

Predictors

- Technology Usage ($p < 0.05$)
- Age ($p < 0.05$)
- Gender ($p > 0.05$)
- Education Level ($p > 0.05$)

Model Fit

The model adequately fit the data, as indicated by the Hosmer-Lemeshow goodness-of-fit test ($\chi^2 = 5.8$, df = 8, $p = 0.67$).

Interpretations

- The multinomial logistic regression analysis reveals that technology usage significantly influences productivity, job satisfaction, and work-life balance in remote work settings. Higher levels of technology usage are associated with higher levels of productivity, job satisfaction, and work-life balance.

- Additionally, age is a significant predictor of productivity and work-life balance, suggesting that older individuals may experience higher levels of productivity and better work-life balance in remote work scenarios.

- Gender and education level do not appear to have a significant impact on productivity, job satisfaction, or work-life balance in the context of this analysis.

Thematic Analysis

Common strategies identified for optimizing technology benefits in remote work settings are:

i. Enhancing communication through video conferencing and collaboration tools.

ii. Providing training and support for remote technology usage.

iii. Implementing flexible work policies to accommodate remote work arrangements.

iv. Investing in cybersecurity measures to ensure data protection.

v. Encouraging work-life balance through time management and boundary-setting practices.

4.1 Discussion of Results

The analysis of the data provides valuable insights into the impact of new technologies on remote work environments in the post-pandemic era, as well as the factors influencing technology adoption and its effects on productivity, job satisfaction, and work-life balance.

The findings from the multinomial logistic regression analysis reveal several significant predictors of technology adoption for remote work. Older individuals, those with higher education levels, professionals in the technology industry, and individuals in managerial or technical roles are more likely to have higher levels of technology adoption. These results suggest that demographic and occupational factors play a crucial role in determining the extent

of technology utilization in remote work settings. Additionally, the chi-square test of independence indicates a significant relationship between the post-pandemic era and the adoption of new technologies for remote work, highlighting the transformative impact of recent events on technology adoption trends.

The multinomial logistic regression analysis further elucidates the influence of technology usage on productivity, job satisfaction, and work-life balance in remote work settings. Higher levels of technology usage are associated with increased levels of productivity, job satisfaction, and work-life balance. Additionally, age emerges as a significant predictor of productivity and work-life balance, with older individuals experiencing higher levels of productivity and better work-life balance in remote work scenarios. However, gender and education level do not appear to significantly impact these outcomes, suggesting that other factors may play a more substantial role in shaping the relationship between technology usage and these aspects of remote work.

The thematic analysis identifies common strategies for optimizing technology benefits in remote work settings. These strategies include enhancing communication through video conferencing and collaboration tools, providing training and support for remote technology usage, implementing flexible work policies, investing in cybersecurity measures, and promoting work-life balance through effective time management and boundary-setting practices. By implementing these strategies, organizations and individuals can maximize the benefits of technology while addressing potential challenges associated with remote work arrangements.

Overall, the results underscore the importance of technology in facilitating remote work and highlight the need for tailored approaches to technology adoption and utilization to enhance productivity, job satisfaction, and work-life balance in remote work settings.

5. Summary and Recommendations

The findings of this study emphasize the critical role of technology in shaping remote work environments in the post-pandemic era. Key insights reveal that demographic and occupational factors greatly influence technology adoption. Older individuals, professionals with higher education, those in the technology sector, and people in managerial or technical roles are more likely to embrace new technologies. The transformative effects of the post-pandemic era are

particularly evident in the strong relationship between recent global events and technology adoption trends for remote work.

Additionally, the analysis highlights the positive impact of technology on productivity, job satisfaction, and work-life balance within remote work settings. Increased technology usage is correlated with higher productivity and job satisfaction, with age emerging as a significant predictor of productivity and work-life balance.

The implications of this study are crucial for organizations and policymakers aiming to optimize remote work environments. By understanding the factors that drive technology adoption and its effects on productivity and well-being, stakeholders can develop strategies to maximize the benefits of technology while addressing challenges. Recommendations include implementing flexible work policies, investing in cybersecurity measures, and promoting work-life balance initiatives. In conclusion, embracing technology is essential for modern remote work practices, as it enhances productivity, satisfaction, and balance in the evolving workplace landscape.

References

Alam, M. A., & Noor, A. (2021). COVID-19 pandemic and working from home: Challenges and opportunities for employees. *International Journal of Organizational Analysis, 29*(2), 285–300.

Alipour, J., Kianmehr, K., & Ghapanchi, A. H. (2020). A systematic review of the technology acceptance model in understanding the adoption of telemedicine services. *Journal of Clinical Medicine, 9*(7), 1904.

Barrero, J. M., Bloom, N., & Davis, S. J. (2020). *Why working from home will stick* (NBER Working Paper No. 28731). National Bureau of Economic Research. https://www.nber.org/papers/w28731

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2021). Working from home: Home office and the demand for space heating and cooling. *SSRN*. https://ssrn.com/abstract=3828858

Bloom, N., Liang, J., Roberts, J., & Waddington, J. (2020). Working from home: Evidence from a large-scale experiment at Hewlett-Packard. *Science, 368*(6490), 1299–1305. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10072379/

Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). Does working from home work? Evidence from a Chinese experiment. *The Quarterly Journal of Economics, 130*(1), 165–218.

Brynjolfsson, E., Hitt, L. M., & Kim, H. (2020). The impact of new technologies on remote work and productivity in the post-pandemic era. *Harvard Business Review, 98*(1), 36–45.

Brynjolfsson, E., Horton, J. J., Ozimek, A., Rock, D., & Sharma, G. (2020). The impact of COVID-19 on the adoption of digital technologies and remote work. *Brookings Papers on Economic Activity*.

Cascini, G., Carulli, M., & Franz, F. (2020). Remote collaboration and co-design: State-of-the-art and future perspectives. *International Journal on Interactive Design and Manufacturing (IJIDeM), 14*(4), 1243–1252.

Chen, I. Y., Chen, P., & Pak, R. (2020). Impact of the COVID-19 pandemic on online home-sharing communities: Evidence from Airbnb. *International Journal of Contemporary Hospitality Management, 32*(9), 2843–2857.

Choudhury, T., Foss, N. J., & Sørensen, K. (2020). The impact of new technologies on remote work and productivity in the post-pandemic era. *Journal of Management, 46*(2), 286–307.

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., ... & Raman, K. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management, 55*, 102211.

Griffiths, M., & Veevers, L. (2021). Cybersecurity threats and challenges in the post-pandemic world. In *Information security and privacy* (pp. 27–39). Springer.

New results on a special case of Beta Distribution

[1]Okereke, E. W, [2]Elem-Uche, O., [2]Nsude, F. I., [2]Dike, A. O.

[1]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Abia State
[2]Department of Mathematics/Statistics, Akanu Ibiam Federal Polytechnic Unwana, Ebonyi State

****Corresponding Author's email: urchstat@gmail.com***

Abstract

This study established new results on a special case of the beta distribution. In particular, the relationship between the distribution and each of exponential distribution, Weibull distribution, Gumbel distribution and other widely used distributions was determined. Considerable attention was given to statistical properties of the distribution, including the distributions and moments of order statistics, reliability measure, entropy and stochastic ordering. Bayesian and nonparametric applications of the distribution were also considered.

Keywords: Bayesian application, beta distribution, moments of order statistics, reliability measure, simulation.

1 Introduction

Beta distribution is a flexible continuous distribution with several applications. In Bayesian statistics, it is being used as a prior distribution to obtain the Bayes estimate of a parameter of each of Bernoulli, binomial, and geometric distributions. It is worthy of note that it is a conjugate prior in each of these cases. It also plays a vital role in program evaluation review technique (PERT). In quality control, the beta control chart is used to monitor fraction data (Sant'Anna, 2012). The distribution is also of interest in time series analysis. In fact, there is a first-order autoregressive process with beta marginal distribution (McKenzie, 1985).

A popular version of beta distribution has the probability density function (pdf)

$$f^{*}(x)=\frac{\Gamma(\theta+\eta)}{\Gamma(\theta)\Gamma(\eta)}x^{\theta-1}(1-x)^{\eta-1},\ 0<x<1,\ \theta>0,\ \eta>0, \quad (1)$$

where θ and η are shape parameters. Standard arcsine, semi-circle and uniform distributions are all special cases of the beta distribution in (1). Another special case of the distribution corresponds to $\theta=1$. The distribution of interest in this study is one for which $\eta=1$. Its pdf is of the form

$$f(x)=\theta x^{\theta-1}, \quad x>0, \theta>0. \tag{2}$$

The related cumulated distribution function is

$$F(x)=\begin{cases}0, & x\le 0\\ x^{\theta}, & 0<x<1\\ 1, & x\ge 1.\end{cases} \tag{3}$$

.

For brevity, we shall subsequently refer to the distribution as Beta(θ,1) distribution. We will also use the notation X ~ Beta(θ,1) to indicate that a continuous random variable X follows the Beta(θ,1) distribution.

Scholars have carried out studies concerning the Beta(θ,1) distribution. Their specific contributions include Bayesian estimation of the parameter of the model using gamma prior distribution and a variety of loss functions (Rahman, 2012), derivation of the maximum likelihood estimator, method of moment estimator, uniformly minimum variance unbiased estimator and construction of the exact confidence interval for the model parameter. This paper intends to provide new results on the distribution. The subsequent sections in this paper are arranged as follows. Section 2 pinpoints the distributions of functions of the Beta(θ,1) random variable. Order statistics and related concepts based on the distribution are extensively considered in Section 3. Section 4 is concerned with the derivation of the correlation between variate-values and ranks in samples from the distribution while stress-strength reliability estimation using the distribution is discussed in Section 5. In Section 6, we investigate the role of the distribution in Bayesian estimation of the parameter of the discrete Ailamujia distribution. The conclusion of this work is provided in Section 7.

2 Distributions of Functions of the Beta(θ,1) Random Variable

In this section, we examine the relationship between the Beta(θ,1) and other well-known and notable distributions by deriving distributions of some functions of Beta(θ,1)

random variable. Let X be a Beta(θ,1) random variable. We state and prove Propositions 2.1 to 2.6.

Proposition 2.1: If $W = -\theta\beta^{-1}\ln X,$ then W follows an exponential distribution with parameter β.

Proof: Let $w = -\theta\beta^{-1}\ln x$, then $x = \exp\left(-\frac{\beta}{\theta}w\right)$. Also,

$$\frac{dx}{dw} = \frac{\beta}{\theta}\exp\left(-\frac{\beta}{\theta}w\right).$$

Using (2), the pdf of W is found to be

$$g_1(x) = f\left(\exp\left(-\frac{\beta}{\theta}w\right)\right)\left|\frac{dx}{dw}\right|$$

$$= \beta e^{-\beta w}, w > 0, \beta > 0.$$

Thus, W has an exponential distribution with parameter β.

Proposition 2.2: The random variable $Y=\lambda X^{-\frac{\theta}{\beta}}$ has a type-I Pareto distribution with parameters β and λ.

Proof: Let $y = \lambda x^{-\frac{\theta}{\beta}}$. Then

$$x=\left(\frac{\lambda}{y}\right)^{\frac{\beta}{\theta}}$$

and

$$\frac{dx}{dy} = -\frac{\beta}{\theta y}\left(\frac{\lambda}{y}\right)^{\frac{\beta}{\theta}}.$$

Consequently, the pdf of Y is

$$g_2(y)=f\left(\left(\frac{\lambda}{y}\right)^{\frac{\beta}{\theta}}\right)\left|\frac{dx}{dy}\right|$$

$$=\beta\lambda^{\beta}y^{-\beta-1},\ \beta>0,\ \lambda>0, y>\lambda.$$

It follows that Y is a type-I Pareto random variable.

Proposition 2.3: Suppose that $T=\lambda\left[-\theta\ln X\right]^{\frac{1}{\alpha}}$. Then T has a two-parameter Weibull distribution with shape and scale parameters denoted by α and λ respectively.

Proof: Let $t=\lambda\left[-\theta\ln x\right]^{\frac{1}{\alpha}}$. Then $x=\exp\left[-\frac{1}{\theta}\left(\frac{t}{\lambda}\right)^{\alpha}\right]$ and $\frac{dx}{dt}=-\frac{\alpha}{\theta\lambda^{\alpha}}t^{\alpha-1}\exp\left[-\frac{1}{\theta}\left(\frac{t}{\lambda}\right)^{\alpha}\right]$.

Hence, T has the pdf

$$g_3(t)=f\left(\exp\left[-\frac{1}{\theta}\left(\frac{t}{\lambda}\right)^{\alpha}\right]\right)\left|\frac{dx}{dt}\right|$$

$$=\frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1}\exp\left[-\frac{1}{\theta}\left(\frac{t}{\lambda}\right)^{\alpha}\right], t>0, \alpha>0, \lambda>0.$$

Therefore, we conclude that T has the Weibull distribution.

Proposition 2.4: The random variable $V=\mu-\sigma\ln\left[-\theta\ln X\right]$ is Gumbel distributed with location parameter μ and scale parameter σ.

Proof: Let $v = \mu - \sigma\ln\left[-\theta\ln x\right]$. Then $x = \exp\left[-\frac{1}{\theta}\exp\left(\frac{v-\mu}{\sigma}\right)\right]$.

$$\frac{dx}{dv} = \frac{1}{\theta\sigma}\exp\left(\frac{v-\mu}{\sigma}\right)\exp\left[-\frac{1}{\theta}\exp\left(\frac{v-\mu}{\sigma}\right)\right].$$

Hence, V has the pdf

$$g_4(v) = f\left(\exp\left[-\frac{1}{\theta}\exp\left(\frac{v-\mu}{\sigma}\right)\right]\right)\left|\frac{dx}{dv}\right|$$

$$= \frac{1}{\sigma}\exp\left(-\left(\frac{v-\mu}{\sigma}\right)+\exp\left[-\frac{1}{\theta}\exp\left(\frac{v-\mu}{\sigma}\right)\right]\right), -\infty < v < \infty, -\infty < \mu < \infty, \sigma > 0,$$

which is the pdf of the Gumbel distribution.

Proposition 2.5: If $Z = \mu - \sigma\ln\left[X^{-\theta} - 1\right]$, then Z follows the logistic distribution with location parameter μ and scale parameter σ.

Proof: Let $z = \mu - \sigma\ln\left[x^{-\theta} - 1\right]$. Then $x = \left[1+\exp\left(-\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\theta}}$ and

$$\frac{dx}{dz} = \frac{1}{\theta\sigma}\exp\left(-\frac{z-\mu}{\sigma}\right)\left[1+\exp\left(-\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\theta}-1}.$$

The pdf of Z is

$$g_5(v) = f\left(\left[1+\exp\left(-\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\theta}}\right)\left|\frac{dx}{dz}\right|$$

$$= \frac{\exp\left(-\frac{z-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(-\frac{z-\mu}{\sigma}\right)\right]^2}, -\infty < z < \infty, -\infty < \mu < \infty, \sigma > 0.$$

This ends the proof.

Proposition 2.6: Given that $S = \alpha\left[X^{-\theta} - 1\right]^{-\frac{1}{\beta}}$, then S is a log-logistic random variable.

Proof: Let $s = \alpha\left[x^{-\theta} - 1\right]^{-\frac{1}{\beta}}$,. Then $x = \alpha\left[\frac{s^{\beta}}{\alpha^{\beta} + s^{\beta}} \right]^{\frac{1}{\theta}}$. Consequently,

$$\frac{dx}{ds} = \frac{\beta}{\theta} s^{\frac{\beta}{\theta}-1}\left[\alpha^{\beta} + s^{\beta}\right]^{-\frac{1}{\theta}-1}\alpha^{\beta}.$$

Thus, the pdf of S is

$$g_5(v) = f\left(\alpha s^{\frac{\beta}{\theta}}\left[\alpha^{\beta} + s^{\beta}\right]^{-\frac{1}{\theta}} \right)\left|\frac{dx}{ds}\right|$$

$$= \left(\frac{\beta}{\alpha}\right)\left(\frac{s}{\alpha}\right)^{\beta-1}\left(1+\left(\frac{s}{\alpha}\right)^{\beta}\right)^{-2}, \; s>0, \alpha > 0, \beta > 0,$$

indicating that S has a log-logistic distribution with scale parameter α and shape parameter β .

3 Order Statistics and Related Concepts

Order statistics are important in both parametric and nonparametric inferential procedures. Let $X_{(1)}, X_{(2)}, ..., X_{(n)}$ be order statistics corresponding to a random sample of size n $X_1, X_2, ..., X_n$ from a Beta(θ,1) distribution. The pdf of the kth order statistic is

$$f_{X_{(k)}}(x) = \frac{n!}{(n-k)!(k-1)!} f(x) F^{k-1}(x)\left(1 - F(x)\right)^{n-k}$$

$$f_{X_{(k)}}(x) = \frac{n!\theta}{(n-k)!(k-1)!} x^{k\theta-1}\left(1 - x^{\theta}\right)^{n-k}. \qquad (4)$$

In particular, the pdf of the first order statistic $X_{(1)}$ is obtained by substituting k=1 into (3). Thus, we have

$$f_{X_{(1)}}(x) = n\theta x^{\theta-1}\left(1 - x^{\theta}\right)^{n-1}. \qquad (5)$$

If we substitute k=n into (3), then the pdf of nth order statistic $X_{(n)}$ becomes

$$f_{X_{(n)}}(x) = n\theta x^{n\theta-1}. \qquad (6)$$

In accordance with (4) and (5), $X_{(1)}$ and $X_{(n)}$ have the Kumaraswamy distribution and Beta(θ,1) distribution respectively. The rth raw moment of the kth order statistic based on Beta(θ,1) distribution is given by

$$E\left(X_{(k)}^{\,r}\right) = \frac{n!\theta}{(n-k)!(k-1)!}\int_0^1 x^{k\theta+r-1}\left(1-x^{\theta}\right)^{n-k}\,dx.$$

Let $y = x^{\theta}$. Then

$$E\left(X_{(k)}^{\,r}\right) = \frac{n!}{(n-k)!(k-1)!}\int_0^1 y^{k+\frac{r}{\theta}-1}\left(1-y\right)^{n-k}\,dy$$

$$= \frac{\Gamma(n+1)\Gamma\left(k+\frac{r}{\theta}\right)}{\Gamma(k)\Gamma\left(n+\frac{r}{\theta}+1\right)}. \qquad (7)$$

The joint pdf of $X_{(1)}$ and $X_{(n)}$ based on random sample from a continuous distribution is known and can be used to find the distribution of the sample range R. It is given by

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)f(x)f(y)\left(F(y)-F(x)\right)^{n-2}.$$

In the case of Beta(θ,1) distribution,

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)\theta^2 x^{\theta-1}y^{\theta-1}\left(y^{\theta}-x^{\theta}\right)^{n-2},\ 0<x<y<1.$$

To derive the pdf of the sample range $R = X_{(n)} - X_{(1)}$, let $r = x_{(n)} - x_{(1)}$ and $q = x_{(n)} = h_2^{-1}(r,q)$. Then

$x_{(1)} = q - r = h_1^{-1}(r,q)$. The Jacobian of the transformation is

$$J=\begin{vmatrix} \frac{dh_1^{-1}(r,q)}{dr} & \frac{dh_1^{-1}(r,q)}{dq} \\ \frac{dh_2^{-1}(r,q)}{dr} & \frac{dh_2^{-1}(r,q)}{dq} \end{vmatrix} = -1.$$

The joint pdf of R and Q is

$$f_{R,Q}(x,y) = f_{X_{(1)},X_{(n)}}(q-r,q)|J|.$$

$$= n(n-1)\theta^2(q-r)^{\theta-1}q^{\theta-1}\left(q^{\theta}-(q-r)^{\theta}\right)^{n-2},\ 0<r<q<1.$$

$$= n(n-1)\theta^2 q^{n\theta-2}\left(1-\frac{r}{q}\right)^{\theta-1}\left[1-\left(1-\frac{r}{q}\right)^{\theta-1}\right]^{n-2}$$

$$= n(n-1)\theta^2 q^{n\theta-2}\sum_{i=0}^{n-2}\binom{n-2}{i}(-1)^i\left(1-\frac{r}{q}\right)^{\theta(i+1)-1}.$$

Let $u=\frac{r}{q}$. The marginal distribution of R is

$$f_R(r) = n(n-1)\theta^2 r^{n\theta-1}\sum_{i=0}^{n-2}\binom{n-2}{i}(-1)^i\int_r^1 u^{(1-n\theta)-1}(1-u)^{\theta(i+1)-1}\,du.$$

In terms of the beta and incomplete beta functions, we have

$$f_R(r) = n(n-1)\theta^2 r^{n\theta-1}\sum_{i=0}^{n-2}\binom{n-2}{i}(-1)^i\left[B\left((1-n\theta),\theta(i+1)\right)-B\left(r,(1-n\theta),\theta(i+1)\right)\right], 0<r<1, \quad (8)$$

Where the beta function $B\left((1-n\theta),\theta(i+1)\right)$ and incomplete beta function $B\left(r,(1-n\theta),\theta(i+1)\right)$ are defined as

$$B\left((1-n\theta),\theta(i+1)\right)=\int_0^1 u^{(1-n\theta)-1}\left(1-u\right)^{\theta(i+1)-1}du \quad \text{and}$$

$$B\left(r,(1-n\theta),\theta(i+1)\right)=\int_0^r u^{(1-n\theta)-1}\left(1-u\right)^{\theta(i+1)-1}du.$$

4 Correlation between Variate-Values and Ranks in a Sample from the Beta(θ,1) distribution

Many nonparametric statistical methods are predicated on the ranks of the observations that make up a data set instead of the actual data. Hence, it is worthwhile to examine the direction and magnitude of the association between variables and ranks (Zhang, 2021). The coefficient of correlation is useful in determining the amount of information that is lost by using ranks instead of actual observations to make inference (Maturi and Elsayigh, 2009). The correlation between the variate-values X_i and their ranks R_i, $\rho_{Xi,Ri}$ is defined as (Stuart, 1954)

$$\rho_{Xi,Ri}=\left[\frac{12(n-1)}{(n+1)Var(X)}\right]^{\frac{1}{2}}\left[E(XF(X))-\frac{1}{2}E(X)\right] \tag{9}$$

Considering the Beta(θ,1) distribution, we have

$$E(X)=\theta\int_0^1 x^{\theta}dx=\frac{\theta}{\theta+1}, \tag{10}$$

$$Var(X)=E(X^2)-\left(E(X)\right)^2=\theta\int_0^1 x^{\theta+1}dx-\left(E(X)\right)^2$$

$$=\frac{\theta}{\theta+2}-\left(\frac{\theta}{\theta+1}\right)^2=\frac{\theta}{(\theta+2)(\theta+1)^2} \tag{11}$$

and

$$E(XF(X))=\theta\int_0^1 x^{2\theta}dx=\frac{\theta}{2\theta+1}. \tag{12}$$

Substituting (9), (10) and (11) into (8), the correlation between the variate-values X_i and their ranks R_i based on a sample from a Beta$(\theta,1)$ distribution becomes

$$\rho_{Xi,Ri} = \left[\frac{3\theta(n-1)(\theta+2)}{(n+1)(2\theta+1)^2}\right]^{\frac{1}{2}},\ \theta > 0. \qquad (13)$$

Using the Mathematica, the minimum value and maximum values of $\rho_{Xi,Ri}$ in (13) are found to be 0 and 0.9992 respectively. That is $0 \le \rho_{Xi,Ri} \le 0.9992$.

5 Stress-Strength Reliability

In reliability, the stress-strength model describes the life of a component which has a random strength X which is subjected to random stress Y. Dhanya and Jeevavand (2018) considered the estimation of stress-strength reliability parameter based on the power function distribution, when the data on strength is record values. Let X and Y be two independent random variables such that $X \sim \text{Beta}(\theta_1,1)$ and $Y \sim \text{Beta}(\theta_2,1)$. Suppose that the pdf and cdf of X and Y are $f_X(x)$ and $F_Y(y)$ respectively. The stress-strength reliability parameter is

$$R^* = \int_0^1 f_X(x)F_Y(x)dx = \theta_1\int_0^1 x^{\theta_1+\theta_2-1}dx = \frac{\theta_1}{\theta_1+\theta_2}. \qquad (14)$$

Consequently, R^* is a function of θ_1 and θ_2. Hence, with maximum likelihood estimates of θ_1 and θ_2, one can obtain the maximum likelihood estimate of R^*. Let $x_1, x_2, ..., x_n$ denote n sample observations from a $\text{Beta}(\theta_1,1)$ distribution and $y_1, y_2, ..., y_n$ denote n sample observations from a $\text{Beta}(\theta_2,1)$ distribution. Then the likelihood function is

$$L = \prod_{i=1}^{n}\left[\theta_1 x_i^{\theta_1-1}\right]\prod_{j=1}^{m}\left[\theta_2 y_j^{\theta_2-1}\right].$$

The associated log-likelihood function is

$$\ln L = n\ln(\theta_1) + (\theta_1-1)\sum_{i=1}^{n}\ln(x_i) + m\ln(\theta_2) + (\theta_2-1)\sum_{j=1}^{m}\ln(y_i).$$

Partial derivative of $\ln L$ with respect to each θ_1 and θ_2 are given below:

$$\frac{d\ln L}{d\theta_1} = \frac{n}{\theta_1} + \sum_{i=1}^{n} \ln(x_i)$$

$$\frac{d\ln L}{d\theta_2} = \frac{m}{\theta_2} + \sum_{j=1}^{m} \ln(y_i)$$

Solving the equations $\frac{d\ln L}{d\theta_1} = 0$ and $\frac{d\ln L}{d\theta_1} = 0$ simultaneously for the maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ_1 and θ_2 respectively, we obtain $\hat{\theta}_1 = -\frac{n}{\sum_{i=1}^{n} \ln(x_i)}$ and

$$\hat{\theta}_2 = -\frac{m}{\sum_{j=1}^{n} \ln(y_i)}.$$

Hence, the maximum likelihood estimate of R^* is $R^* = \frac{\hat{\theta}_1}{\hat{\theta}_1 + \hat{\theta}_2}$.

In a simulation exercise to investigate the performance of the maximum likelihood estimate of R^*, 1000 samples are drawn independently from the $Beta(\theta_1, 1)$ and $Beta(\theta_2, 1)$ distributions based on the parameter values $(\theta_1, \theta_2) : (0.5,1), (1,0.5), (1,1.5)$ and the sample sizes $(n, m) : (10,10), (20,20), (30,30), (100,100)$. Table 1 contains the simulation results., which are the various maximum likelihood estimates of R^*, average biases (ABs) and mean squared errors (MSEs). As the sample size increases, the average bias and mean squared error tend to zero while the average maximum likelihood estimates (AMLEs) of R^* tends to its true value.

Table 1: Average MLE of R^* and the Associated Average Bias and Mean Squared Error

$\theta_1 = 0.5,\ \theta_2 = 1$ and R=0.3333			
(n,m)	AMLE	AB	MSE
(10,10)	0.3425	0.0092	0.0845
(20,20)	0.3409	0.0075	0.0565
(30,30)	0.3352	0.0019	0.0036
(100,100)	0.3328	-0.0005	0.0003
$\theta_1 = 1,\ \theta_2 = 0.5$ and R=0.6667			
(10,10)	0.6609	-0.0057	0.0329
(20,20)	0.6633	-0.0033	0.0101
(30,30)	0.6647	-0.0019	0.0036
(100,100)	0.6650	-0.0017	0.0029
$\theta_1 = 3,\ \theta_2 = 2$ and R=0.6000			
(10,10)	0.5969	-0.0031	0.0098
(20,20)	0.5982	-0.0018	0.0033
(30,30)	0.5989	-0.0011	0.0011
(100,100)	0.6001	0.0001	2.0294×10^{-5}

6 Application of $\text{Beta}(\theta,1)$ Distribution in Bayesian Inference

Bayesian point estimation of a parameter of a probability distribution is essential in solving many statistical problems. This is because Bayes estimator is often a better alternative to the maximum likelihood estimator, among other classical estimators (Supharaknosakun, 2021). Studies on the Bayesian approach to estimating the parameter of a $\text{Beta}(\theta,1)$ distribution has been considered in the past (Rahman et al., 2012; Hanif et al., 2015). In this section, we focus on the role of the distribution as a prior distribution in the Bayesian estimation of the parameter of the discrete Ailmujia distribution. Statistically speaking, a random variable X is said to follow the discrete Ailmujia distribution with parameter η if its probability mass function is (Ahmad et al., 2022)

$$p(x)=\frac{(1-\eta)^2}{\eta}x\eta^x, 1,2,3,\ldots \quad (14)$$

Some properties of the distribution were derived by the author. Specifically, the maximum likelihood estimate of the parameter of the distribution is

$$\eta=\frac{1-\bar{x}}{1+\bar{x}}, \quad (15)$$

which is also equal to the associated method of moments estimator of η.

Let $L(\eta)$ be the likelihood function corresponding to the discrete Ailamujia distribution. This implies that

$$L(\mathbf{x}|\eta)=\frac{(1-\eta)^{2n}}{\eta^{n}}\left(\prod_{i=1}^{n}x_i\right)\eta^{\sum_{=1}^{n}x_i}. \tag{16}$$

If the prior pdf of η is the $\text{Beta}(\eta,1)$ distribution with pdf

$$\pi(\eta)=\theta\eta^{\theta-1},\quad \eta>0,\theta>0. \tag{17}$$

The posterior distribution is given by

$$\pi(\eta|\mathbf{x})=\frac{L(\mathbf{x}|\eta)\pi(\eta)}{\int_0^1 L(\eta)\pi(\eta)d\eta}=\frac{\frac{(1-\eta)^{2n}}{\eta^{n}}\left(\prod_{i=1}^{n}x_i\right)\eta^{\sum_{=1}^{n}x_i}\theta\eta^{\theta-1}}{\int_0^1\frac{(1-\eta)^{2n}}{\eta^{n}}\left(\prod_{i=1}^{n}x_i\right)\eta^{\sum_{=1}^{n}x_i}\theta\eta^{\theta-1}d\eta}$$

$$=\frac{\eta^{\sum_{=1}^{n}x_i+\theta-n-1}(1-\eta)^{2n}}{\int_0^1\eta^{\sum_{=1}^{n}x_i+\theta-n-1}(1-\eta)^{2n}d\eta}=\frac{\Gamma\left(\theta+n+\sum_{=1}^{n}x_i+1\right)\eta^{\sum_{=1}^{n}x_i+\theta-n-1}(1-\eta)^{2n}}{\Gamma\left(\theta-n+\sum_{=1}^{n}x_i\right)\Gamma(2n+1)}. \tag{18}$$

Therefore, $\pi(\eta|\mathbf{x})$ □ $\text{Beta}\left(\sum_{=1}^{n}x_i+\theta-n,2n+1\right)$.

Under the squared loss function, the Bayes estimate t of η is the posterior mean defined by

$$t=E(\eta|\mathbf{x})=\frac{\Gamma\left(\sum_{=1}^{n}x_i+\theta+n+1\right)}{\Gamma\left(\sum_{=1}^{n}x_i+\theta-n\right)\Gamma(2n+1)}\int_0^1\eta^{\sum_{=1}^{n}x_i+\theta-n}(1-\eta)^{2n}\,d\eta.$$

$$= \frac{\Gamma\left(\sum_{=1}^{n} x_i + \theta + n + 1\right)\Gamma\left(\sum_{=1}^{n} x_i + \theta - n + 1\right)\Gamma(2n+1)}{\Gamma\left(\sum_{=1}^{n} x_i + \theta - n\right)\Gamma\left(\sum_{=1}^{n} x_i + \theta + n + 2\right)\Gamma(2n+1)}$$

$$= \frac{\sum_{=1}^{n} x_i + \theta - n}{\sum_{=1}^{n} x_i + \theta + n + 1}. \qquad (19)$$

In accordance with the quadratic loss function $L(t^*, \eta) = \left(\frac{t^* - \eta}{\eta}\right)^2$, the Bayes estimate t^* of η is obtained by solving the equation

$$\frac{\partial}{\partial t^*}\int_0^1 \left(\frac{t^* - \eta}{\eta}\right)^2 \pi(\eta|\mathbf{x})d\eta = 0.$$

That is

$$\frac{\partial}{\partial t^*}\int_0^1 \left(\frac{t^* - \eta}{\eta}\right)^2 \frac{\Gamma\left(\theta + n + \sum_{=1}^{n} x_i + 1\right)\eta^{\sum_{=1}^{n} x_i + \theta - n - 1}(1-\eta)^{2n}}{\Gamma\left(\theta - n + \sum_{=1}^{n} x_i\right)\Gamma(2n+1)} d\eta = 0. \qquad (20)$$

If we solve for t^* in (20), we have

$$t^* = \frac{\sum_{=1}^{n} x_i + \theta - n\text{-}2}{\sum_{=1}^{n} x_i + \theta + n\text{-}1}.$$

7 Conclusion

We have derived some results corresponding to a special case of beta distribution. In particular, the relationships between the distribution and exponential, Weibull, Gumbel, log-logistic, Pareto and logistic distributions are theoretically established. This simply indicates that it is possible to transform a random variable that follows the distribution under consideration into

any of the exponential, Weibull, Gumbel, log-logistic, Pareto and logistic distributed random variables. The consistency property of the maximum likelihood estimate of the stress-strength parameter based on the distribution is illustrated through a simulation procedure. The coefficient of correlation between the variate-values X_i and their ranks R_i associated with a random sample from the distribution is derived and shown to be bounded below by 0 and above by 0.9992. Furthermore, Bayesian estimation of the parameter of the discrete Ailamujia distribution using the distribution as prior distribution leads to a posterior distribution that is a beta distribution. Thus, beta distribution is a conjugate prior for the parameter of the discrete Ailamujia distribution.

References

Ahmad, A., Ahmad, A., & Ozel, G. (2022). One parameter discrete Ailamujia distribution with statistical properties and biodiversity and abundance data applications. *Punjab University Journal of Mathematics, 54*(2), 111-125.

Dhanya, M., & Jeevavand, E. S. (2018). Stress-Strength reliability of power function distribution based on records. *Journal of Statistics Applications and Probability*, *7*(1), 39-48.

Hanif, S., Al-Ghamdi, S. D., Khan, K., & Shahbaz, M. Q. (2015). Bayesian estimation for parameters of power function distribution under various priors. *Mathematical Theory and Modeling*, *5*(13), 106—112.

Maturi, T. A., & Elsayigh, A. (2009). The correlation between variate-values and ranks in samples from complete fourth power exponential distribution. *Journal of Mathematics Research*, 1(1), 14 -18.

Rahman, H., Roy, M. K., & Baizid, A. R. (2012). Bayes estimation under conjugate prior for the case of power function distribution. *American Journal of Mathematics and Statistics*, *2*(3), 44-48.

Stuart, A. (1954). The correlation between variate-values and ranks in samples from a continuous distribution. *British Journal of Statistical Psychology*, *7*, 37–44.

Supharakonsakun, Y. (2021). Bayesian approaches for Poisson distribution parameter estimation. *Emerging Science Journal*, *5*(5), 755 – 774.

Zhang, C. (2021). Further examples related to correlations between variables and ranks. *The American Statistician*, *75*(2), 226–229.

On the Prediction Variance Performance of Replicated Minimum-Run Resolution V Designs

[1,*]F. I. Nsude, [2]Elem Uche and [1]E.C. Ukaegbu

[1]State University of Medical and Applied Sciences, Igbo-Eno, Enugu State

[2]Department of Maths/Statistics, Akanu Ibiam Federal Polytechnic Unwana

****Corresponding Author's email: ijelink@gmail.com***

Abstract

The Minimum-run resolution V (MinResV) central composite designs used in response surface exploration offers smaller number of experimental runs when compared to the standard central composite design (CCD) which is an advantage when considering the cost of experimentation. However, the evaluation of the MinResV design has focused on replicating only the centre point for error estimation. Replicating the other portions of the design and the implications have not yet been considered in previous related studies. The cube and star portions of the Minimum-run resolution V (MinResV) designs have been replicated, evaluated and compared using the A-, D-, G- and I-optimality criteria. The stability and prediction capabilities of the designs generated by partial replication of the portions of the MinResV designs are tracked using the fraction of design space (FDS) graph. In this study, these graphs were plotted for the scaled and unscaled prediction variances of the partially replicated variations of the MinResV designs in the spherical regions. For the $k = 6, 7, 8, 9$ and 10 factors considered in this study, the MinResV CCDs with star-replicated designs performed better than the cube-replicated designs in terms of minimum spread of scaled and unscaled prediction variances.

1. Introduction

Minimum-Run resolution V (MinResV) designs are equi-replicated two-level irregular fractions of resolution V, often used for screening and estimations. Oehlert and Whitcomb (2002) developed this class of response surface designs as a smaller alternative to standard Central Composite designs (CCD). Generally, CCDs are known to consist of regular fractions (called the "cube") of at least resolution V, axial runs (called the "star"), and centre runs. In the construction of MinResV composite designs, the regular fraction or cube portion of the CCD is replaced by the MinResV fraction but still consists of the same number of star runs and centre

points: see, for example, Li et al (2009). This design usually has smaller number of experimental runs when compared to the standard CCD especially as the number of factors, k, increases.

The design matrix, X, of the MinResV design for six factors with one centre point is given as

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \left[\overbrace{\begin{array}{cccccccccccccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{array}}^{min\,resV fraction} \overbrace{\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -\alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\alpha & \alpha & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\alpha & \alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\alpha & \alpha \end{array}}^{starrun} \overbrace{\begin{array}{c} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array}}^{centrerun} \right]$$

Traditionally, only the centre point of the CCD is replicated n_0 times for the estimation of experimental error. However, in reality, by replicating only the centre point, variability may not be adequately accounted for since variability is not expected to be constant throughout the design region. As Dykstra (1960) rightly pointed out, if variability increases elsewhere aside the centre of the design region, the associated experimental error may be too small to properly evaluate the model parameters. On this note, some authors have suggested replication at other portions of the design to adequately estimate the experimental error: see, for example, Giovannitti-Jensen and Myers (1989).

In this study, we focus on partially replicating the cube and star portions of the MinResV CCD for error estimation. The objective for the partial replication of the portions of this design in this study is to assess the behaviour of the design's scaled and unscaled prediction variances as well as the prediction capabilities when other portions of the designs except the centre point are replicated. To achieve this, the maximum and average points of the scaled prediction variances of the designs were studied using the *G*- and *I*-optimality criteria while the fraction of design space (FDS) graphs were used to track the scaled and unscaled prediction variances throughout the design space.

2. Literature Review

In this section, we start by introducing the model that is used to describe the relationship between the variables under study. Generally, the response surface model of equation (1) is used to describe the relationship between the response variable, y, and the k prediction variables. The model is given by

$$Y = X'\beta + e, \qquad (1)$$

where $Y = N \times 1$ vector of responses, $X = N \times p$ design matrix expanded to model form and which is derived from the $N \times k$ design matrix, β = vector of the unknown p model parameters, and e = random error associated with y; e follows the normal distribution with zero mean and variance, σ^2. Appropriately, each row of X represents an experimental observation involving the different levels of the factors such that the total number of rows accounts for the total number of design runs, N, while each column of X represents a model parameter such that the total number of columns gives the total number of model parameters, p, to be estimated.

However, in many experimental situations, the relationship between the response and predictor variables is more adequately described by a second-order response surface model (see Ukaegbu and Chigbu, 2017),

$$y(x) = \beta_0 + x'\beta + x'Bx + e. \tag{2}$$

Here, x = point in the design space spanned by the design, and $B = k \times k$ matrix with the diagonal elements being the pure quadratic terms and the off-diagonal elements being the mixed quadratic or interaction terms.

Now, every point, x, in the design space has a prediction variance given by

$$Var[\hat{y}(x)] = \sigma^2 x'(X'X)^{-1}x. \tag{3}$$

$x' = [1, x_1, x_2, \ldots, x_k; x_1^2, \ldots, x_k^2; x_1x_2, \ldots, x_{k-1}x_k]$ is the vector of points in the design space expanded to model form. The scaled prediction variance (SPV) is obtained by multiplied by equation (3) by N is and divided by the process variance, σ^2. That is,

$$\frac{Nvar[\hat{y}(x)]}{\sigma^2} = Nx'(X'X)^{-1}x. \tag{4}$$

It is known that the scaling is very useful when comparing competing designs of various sizes and it penalizes larger designs with respect to the smaller ones. For further information on scaling of prediction variance, see Anderson-Cook et al. (2009), Li et al. (2009) and Umegwuagu et al. (2020).

Some practitioners prefer the standardized or unscaled prediction variance (UPV) given in equation (5) for design evaluation and comparison. The UPV is

$$\frac{var[\hat{y}(x)]}{\sigma^2} = x'(X'X)^{-1}x. \quad (5)$$

In this case, the quality of the designs under comparison is not considered as a function of cost, rather, the UPV is useful in comparing designs of different sizes in other to know if the variance of the predicted response is substantially reduced by increasing the number of runs of larger designs. Piepel (2009) and Goos (2009) presented further justification the use of the UPV in design evaluation and comparison.

In this study, we explored the prediction capabilities of variations of the partially replicated MinResV CCDs by tracking the qualities of the SPV and UPV in spherical region. The number of factors under consideration here is $k = 6$, 7, 8, 9 and 10.

3. Methods

3.1 Optimality Criteria

Two optimality criteria that are directly related to the measure of the scaled prediction variance of the design are used to study the maximum and average measures of the SPV. They are the *G*- and *I*-optimality criteria.

G-Optimality

The *G*-optimality criterion is known to minimize the maximum SPV of a design over the region of interest. Symbolically, the *G*-optimality is expressed as

$$G = minmax\left\{\frac{Nvar[\hat{y}(x)]}{\sigma^2}\right\}. \quad (6)$$

I-Optimality

The *I*-optimality is also known as the *V*-optimality or *IV*-criterion and minimizes the normalized average integrated SPV in the region, *R*, of interest. The *I*-optimality is defined as

$$I = min\frac{1}{k}\int_R V[\hat{y}(x)]dx. \quad (7)$$

3.2 Fraction of design space plots

The fraction of design space (FDS) graph, developed by Zahran et al (2003), is used as tool for evaluating the spread of the prediction variances of any design throughout the design region.

This graphical technique has been used extensively in robust design evaluations and designs with mixture components, see for example, Ozol-Godfrey (2004), Anderson-Cook, Borror and Montgomery (2009) and Jang and Anderson-Cook (2011). The stability and prediction capability of a design is measured by the flatness of the graph so that, a design with flatter graphs has stronger stability and higher prediction capability.

3.3 Design replication

In this study, the cube was replicated n_c times and the star was replicated n_s, such that the MinResV CCD uses a total of $N = n_c f + 2n_s k + n_0$ observations or runs for model parameter estimation. The pattern of replication is that, if the cube (C) is replicated, then the star (S) is not replicated and vice versa. In each case, the number of centre points used is $n_0 = 1$. Therefore, the first design variation is C_2S_1, in which there is the replication of the cube portion twice and there is no replication of the star portion. The second design variation is C_1S_2, in which there is the replication of the star portion twice but there is no replication of the cube portion. The other design variations include C_3S_1, C_1S_3, C_4S_1 and C_1S_4. These design variations were compared with the standard MinResV CCD, C_1S_1, where the cube and star portions are not replicated. The replication procedure adopted here is an extension of Draper (1982). For other forms of replication of the CCD and mixed resolution designs, see Dykstra (1960), Borkowski (1995), Borkowski and Lucas (1997), Borkowski and Valeroso (2001).

4. Results and Discussion

In this section, we evaluate the prediction variances of the seven variations of the MinResV designs and compare the results using the two optimality criteria and FDS plots. Statistical packages JMP (version 12) was used to obtain the values of the optimality criteria and the FDS graphs.

4.1 Comparison using I-optimality

The results of the *I*-optimality were presented in Table 1. For $k = 6$, 7, 8, 9 and 10 factors, the star-replicated MinResV CCD gave smaller values of *I*-optimalilty when compared with other design variations. The MinResV CCD with higher replication of star portion, C_1S_3 and C_1S_4 have *I*-optimal values that compete favourably for all the factors under consideration. Though C_1S_4 have slightly smaller *I*-optimal values but C_1S_3 compensated for this with smaller number

of runs. The replicated cube designs and C_1S_1 have poor I-optimal values when compared with the star designs. The *I*-optimal values of these designs get worse as the replication and number of runs increase, making these designs undesirable.

Table 1: Summary Statistics for the Design Optimality Criteria

k	**Design**	**N**	***I*-Optimality**	***G*-Optimality**	***k***	**Design**	**N**	***I*-Optimality**	***G*-Optimality**
6	C_1S_1	35	12.16	17.25	9	C_1S_1	65	13.07	15.11
	C_2S_1	57	17.48	23.48		C_2S_1	111	20.44	22.67
	C_1S_2	47	10.20	15.08		C_1S_2	83	10.02	12.67
	C_3S_1	79	22.25	28.61		C_3S_1	157	28.17	30.67
	C_1S_3	59	9.35	14.02		C_1S_3	101	9.21	12.17
	C_4S_1	101	27.02	33.68		C_4S_1	203	36.32	39.85
	C_1S_4	71	**9.07**	**13.81**		C_1S_4	119	**8.81**	**12.15**
7	C_1S_1	45	10.77	12.49	10	C_1S_1	77	14.99	17.79
	C_2S_1	75	16.28	18.03		C_2S_1	133	23.88	27.25
	C_1S_2	59	8.68	10.77		C_1S_2	97	11.02	13.84
	C_3S_1	105	22.78	25.51		C_3S_1	189	32.19	35.49
	C_1S_3	73	7.90	**10.21**		C_1S_3	117	9.56	12.55
	C_4S_1	135	29.20	32.99		C_4S_1	245	40.73	44.31
	C_1S_4	87	**7.70**	10.40		C_1S_4	137	**8.94**	**12.22**
8	C_1S_1	55	12.57	15.20					
	C_2S_1	93	18.91	21.45					
	C_1S_2	71	9.57	12.11					
	C_3S_1	131	25.61	28.33					
	C_1S_3	87	8.77	11.67					
	C_4S_1	169	32.99	36.64					
	C_1S_4	103	**8.39**	**11.60**					

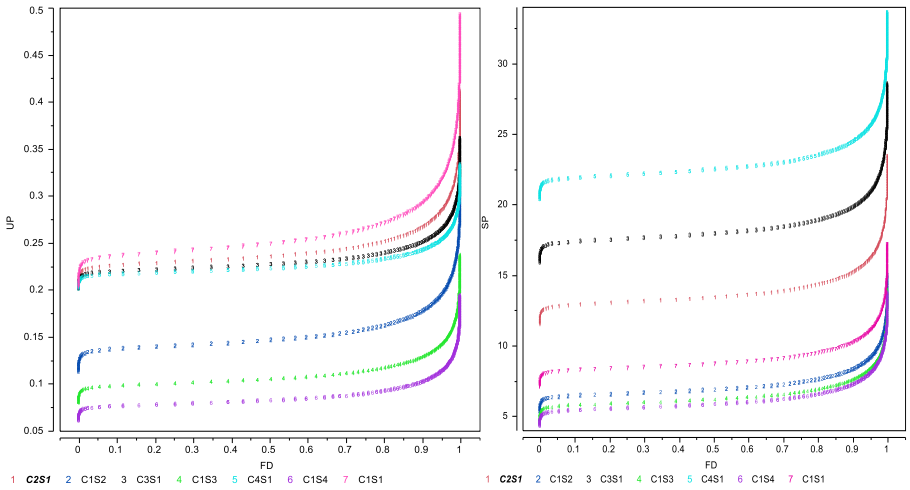
4.2 Comparison with G-optimality Criterion

The results for the *G*-optimality criterion were shown in Table 1. It could be observed that the three MinResV CCD with replication of the star portion have smaller values of the *G*-optimality criterion than the other designs variations. Although these star-replicated MinResV CCDs compete favourably, C_1S_4 has the smallest *G*-optimal values for all the factors being studied. The replicated cube designs and C_1S_1 behave in similar manner with respect to the *I*-optimality criterion, with the *G*-optimal values increasing as the number of runs increases.

4.3 Comparison using fraction of design space plots

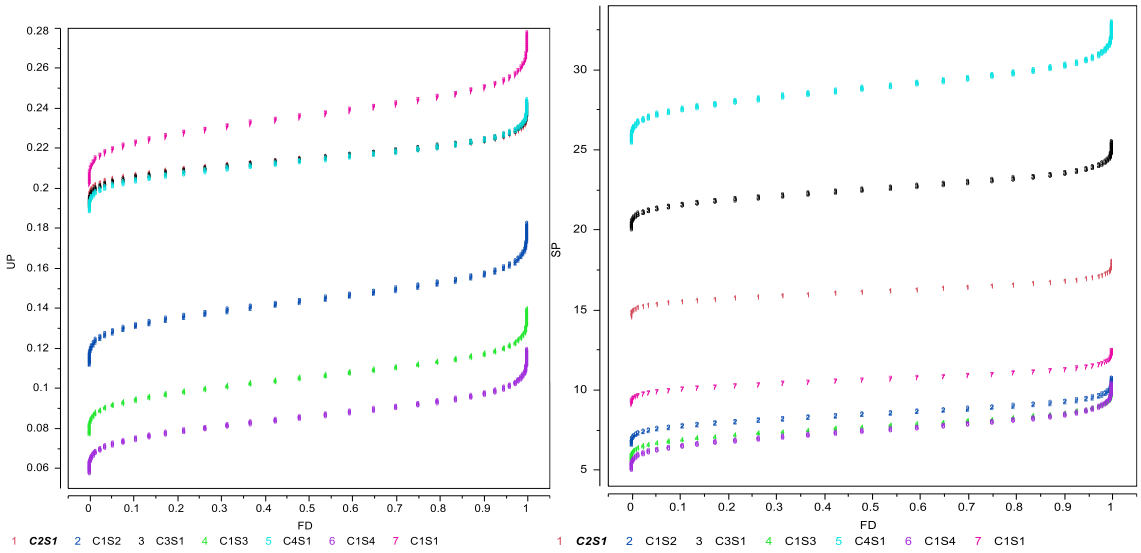
The FDS plots for the unscaled and scaled prediction variances are displayed in Figures 1, 2, 3, 4 and 5 for 6, 7, 8, 9 and 10 factors, respectively. For all the factors and for both the scaled

and unscaled prediction variances, the FDS plots show that star-replicated MinResV CCD options displayed flatter graphs closer to the horizontal line, which indicates spread of minimum prediction variances for almost all the fractions of the design space. The FDS plots for the SPV of the higher replicated star designs, C_1S_3 and C_1S_4, show that these two designs compete favourably with little dispersion, though C_1S_4 tends to be better with smaller prediction variances.



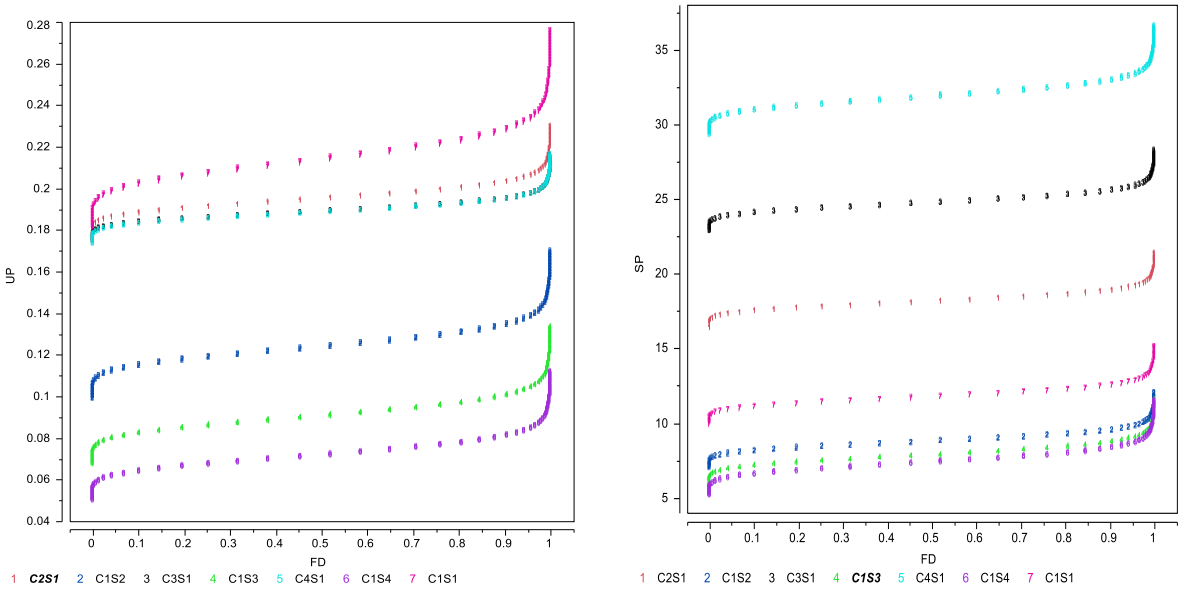
(a) (b)

Figure 1: FDS plots for (a) Unscaled and (b) Scaled Prediction Variances, $k = 6$ factors



(a) (b)

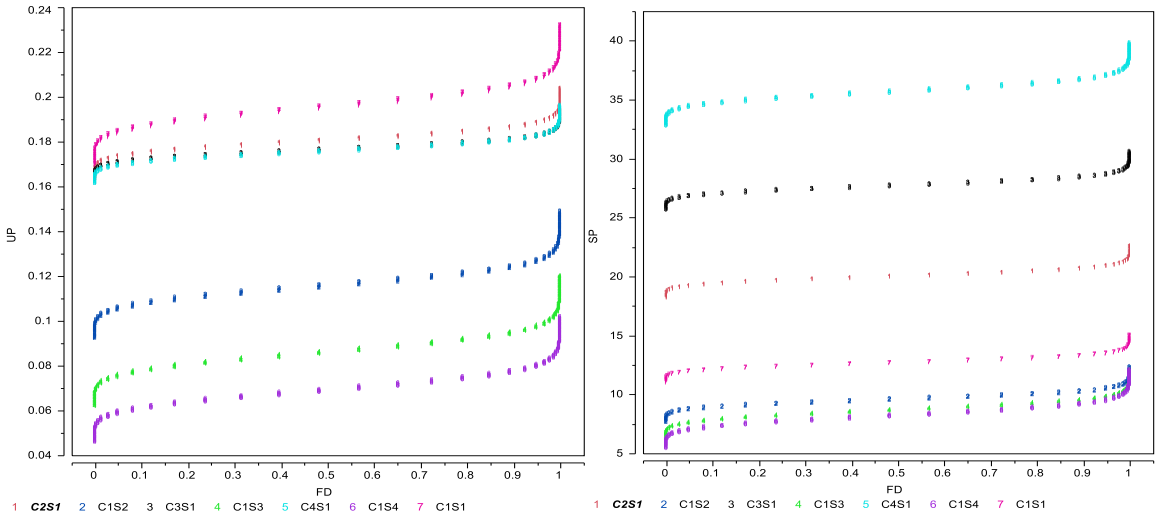
Figure 2: FDS plots for (a) Unscaled and (b) Scaled Prediction Variances, $k = 7$ factors



(a) (b)

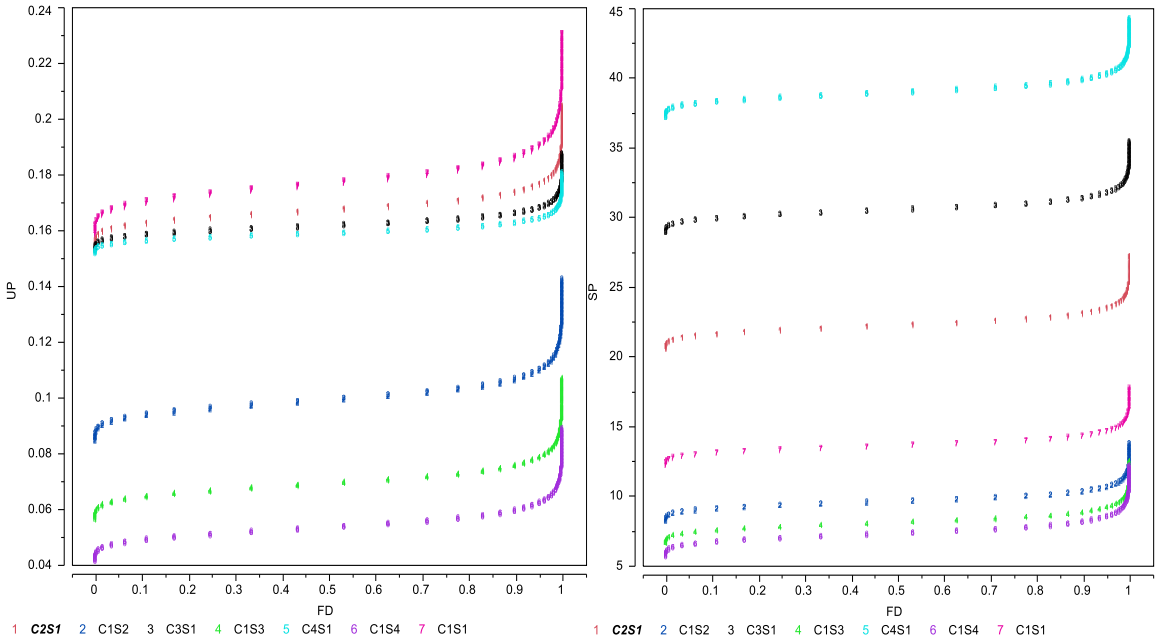
Figure 3: FDS plots for (a) Unscaled and (b) Scaled Prediction Variances, $k = 8$ factors.

On the other hand, the design, C_1S_1, seems to be the worst of the seven under the UPV FDS plots but perform better than the replicated cube designs under the SPV FDS plots. The replicated cube designs display poor prediction capability for all the factors under consideration with high prediction variances for the scaled and unscaled prediction variances. As can be seen from Figures 1, 2, 3, 4 and 5, the higher replicated cube designs, C_3S_1 and C_4S_1, tend to compete favourably for UPV but exhibit serious dispersions for SPV.



(a) (b)

Figure 4: FDS plots for (a) Unscaled and (b) Scaled Prediction Variances, $k = 9$ factors



(a) (b)

Figure 5: FDS plots for (a) Unscaled and (b) Scaled Prediction Variances, $k = 10$ factors

5. Summary and Recommendations

The star-replicated MinResV CCD options performed better than the other designs variations when judged with the *G*- and *I*-optimality criteria. This is also true when evaluated using the FDS plots. The FDS graphs show that for all the factors under study, C_1S_3 and C_1S_4 displayed better stability and higher prediction capabilities than the other designs variations. The design, C_1S_4, is recommended if the practitioner has the resources for the number of observations obtainable for the design. In terms of number of design runs, we recommend C_1S_3 since this design competes well with C_1S_4 for the three conditions under which the designs are judged. On the other hand, replicating the cube portion of the MinResV CCDs for $k = 6$ to 10 factors is not advantageous for all the factors under consideration.

References

Anderson-Cook, C.M., Borror, C.M. and Montgomery, D.C. (2009), Response Surface Design Evaluation and Comparison, *Journal of Statistical Planning and Inference*, vol.139, pp. 629-641.

Borkowski, J.J (1995), Spherical Prediction Properties of Central Composite and Box-Behnken Designs, *Technometrics*, vol. 37 (4), pp. 399-410.

Borkowski, J.J. and Lucas, J.M. (1997), Designs of Mixed Resolution for Process Robustness Studies, *Technometrics*, 39, 1, pp. 63-70.

Borkowski, J.J and Valeroso, E.S. (2001), Comparison of Design Optimality Criteria of Reduced Models for Response Surface Designs in the Hypercube, *Technometrics*, 43 (4), pp. 468 – 477.

Draper, N.R. (1982), Centre Points in Second-Order Response Surface Designs, *Technometrics*, vol. 24 (2), pp. 127-133.

Dykstra, O. Jr. (1960), *Partial* Duplication of Response Surface Designs, *Technometrics*, vol. 2 (2), pp. 185-195.

Giovannitti-Jensen, A. and Myers, R.H. (1989), Graphical Assessment of the Prediction Capability of Response Surface Designs, *Technometrics*, vol. 31 (2), pp. 159-171.

Goos, P. (2009), Discussion of "Response Surface Design Evaluation and Comparison", *Journal of Statistical Planning and Inference*, vol.139, pp. 657-659.

Jang, D.H. and Anderson-Cook, C.M. (2011), Fraction of Design Space Plots for Evaluating Ridge Estimators in Mixture Experiments, *Quality and Reliability International*, 27, pp. 27-34.

Li, J., Liang, L., Borror, C.M., Anderson-Cook, C.M. and Montgomery, D.C. (2009), Graphical Summaries to Compare Prediction Variance Performance for Variations of the Central Composite Design for 6 to 10 Factors, *Quality Technology and Quantitative Management*, vol. 6(4), pp. 433-449.

Oehlert, G. W. and Whitcomb, P. (2002), Small, Efficient, Equireplicated Resolution V Fractions of 2^k Designs and Their Application to the central Composite Designs. 46th Annual Fall Technical Conference of the ASQ and ASA.

Ozol-Godfrey, A. (2004), *Understanding Scaled Prediction Variance Using Graphical Methods for Model Robustness, Measurement Error and Generalized Linear Models for Response Surface designs*, Unpublished PhD Dissertation, Virginia Polytechnic Institute and State University

Piepel, G.F. (2009), Discussion of "Response Surface Design Evaluation and Comparison" by C.M. Anderson-Cook, C.M. Borror and D.C. Montgomery, *Journal of Statistical Planning and Inference*, vol. 139, pp. 653-656.

Ukaegbu, E.C. and Chigbu, P.E. (2017). Modified Exact Single-Value Criteria for Partial Replications of the Central Composite Design, *American Journal of Theoretical and Applied Statistics*, 6(4), 205-208.

Umegwuagu, C.N., Chigbu, P.E. and Ukaegbu, E.C. (2020). Assessment of composite mixed resolution designs in spherical regions, *Journal of the Nigerian Statistical Association*, 32, 29 -38.

Zahran, A., Anderson-Cook, C.M. and Myers, R.H. (2003), Fraction of Design Space to Assess the Prediction Capability of Response Surface Designs, *Journal of Quality Technology*, 35, pp. 377-386.

Comparison of performance metrics in the extension of Artificial Neural Network to relative survival

***Rasheedah SADIQ**

Department of Statistics, University of Abuja, Abuja, Nigeria
National Bureau of Statistics

***Corresponding Author's email: sheedahtwinkled@gmail.com**

Abstract

The predictive ability of a classifier is typically measured using Accuracy (or Misclassification Error Rate (MER), which is 1 minus Accuracy), but in the 1970's, Area under the Reciever Operating Characteristics (ROC) curve simply known as AUC was been proposed as an alternative for evaluating performance. Many papers have been published as to which of them is better in prediction. In this paper, Artificial Neural Network (ANN) was extended to Relative Survival and these performance measures were compared on varied factors of the ANN architecture. The factors were learning algorithms, dataset allocation and hidden nodes. The real-life dataset was obtained from the R package. The Multi-layer Perceptron (MLP) neural network topology was used and analysis was carried out using R statistical software (www.cran-r.org).

1. INTRODUCTION

To select a proper model for the analysis of survival data requires model assumptions, such as proportionality of hazards and independent time of occurrence. If the data is complicated, it may make using the models problematic and restricted (Kutner et al., 2004). One way to tackle such problems is to apply Artificial Neural Network (ANN) models that have been increasingly used in recent decades to survival analysis (De Laurentiis and Ravdin., 1994; Biganzoli et al., 1998; (Akbar Biglarian, et al., 2013; Sharaf and Tsokos., 2015). These models are distribution-free and consequently free of assumptions. Another issue is that survival analysis does not differentiate the effects of prognostic factors on disease-specific mortality from their effects on overall mortality, which may lead to wrong interpretation of results (Quantin C et al., 1999). A limitation of standard survival analysis leading to the use of cause-specific survival analysis. However, in population-based cancer studies, cause of death may either not be available from death certificates or often erroneously recorded (Begg and Schrag, 2002). In addition, the decision of whether a death is due or not due to a particular disease is difficult, opinions may differ and it may be tough to reach an indisputable conclusion, thereby preventing the cause-

specific survival analysis. Therefore, when analyzing survival data in population-based cancer studies, mortality due to other causes is commonly incorporated using "relative survival methods" (Cutler and Aztell, 1969). The application of statistical models for relative survival increased (Dickman *et al*, 2004; Esteve *et al.*,1990; Hakulinen and Tenkanen, 1987; Sasieni., 1996; Bolard *et al*, 2002; Giorgi *et al*, 2003; Lambert *et al.*, 2005). However, all these models are characterized by the earlier stated lapses above, in addition, when interactions exist amongst variables, the relationship of variables to the outcome is complex and unknown, then a non-linear method is necessary for the estimation of Relative Survival, hence ANN (Rasheedah Sadiq., et al 2016). ANN as a classifier, can be evaluated by several performance measures, but in this paper, two were compared, namely AUC and Accuracy. Accuracy is the ratio of correctly predicted observations while AUC measures how true positive rate (sensitivity/recall) and false positive rate (1- specificity) trades off, several papers have compared their evaluation abilities but this paper wants to examine the comparison within the scope of ANN in Relative survival.

2. MATERIALS AND METHODS

We use the dataset that were collected in the study carried out at the University Clinical Centre in Ljublana and contains 1040 Acute Myocardial Infarction patients diagnosed between 1982 and 1986 and followed up until 1997, that was included in the package relsurv in R. The file name is rdata. During this time 547 deaths occurred and as the causes of death are not given, this is a good example of the need for the relative survival methodology. The cases where the event of interest has not occurred due to patient lost to follow up or the patient's death for other reasons or alive as at the end of the study are regarded as censored data. The organization of the data is as follows:

>rdata [1:2,]

	time	cens	age	sex	year	agegr
1	2657	1	68	2	24-Jun-82	62-70
2	1097	1	63	2	31-Aug-82	62-70

Time is measured in days and year of infarction is expressed in R date format. Age is measured in years and a categorical variable agegr containing four age categories ("under 54", "54–61", "62–70", "71–95") is formed.

The censoring indicator is specified in variable cens and is coded 0 (censoring) and 1 (event). The population data is the Slovenian population census table since the study was performed in Slovenia. The data were used to run the
simulation using 5% and 10% hold out samples to validate after training.

3. RESULTS

Descriptive Statistics

Table 1: Cases available in analysis

Events	547	52.6%
Censored	493	47.4%
Total	1040	100.0%

Table 2: Cases available by sex

		Censored	Event	Total
Male	Count	391	360	751
	% within sex	52.1%	47.9%	100.0%
Female	Count	102	187	289
	% within sex	35.3%	64.7%	100.0%
Total	Count	493	547	1040
	% within sex	47.4%	52.6%	100.0%

Table 3: Cases available by age group

		Censored	Event	Total
<54	Count	186	84	270
	% within agegr	68.9%	31.1%	100.0%
54-61	Count	157	99	256
	% within agegr	61.3%	38.7%	100.0%
62-70	Count	100	155	255
	% within agegr	39.2%	60.8%	100.0%
71-95	Count	50	209	259
	%within agegr	19.3%	80.7%	100.0%
Total	Count	493	547	1040

Table 4: Performance measures of the three algorithms for varying hidden nodes at different hold-out samples

		Back-propagation Algorithm			Resilient Back-propagation Algorithm			Scaled Conjugate Algorithm		
		H =1	H =5	H =10	H =1	H =5	H =10	H =1	H =5	H =10
5%	AUC	0.8078	0.8195	0.8135	0.8180	0.8127	0.8054	0.8241	0.8243	0.8189
	ACC	0.8097	0.8190	0.8127	0.8144	0.8125	0.8038	0.8221	0.8202	0.8163
10%	AUC	0.8103	0.8117	0.8062	0.8288	0.8225	0.8172	0.8274	0.8162	0.8239
	ACC	0.8087	0.8103	0.8062	0.8269	0.8207	0.8144	0.8274	0.8168	0.8216
20%	AUC	0.8173	0.8056	0.8061	0.8048	0.8048	0.8047	0.8154	0.8085	0.8137
	ACC	0.8173	0.8052	0.8057	0.8188	0.8076	0.8029	0.8139	0.8075	0.8130

4. DISCUSSION

In this work, 1040 myocardial infarction patients were diagnosed between 1982 and 1986 and followed up until 1997. Table 1 shows that 52.6% of the patients (547) experienced the event of interest (death) and the remaining 47.4% (493) were censored (alive or lost to follow-up).

Table 2 indicates that males were 2.5 times more than females in the study. 47.9% and 64.7% respectively experienced the event of interest, giving a hazard ratio (HR) of 0.74 implying that males had a lower mortality rate than females.

Table 3 showed that, of the 547 patients that experienced the event of interest (death) 15.35% were less than 54 years old, 18.09% were in the 54-61 years age bracket, 28.33% were in the age bracket of 62-70 years and 38.21% were in the age bracket of 71-95 years showing that there were more deaths in the older age brackets.

Table 4 showed that from the simplest architecture of the ANN relative survival design which employed the use of one hidden node, to increasing the complexity of the architecture due to the increase of hidden nodes, AUC was not so different from ACC for all three algorithms.

5. CONCLUSION

The values of the performance measures were not so different in their use in ANN for Relative Survival.

REFERENCES

Biglarian, A., Bakhshi, E., Baghestani, A. R., Gohari, M. R., Rahgozar, M., & Karimloo, M. (2013). Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics*, *2013*.

Tavakoli, A. R., & Seifi, A. R. (2016). Adaptive self-tuning PID fuzzy sliding mode control for mitigating power systems oscillations. *Neurocomputing*, *218*, 146–153.

Biganzoli, E. M., & Borrachi, P. (2005). The Partial Logistic Artificial Neural Network (PLANN): A tool for the flexible modelling of censored survival data. In *Proceedings of the European Conference on Emergent Aspects in Clinical Data Analysis (EACDA'05)*.

Sadiq, R., Oyejola, B. A., & Abiodun, A. A. (2016). Comparison of artificial neural networks and Cox regression models in population-based survival. *International Journal of Advanced Research in Science, Engineering and Technology*, *3*(3), 1718–1724.

De Laurentiis, M., & Ravdin, P. M. (1994). Survival analysis of censored data: Neural network analysis detection of complex interactions between variables. *Breast Cancer Research and Treatment*, *32*, 113–118.

Biganzoli, E., Boracchi, P., Mariani, L., & Marubini, E. (1998). Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Statistics in Medicine*, *17*(10), 1169–1186.

Soleimani, F., Teymouri, R., & Biglarian, A. (2013). Predicting developmental disorder in infants using an artificial neural network. *Acta Medica Iranica*, *51*(6), 347–352.

Sharaf, T., & Tsokos, C. P. (2015). Two artificial neural networks for modeling discrete survival time of censored data. *Advances in Artificial Intelligence*, *2015*(1), Article 270165.

Quantin, C., Abrahamowicz, M., Moreau, T., Bartlett, G., MacKenzie, T., Tazi, M. A., Lalonde, L., & Faivre, J. (1999). Variation over time of the effects of prognostic factors in a population-based study of colon cancer: Comparison of statistical models. *American Journal of Epidemiology*, *150*(11), 1188–1200.

Bach, P. B., Schrag, D., Brawley, O. W., Galaznik, A., Yakren, S., & Begg, C. B. (2002). Survival of blacks and whites after a cancer diagnosis. *JAMA*, *287*(16), 2106–2113.

Cutler, S. J., Axtell, L. M., & Schottenfeld, D. (1969). Adjustment of long-term survival rates for deaths due to intercurrent disease. *Journal of Chronic Diseases*, *22*(6/7), 485–491.

Dickman, P. W., Sloggett, A., Hills, M., & Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine, 23*(1), 51–64.

Geerligs, L. J., Anderegg, V. F., Holweg, P. A. M., Mooij, J. E., Pothier, H., Esteve, D., Urbina, C., & Devoret, M. H. (1990). Frequency-locked turnstile device for single electrons. *Physical Review Letters*, *64*(22), 2691.

Hakulinen, T., & Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *36*(3), 309–317.

Sasieni, P. D., Cuzick, J., & Lynch-Farmery, E. (1996). Estimating the efficacy of screening by auditing smear histories of women with and without cervical cancer. *British Journal of Cancer*, *73*(8), 1001–1005.

Bolard, M., & Boichard, D. (2002). Use of maternal information for QTL detection in a (grand) daughter design. *Genetics Selection Evolution*, *34*(3), 335–352.

Giorgi, A. P., & Giorgi, B. M. (2003). The descriptive phenomenological psychological method. *American Psychological Association*.

Lambert, D. M., García-Dastugue, S. J., & Croxton, K. L. (2005). An evaluation of process-oriented supply chain management frameworks. *Journal of Business Logistics, 26*(1), 25–51.

Neter, J., Kutner, M. H., & Nachtsheim, C. J. (2004). *Applied linear statistical models* (4th ed.). McGraw-Hill.

Variance-Range Hierarchical Clustering Method with Applications

*Eriobu N. O[1] and Abidoye A. O[2].

[1]Department of |Statistics, Nnamdi Azikiwe University, Awka, Anambra State.
[2]Department of |Statistics, University of Ilorin, Ilorin, Kwara State.

*Correspondence: no.eriobu@unizik.edu.ng

Abstract

In this paper, new procedure of linkage hierarchical cluster algorithm, is set out to present a more robust algorithm for proper classification. Classification method is given a better if it can be done with smaller steps or algorithm. A classification method with smaller number of steps is better than algorithm with higher number of steps if the two can achieve the same criterion. Based on this argument, the need for presentation of new classification algorithm is necessary which will equally reduce the apparent error rate in the classification. Instead of the commonly used methods such as single, complete and average linkages, variance approach for the generation of classification matrix was proposed as it reduces the effect of extreme values in the classification. Also, range as a measure of dispersion was used in the construction of dendogram. This resulted to a technique called Variance-Range (VR) |Hierarchical clustering method. To show the superiority and usefulness of the method, it was compared with some of the existing clustering methods using real-life dataset. |It was observed that VR Hierarchical clustering method is better for the classification of objects as it has similar dendogram with fewer steps in the classification procedure.

Keywords: Clustering, Variation, Similarity, Dendogram, Hierarchical, Dispersion

.

1. Introduction

Nature of similarity measure plays an important role in the failure or success of clustering method (May and Moe, 2016). Inter object similarity is a measure of correspondence or resemblance between objects to be clustered. Just as correlation matrix between variables is used in factor analysis, the characteristics are combined into a similarity measure calculated for all pairs of objects. Clustering is one of the most common methods of unsupervised learning in which data is segmented based on the similarity of instances (Hamidi, Akbari, and Motameni 2019). Cluster analysis procedure then proceeds to group similar objects together into clusters. Inter object similarity can be measured in a variety of ways but three methods dominate the applications of cluster analysis (Krishnamurthy, 2006); correlation measures, distance measures and association measures. Each of the methods represents a particular

perspective on similarity, dependent on both its objectives and the types of data. Both the correlation and distance measures require metric data, while the association measures are for non-metric data.

Correlation Measures can also be adopted. The inter object measure of similarity that probably comes to mind first, is the correlation coefficient between a pair of objects measured on several variables. The correlation between the two columns of numbers is the correlation (or similarity) between the profiles of the two objects. High correlation indicates similarity and low correlations denote a lack of it. A correlation measure of similarity does not look at the magnitude of the values but instead the patterns of values. But correlation measures are rarely used because emphasis in most applications of cluster analysis is on the magnitudes of the objects, not the patterns of values.

Distance $(x, y) = \sum_{i=1}^{n} \frac{(x_i - \mu_x)(y_i - \mu_y)}{(n-1)\sigma_x \sigma_y}$, where μ_x is the mean of x and σ_x is its standard deviation.

Distance measures are normally used to measure the similarity or dissimilarity between two data objects. Joining or tree clustering methods use the dissimilarities or distances between objects when forming the clusters. These distances can be based on a single dimension or multiple dimensions. Clusters based on correlation measures may not have similar values but instead have similar patterns. Distance-based clusters have more similar values across the set of variables but the patterns can be quite different. Some of the distance measures are:

- Euclidean distance:

  Euclidean distance is a standard matrix for geometrical problems and widely used in clustering problems. It is the ordinary distance between two points and can be easily measured with a ruler in a two or three-dimensional space. It works well when a data set has compact or isolated clusters (Sruthi and Reddy, 2013).

  Distance $(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

  This is computed from the raw data and not from consistent data.

- Squared Euclidean distance:

  One may want to square the standard Euclidean distance in order to place gradually greater weight on objects that are further apart. This distance is computed as:

  Distance $(x_i, y_i) = \sum_{i=1}^{n} (x_i - y_i)^2$

- City-Block (Manhattan) distance:

  This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, it is noted that in this measure, the effect of single large differences (outliers) is reduced since they are not squared. The city-block distance is computed as:

  $$\text{Distance}\ (x_i, y_i) = \sum_{i=1}^{n} |x_i - y_i|$$

There are various types of clustering;

i. Hierarchical clustering or Hierarchical cluster analysis: is a method of cluster analysis which seeks to build a hierarchy of clusters *(Rokach, Lior, and Oded, 2005).* It is also constructing the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. Hierarchical clustering produces a hierarchical tree of clusters called dendrogram. A dendrogram describing the relationships between all observations in a data set is useful for understanding the hierarchical relationships in the data, in many situations a discrete number of specific clusters is needed (Myatt, 2009). Examples of Hierarchical method include Agglomerative and Divisive. The weakness of hierarchical clustering is instability (unsteadiness) to scale well and can never undo what was done previously that is there is no back-tracking capability.

ii. Partitional clustering: is a method that divides the large object into (groups) clusters and each cluster containing at least one element. It follows an iterative process by use of this process; we can relocate the object from one group to another more relevance group. This method is effective for small to medium sized data sets. Examples of partitioning methods include k-means and k-medoids (Jiawei and Micheline, 2007).

iii. Density-based method: This method assumes that the points that belong to each cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993). Density based method provide high quality performance but it depends on two specified parameters, r and Minpts (Bharat and Manan, 2012).

2. Literature review

Martin et al (1996) proposed a clustering technique called DBSCAN. The method was developed in discovering clusters of arbitrary shape because the shape of clusters in spatial datasets may be spherical, draw-out, linear, elongated etc. The method DBSCAN is from the

word Density based spatial clustering of applications with Noise. The researchers use neighborhood and distance in the clustering of object for effective heuristic parameters. The method was compared with CLARANS and it was concluded that DBSCAN performed better in handling large dataset and also significantly more effective in discovering clusters of arbitrary shape.

Tian, Raghu and Miran (1996) proposed a clustering technique called BIRCH. The method was referred to as an efficient clustering method that can be used for a very large dataset. The method birch is from the word balanced iterative reducing and clustering using hierarchies. It was stated that the proposed method can handle noise which is a situation where data points are not part of the underlying pattern or structure of the given set of observations. The method was compared with an existing clustering method called CLARANS and it was concluded that Birch is better than Clarans in handling large dataset.

Ding and He (2002) provide a comprehensive analysis and experiments on divisive and agglomerative clustering. Similarities were used instead of the traditional distances. For divisive clustering, 4 new cluster selection criteria was introduced, the average similarity, the cluster cohesion, avg-cohesion, and temporary objective. Based on the 4 new cluster selection criteria introduced on internet newsgroups, it is shown that average similarity and similarity-cohesion selection perform well. For agglomerative clustering, MinMax linkage was introduced and compared with single-linkage, complete linkage and average linkage. MinMax linkage were found to be most effective than other in merging clusters.

Almeida et al 2007 presented a new method of hierarchical cluster analysis capable of detecting outlier and automatic clustering. This technique is based on agglomerative hierarchical clustering which is one of the most frequent approaches in unsupervised clustering. The algorithm comprises of three steps namely; treatment of outlier (outlier control), blocks identifier and group of blocks using similarity approach. The resulting classification method was validated by comparing it with some of the traditional methods. It was observed that the new method outperform some of the traditional methods in terms of clustering objects.

Basu and Murty (2013) developed am hierarchical clustering method called CUES. The method CUES is from the Clustering Using Extensive Similarity. The method were developed using a new cluster distance measure to identify two dissimilar clusters, it will never merger them but can stopped the algorithm if the distance between two clusters became very high. The

method was compared with existing methods (single-link hierarchical, average–link hierarchical, bisecting k-means, buckshot, k nearest neighbor and it was found that CURE perform significantly better than other existing clustering.

Yogita and Harish (2013) stated some improved hierarchical clustering algorithms like CURE, BIRCH, CHEMELEON, Linkage, Leaders–Subleaders, Bisecting k-Means that overcome the limitations that exist in pure hierarchical clustering algorithms. They went further to give some criteria on the basis of which one can also determine the best among these mentioned algorithms.

May and Moe (2016) presented a modified agglomerative hierarchical clustering (MAHC) which is an improvement on commonly used hierarchical clustering method. The modification was done to address the shortcomings of AHC such as precision, measurement and dimensionality. Instead of frequency of occurrency, the modified method used item based collation for the construction of distance matrix. It was concluded that MAHC has shorter algorithm and grouped items in clusters better than the AHC method.

Pelin and Derya (2017) proposed a hierarchical clustering method called K- Linkage which evaluates the distance between two clusters by calculating the average distance between k pairs of observations. The researchers introduces two concepts k-min linkage considers k minimum (closet) pairs from points in the first cluster to points in the second cluster, k-max linkage takes into account k maximum (farthest) pairs of observations. The improved hierarchical clustering algorithm based on k-linkage was executed on five well-known benchmark datasets with varying k values to demonstrate its efficiency. The results show that the proposed k-linkage method can often produce clusters with better accuracy, compared to the single, complete, average and centroid linkages.

Summary of Literature

It can be observed that clustering has unlimited application evidently from the literature as it has been used in different ways including grouping of locations or discovery of most common crime in locations. Also, comparison of the methods of clustering revealed that the strength of the methods varies significantly. This implies some of the existing methods can be referred to as better method than some other existing methods.

Since it is possible to detect a better method among existing hierarchical clustering methods, there is room for an improvement especially on the weak methods (see Olowoyori, 1994) for better performance.

Algorithm of hierarchical clustering can also be used in rating the methods as short steps with same achievement can be deemed to be better than a lengthy step of algorithm if both can perform the same assignment.

3. Methodology

Algorithm of the Proposed Methods

Variance-Range Classification Technique

1. Calculate variance to get variance matrix.
2. Locate the smallest value in the matrix.
3. Combine the point (say AB) with other points to get two values.
4. Calculate the range of the two values obtained.
5. Do same for all the points to get a reduced matrix.
6. Repeat step 2, 3 and 4, compile the new reduced matrix till the lowest possible matrix is formed.
7. Then, form the dendogram.

4. Comparison of the Proposed and Existing Methods

a. PROPOSED METHOD

Below is hypothetical data for the comparing of proposed method (Variance-Range) and existing one. The data was extracted from the publication Onyeagu 2003.

	$\mathbf{X_1}$	$\mathbf{X_2}$
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Using the formula for variance in equation 1 for computation of entries of similarities matrix

$$D_{\text{var}} = \sum_{i=1}^{n} \left(\frac{x_i - y_i}{n-1} \right)^2 \qquad (1)$$

$$V_{AA} = \frac{(1-1)^2 + (1-1)^2}{6} = 0$$

$$V_{AB} = \frac{(1-1.5)^2 + (1-1.5)^2}{6} = 0.08$$

$$V_{AC} = \frac{(1-5)^2 + (1-5)^2}{6} = 5.33$$

$$V_{AD} = \frac{(1-3)^2 + (1-4)^2}{6} = 2.17$$

$$V_{AE} = \frac{(1-4)^2 + (1-4)^2}{6} = 3$$

$$V_{AF} = \frac{(1-3)^2 + (1-3.5)^2}{6} = 1.71$$

$$V_{BC} = \frac{(1.5-5)^2 + (1.5-5)^2}{6} = 2.33$$

$$V_{BD} = \frac{(1.5-3)^2 + (1.5-4)^2}{6} = 1.42$$

$$V_{BE} = \frac{(1.5-4)^2 + (1.5-4)^2}{6} = 1.67$$

$$V_{BF} = \frac{(1.5-3)^2 + (1.5-3.5)^2}{6} = 1.04$$

$$V_{CD} = \frac{(5-3)^2 + (5-4)^2}{6} = 0.83$$

$$V_{CF} = \frac{(5-3)^2 + (5-3.5)^2}{6} = 1.04$$

$$V_{DE} = \frac{(3-4)^2 + (4-4)^2}{6} = 0.17$$

$$V_{DF} = \frac{(3-3)^2 + (4-3.5)^2}{6} = 0.04$$

$$V_{EF} = \frac{(4-3)^2 + (4-3.5)^2}{6} = 0.21$$

Below is the similarity matrix (variance matrix)

	A	B	C	D	E	F
A	0	0.08	5.33	2.17	3	1.71
B	0.08	0	2.33	1.42	1.67	1.04
C	5.33	2.33	0	0.83	0.33	1.04
D	2.17	1.42	0.83	0	0.17	**0.04**
E	3	1.67	0.33	0.17	0	0.21
F	1.71	1.04	1.04	0.04	0.21	0

$V(D, F)_A = R(d_{DA}, d_{FA})$

$R(2.17, 1.71) = 0.46$

$V(D, F)_B = R(d_{DB}, d_{FB})$

$R(1.42, 1.04) = 0.38$

$V(D,F)_C = R(d_{CD}, d_{CF})$

$R(0.83, 1.04) = 0.21$

$V(D,F)_E = R(d_{DE}, d_{EF})$

$R(0.17, 0.21) = 0.04$

	DF	A	B	C	E	DF
DF	0	0.46	0.38	0.21	**0.04**	
A	0.46	0	**0.08**	5.33	3.00	
B	0.38	0.08	0	2.33	1.67	
C	0.21	5.33	2.33	0	0.33	
E	0.04	3.00	1.67	0.33	0	

$V(DF, E)_A = R(d_{ADF}, d_{EA})$

$R(0.46, 3.00) = 2.54$

$V(DF, E)_B = R(d_{DFB}, d_{EB})$

$R(0.38, 1.67) = 1.29$

$V(DF, E)_C = R(d_{DFC}, d_{CE})$

$R(0.21, 0.33) = 0.12$

	DFE	A	B	C	DFE
DFE	0	2.54	1.29	0.12	
A	2.54	0	**0.08**	5.33	
B	1.29	0.08	0	2.33	
C	0.12	5.33	2.33	0	

$V(A, B)_{DFE} = R(d_{ADFE}, d_{BDFE})$

$R(2.54, 1.29) = 1.25$

$V(A, B)_C = R(d_{AC}, d_{BC})$

$R(5.33, 2.33) = 3.00$

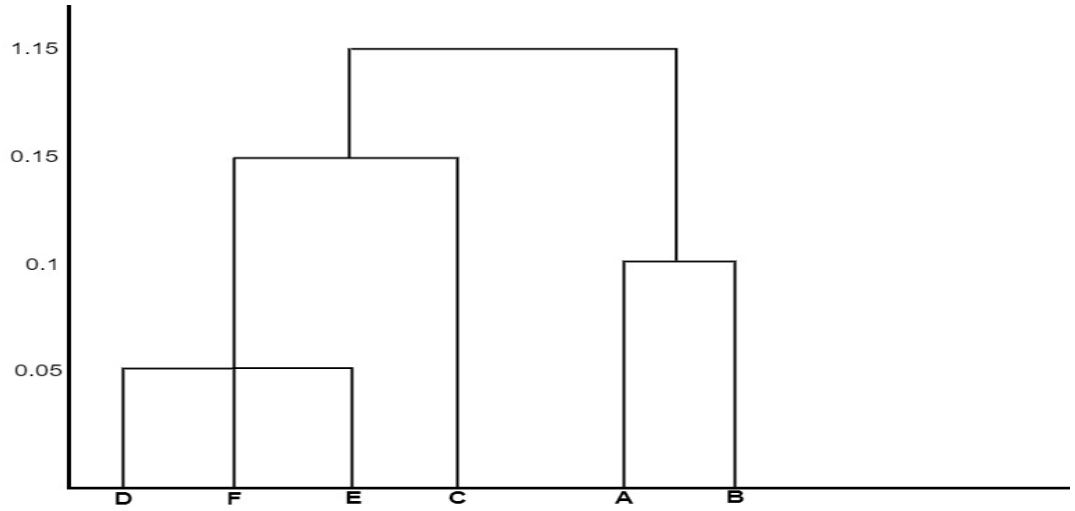
	AB	DEF	C
AB	0	1.25	3.00
DFE	1.25	0	**0.12**
C	3.00	0.12	0

$V(DEF, C)_{AB} = R(d_{DEFAB}, d_{ABC})$

R(1.25, 3.00) = 1.75

	CDEF	AB
CDEF	0	1.75
AB	1.75	0

DENDOGRAM



b. EXISTING METHODS

SINGLE LINKAGE

$d(D, F)_A = Min(d_{DA}, d_{FA})$

Min(2.17, 1.71) = 1.71

$d(D, F)_B = Min(d_{DB}, d_{FB})$

Min(1.42, 1.04) = 1.04

$d(D,F)_C = Min(d_{CD}, d_{CF})$

Min(0.83, 1.04) = 0.83

$d(D,F)_E = Min(d_{DE}, d_{EF})$

Min(0.17, 0.21) = 0.17

	DF	A	B	C	E
DF	0	1.71	1.04	0.83	0.17
A	1.71	0	**0.08**	5.33	3.00
B	1.04	0.08	0	2.33	1.6
C	0.83	5.33	2.33	0	0.33
E	0.17	3.00	1.67	0.33	0

$d(A, B)_{DF}$= Min(d_{ADF}, d_{BDF})

Min(1.71, 1.04) = 1.04

$d(A, B)_{C}$ = Min(d_{AC}, d_{BC})

Min(5.33, 2.33) = 2.33

$d(A, B)_{E}$ = Min(d_{AF}, d_{BE})

Min(3.00, 1.67) = 1.67

	AB	DF	C	E
AB	0	1.04	2.33	1.67
DF	1.04	0	2.33	**0.17**
C	0.8	2.33	0	0.33
E	0.17	1.67	0.33	0

$d(DF, E)_{AB}$= Min(d_{DFAB}, d_{EAB})

Min(1.04, 1.67) =1.04

$d(DF, E)_{C}$ = Min(d_{DFC}, d_{EC})

Min(0.83, 0.33) = 0.33

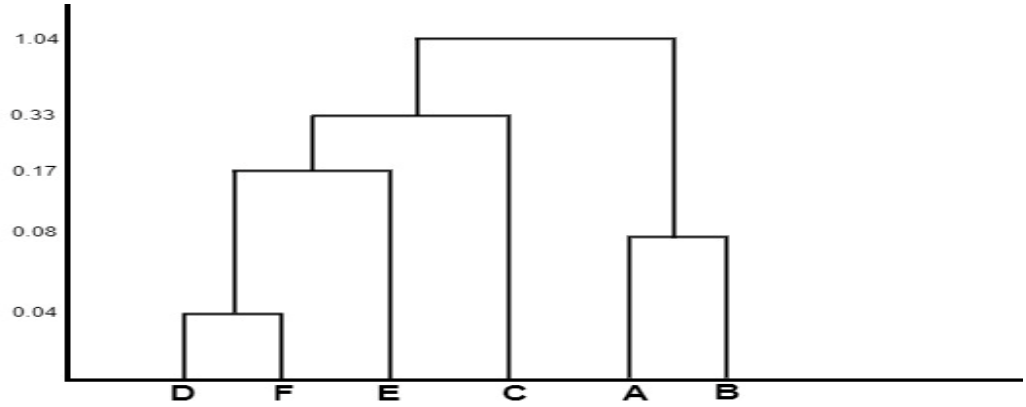
	DEF	AB	C
DEF	0	1.04	**0.33**
AB	1.04	0	2.33
C	0.33	2.33	0

$d(DEF, C)_{AB}$= Min(d_{DEFAB}, d_{ABC})

Min(1.04, 2.33) = 1.04

	DEFAB	C
DEFAB	0	1.04
C	1.04	0

DENDOGRAM



COMPLETE LINKAGE

	A	B	C	D	E	F
A	0	0.08	5.33	2.17	3.00	1.71
B	0.08	0	2.33	1.42	1.67	1.04
C	5.33	2.33	0	0.83	0.33	1.04
D	2.17	1.42	0.83	0	0.17	**0.04**
E	3.00	1.67	0.33	0.17	0	0.21
F	1.71	1.04	1.04	0.04	0.21	0

$d(D, F)_A = Max(d_{DA}, d_{FA})$

Max(2.17, 1.71) = 2.17

$d(D, F)_B = Max(d_{BD}, d_{BF})$

Max(1.42, 1.04) = 1.42

$d(D, F)_C = Max(d_{CD}, d_{CF})$

Max(0.83, 1.04) = 1.04

$d(D, F)_E = Max(d_{DE}, d_{EF})$

Max(0.17, 0.21) = 0.21

	DF	A	B	C	E
DF	0	2.17	1.42	1.04	0.21
A	2.17	0	**0.08**	5.33	3.00
B	1.42	0.08	0	2.33	1.67
C	1.04	5.33	2.33	0	0.33
E	0.21	3.00	1.67	0.33	0

$d(A, B)_{DF} = Max(d_{ADF}, d_{BDF})$

Max(2.17, 1.42) = 2.17

$d(A, B)_C = Max(d_{AC}, d_{BC})$

Max(5.33, 2.33) = 5.33

$d(A, B)_E = Max(d_{AE}, d_{BE})$

Max(3.00, 1.67) = 3.00

	AB	DF	C	E
AB	0	2.17	5.33	3.00
DF	2.17	0	1.04	**0.21**
C	5.33	1.04	0	0.33
E	3.00	0.21	0.33	0

$d(DF, E)_{AB} = Max(d_{ABDF}, d_{EAB})$

Max(2.17, 3.00) = 3.00

$d(DF, E)_C = Max(d_{CDF}, d_{CE})$

Max(1.04, 0.33) = 1.04

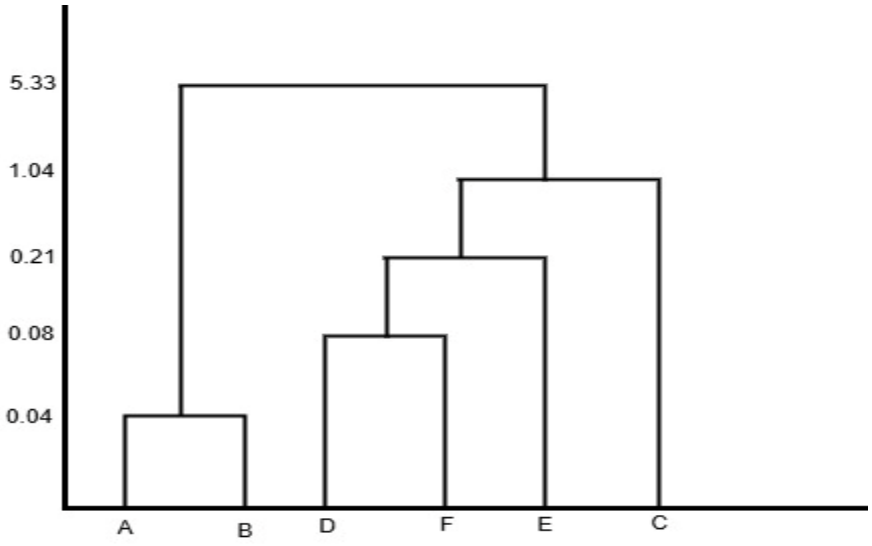
	DFE	AB	C
DFE	0	3.00	**1.04**
AB	3.00	0	5.33
C	1.04	5.33	0

$d(DFE, C)_{AB} = Max(d_{DFEAB}, d_{CAB})$

Max(3.00, 5.33) = 5.33

	CDEF	AB
CDEF	0	5.33
AB	5.33	0

DENDOGRAM



AVERAGE LINKAGE

	A	B	C	D	E	F

A	0	0.08	5.33	2.17	3	1.71
B	0.08	0	2.33	1.42	1.67	1.04
C	5.33	2.33	0	0.83	0.33	1.04
D	2.17	1.42	0.83	0	0.17	**0.04**
E	3	1.67	0.33	0.17	0	0.21
F	1.71	1.04	1.04	0.04	0.21	0

$d(D,F)_A = ½ (d_{DA}, d_{FA})$

½ (2.17+1.71) = 1.94

$d(D,F)_B = ½ (d_{BD}, d_{BF})$

½ (1.42+1.04) = 1.23

$d(D,F)_C = ½ (d_{CD,} d_{CF})$

½ (0.83+1.04) = 0.94

$d(D,F)_E = ½ (d_{DE}, d_{EF})$

½ (0.17+0.21) = 0.19

	DF	A	B	C	E
DF	0	1.94	1.23	0.94	0.19
A	1.94	0	**0.08**	5.33	3.00
B	1.23	0.08	0	2.33	1.67
C	0.94	5.33	2.33	0	0.33
E	0.19	3.00	1.67	0.33	0

$d(A, B)_{DF} = ½ (d_{ADF}, d_{BDF})$

½ (1.94+1.23) = 1.59

$d(A,B)_C = ½ (d_{AC}, d_{BC})$

½ (5.33+2.33) = 3.83

$d(A,B)_E = ½ (d_{AE}, d_{BE})$

½ (3.00+1.67) = 2.34

	AB	DF	C	E
AB	0	1.59	3.83	2.34
DF	1.59	0	0.94	**0.19**
C	3.83	0.94	0	0.33
E	2.34	0.19	0.33	0

$d(DF,E)_{AB} = ½ (d_{DFAB}, d_{EAB})$

½ (1.59+2.34) = 1.94

$d(DF,E)_C = ½ (d_{DFC}, d_{EC})$

½ (0.94+0.33) = 0.64

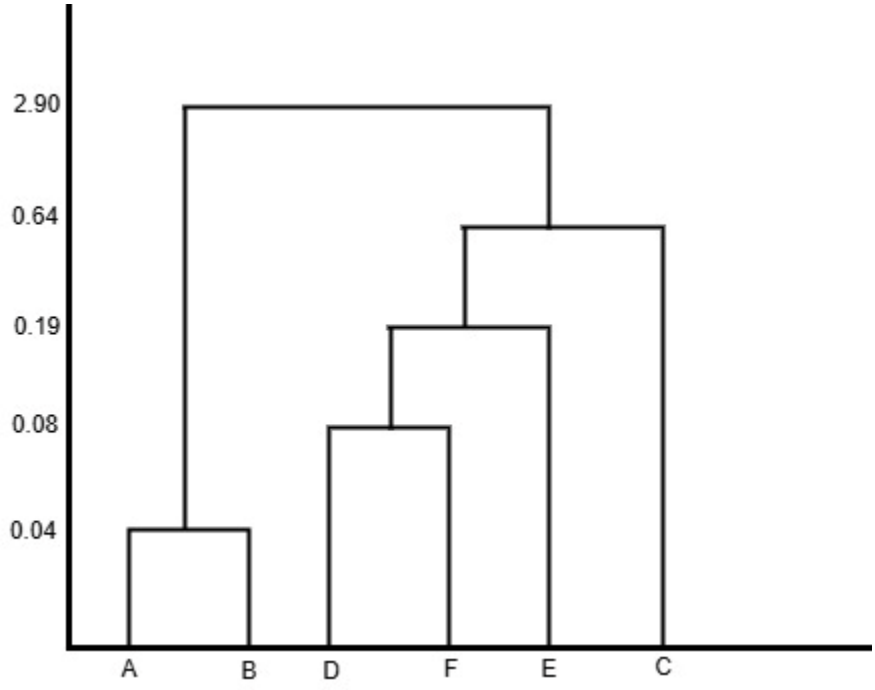
	DEF	AB	C
DEF	0	1.97	∞
AB	1.97	0	3.83
C	**0.64**	3.83	0

$d(DEF, C)_{AB}= ½ (d_{DEFAB}, d_{ABC})$

½ (1.97+3.83) = 2.9

	CDEF	AB
CDEF	0	2.9
AB	2.9	0

DENDOGRAM



5. Discussion of Results

As observed from the results both existing and proposed method generated similarity dendogram which is an indication that the proposed method is capable of classifying the variables or observations the same way as the existing method does. The existing method used Euclidean distance as a measure of similarities but the proposed method used variance as substitute in the algorithm. Range was used for the determination of values used for the dendogram. This implies in the proposed method, measures are variance and range.

6. Summary of Findings and Conclusion

In the classification techniques, as observed in the literature, there exist numerous methods with limitations in terms of accuracy, efficiency, length of algorithm, sensitivity to extreme values as well as complexity in the computation. Better method of classification can be in terms of shorter algorithm, better control of extreme values, less complex etc. |In this study, a less cumbersome with shorter algorithm method of classification was proposed. The method utilizes variance and range which can easily be computed and less rely on statistical assumptions such as normality etc. This implies the method can be referred to as robust because it relies on less number of assumptions which makes it more useful among the existing methods. The suitability and superiority of the method was shown using real life data extract from the work of Onyeagu (2003). From the result of the real life analysis, it can be observed that the proposed methods favourably competed with the existing methods which make it useful for classification purpose. Therefore, Variance-Range Classification technique can be adopted in the classification of observations or variables especially when dealing with problems that have to do with classification.

References

Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, *87*, 208–217.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803–821.

Chaudhari, B., & Parikh, M. (2012). A comparative study of clustering algorithms. *International Journal of Application or Innovation in Engineering & Management*.

Ding, C., & He, X. (2002). Cluster merging and splitting in hierarchical clustering algorithms. In *Proceedings of the International Conference on Data Mining (ICDM 2002)* (pp. 139). IEEE.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). AAAI Press.

Han, J., Kamber, M., & Pei, J. (2007). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hamidi, S. S., Akbari, E., & Motameni, H. (2019). Consensus clustering algorithm based on the automatic partitioning similarity graph. *Data & Knowledge Engineering*, *124*, Article 101754.

Krishnamurthy, P. (2006). *Approaches to clustering gene expression time course data* (Master's thesis). Department of Computer Science and Engineering, State University of New York at Buffalo.

Lwin, M. T., & Aye, M. M. (2016). Modified Agglomerative Hierarchical Clustering (MAHC) algorithm for document clustering. *International Journal of Advances in Electronics and Computer Science*, *3*(8).

Myatt, G. J. (2009). *Making sense of data II*. Wiley.

Neter, J., Kutner, M. H., & Nachtsheim, C. J. (2004). *Applied linear statistical models* (4th ed.). McGraw-Hill.

Onyeagu, S. I. (2003). *A first course in multivariate statistical analysis*. Jasco Press.

Rani, Y., & Rohil, H. (2013). A study of hierarchical clustering algorithm. *International Journal of Information and Computation Technology*, *3*(10), 1115–1122.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321–352). Springer.

Sruthi, K., & Reddy, B. V. (2013). Document clustering on various similarity measures. *International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(8).

Yildirim, P., & Birant, D. (2017). K-Linkage: A new agglomerative approach for hierarchical clustering. *Advances in Electrical and Computer Engineering*, *17*(4), 77–88.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 103–114). ACM.

The Impact of Inflation Targeting and Economic Growth in Emerging Economies from 1980-2023: Should Nigeria adopt?

J. O. Anigwe

Statistics Department, Central Bank of Nigeria, Abuja

joanigwe@cbn.gov.ng

Abstract

This article studied the relationship between inflation and GDP growth in selected emerging economies. The objective is to explore the relationship between the two important macroeconomic indicators, appropriate explanation, and offer adequate policy recommendations to the monetary authority of the Nigerian economy. The study considered data from 14 countries, including New Zealand, the Czech Republic, Poland, Brazil, Chile, South Africa, Colombia, Hungary, Mexico, Peru, the Philippines, South Korea, Thailand, Turkey, and Nigeria. The study considered panel annual data from 1991 to 2022, including nominal GDP, inflation rates, nominal exchange rates, and trade balance. It employed a pooled mean group auto-regressive distributed lag (PMG/ARDL) model. An a priori relationship between inflation and the dependent variables was confirmed, as its increase has a proportional positive impact on economic growth in both the short run and long run. The nominal exchange rates showed an inverse relationship. Foreign exchange rate impact was not statistically significant in the short run, though positive in the long run. However, the trade balance posited an inverse relationship with the dependent variable in the long run but a positive effect in the short run. This calls for a more in-depth study to unravel the range of inflation rates that trigger productivity in the economy to inform the appropriate policy measure to adopt.

1. Introduction

The inflation targeting (IT) regime was first adopted by New Zealand in 1990 and has become an important phenomenon in central banking since then. It is a monetary policy framework adopted by central banks to manage and stabilize inflation rates within a specific target range. The primary goal of IT is to maintain price stability by controlling inflation. Central banks set a specific inflation target (e.g., 2% annually) as their policy objective. However, the apex bank in a certain economic territory sets up certain mechanism on the road to achieving inflation. They use interest rates (such as the policy rate) to influence inflation. When inflation exceeds the target, the central bank raises

interest rates to reduce demand and curb price increases. On the other hand, if inflation is below the target, interest rates are lowered to stimulate economic activity.

Broadly, the subject of inflation targeting is like an adventure, and it drags a basket of necessity which forms the reasons for the practice. Primarily, it is necessary for the maintenance of stable prices in the economy. By targeting a specific inflation rate, central banks aim to prevent excessive price fluctuations, which can disrupt economic activities and erode purchasing power. It provides clarity to market participants, businesses, and households. When people expect stable inflation, they can make informed decisions regarding investments, savings, and consumption. Where this is being implemented, the monetary authority publicly announces their inflation targets and communicate policy decisions for transparency and accountability purposes in order to enhance credibility and foster trust in the central bank's actions. From the ongoing we see that inflation targeting provides a guide to central banks to setting interest rates. When inflation deviates from the target, interest rates are adjusted accordingly to achieve the desired outcome. More so, they also try to maintain some trade-offs and Priorities which aides in the efforts to balance inflation control with other macroeconomic goals such as achieving subsistent economic growth, optimal employment. Striking a balance between inflation control and economic growth can be complex. This is much more so a challenge guarding against structural issue and external shocks in the form of fluctuations in the oil price.

Brazil implemented IT in 1999. Research suggests that Brazilian monetary policy has been procyclical and asymmetric, impacting aggregate demand negatively during downturns (Gilberto, 2007). The main policy implication is that central banks should consider the real effects of monetary policy on output and employment. Chile adopted IT in 1999. Evidence indicates that IT helps Chile avoid high inflation during economic uncertainties while maintaining output growth performance. According to Czech National Bank (CNB), the inflation target has been 2% year-on-year growth in the consumer price index. This target is in line with the practice of the central banks of advanced economies. The CNB regularly assessed the fulfilment of the inflation target on the basis of the statistical data published by the Czech Statistical Office, an institution independent of the central bank. This arrangement enhances the credibility of inflation targeting. The Czech Republic has been practicing IT since 1997. IT

contributes to maintaining price stability and steering inflation expectations, supporting sound economic growth. New Zealand was an early adopter of IT in the late 1980s. The country became the first to adopt this policy and continued the practice which has contributed to stable inflation and economic growth in the country. Poland embraced IT in 1998. While research on Poland's specific case is limited, IT generally helps stabilize inflation and supports growth. South Africa adopted IT in 2000. IT has helped South Africa manage inflation and maintain output growth during economic uncertainties. In summary, IT has been effective in maintaining price stability and supporting economic growth across these countries (Duong, 2022). However, each nation's experience varies based on its unique economic context and policy implementation. In summary, inflation targeting aims to strike a balance between price stability and economic growth, with central banks adjusting interest rates to achieve their inflation targets.

Over the past five years, Nigeria's inflation rate has experienced obvious fluctuations. In 2021, it reached 16.9%, marking a 3.7% increase from the previous year. At the end of 2022 and 2023, it further rose to 18.9% and 28.9%, respectively, representing increases of 36.6% and 35.5%. Overall, inflation has been volatile, impacting purchasing power and economic stability. On the other hand, Nigeria's economic growth has seen ups and downs in recent time. In 2021, the GDP was $440.84 billion, a 2% increase from 2020. However, in 2020, it declined by 8.92% compared to 2019. The highest recent growth rate was in Q3 2020 at 12.12%. Challenges like oil price fluctuations and structural issues have influenced growth. In summary, Nigeria has faced inflationary pressures while navigating economic growth challenges. Policy measures are crucial to stabilize inflation and foster sustainable growth. So as the monetary authority announces the plan to implement inflation targeting, this project seek to ascertain the path of economic growth with inflation and confirm the necessity of the planned monetary strategy.

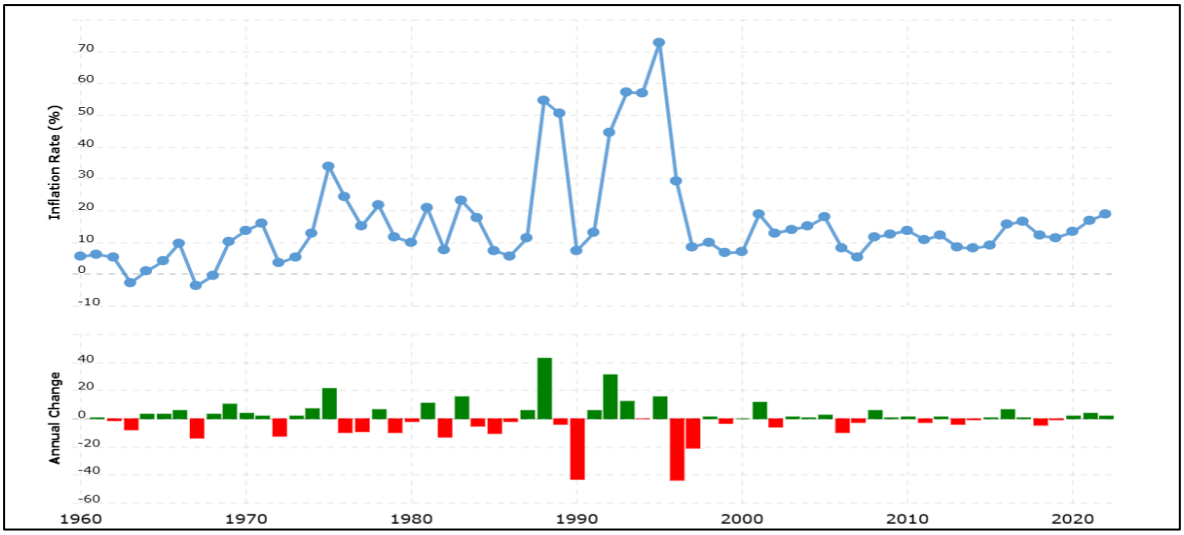


Fig. 1 Nigeria Inflation (1960-2024). *Source:* Macrotrends

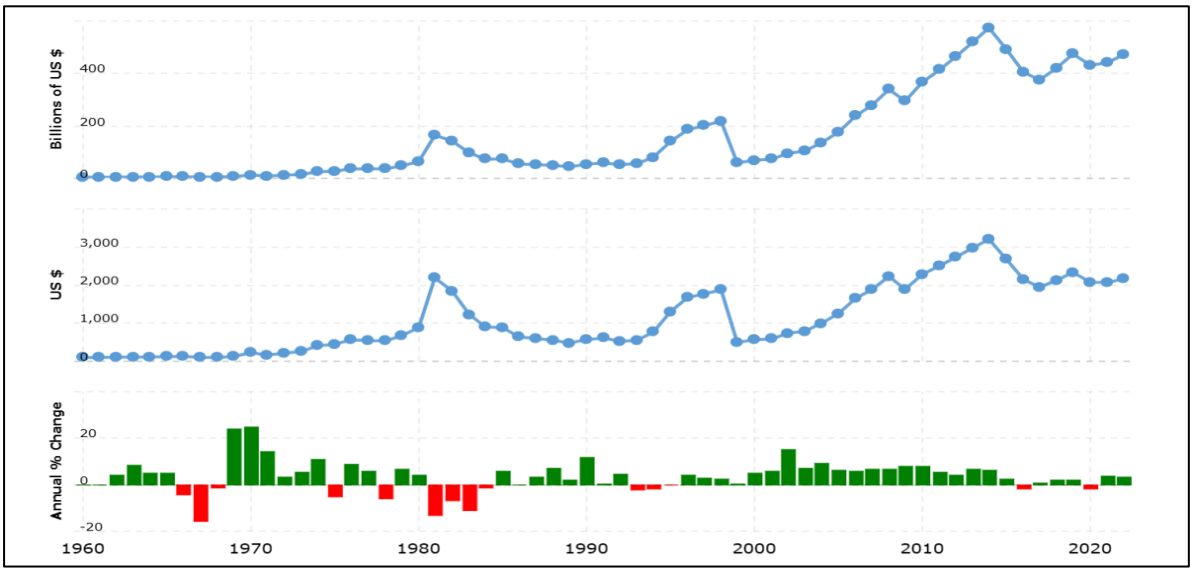


Fig. 2. Nigeria Economic Growth (1960-2024). *Source:* Macrotrends

2. Literature Review

What is the rational for adopting inflation targeting by various economies? Should monetary authorities adopt this economic strategy remains a question as some battle with galloping inflation and issues with the economic growth? The goal of the inflation targeting regime is to provide a kind of monetary strategy that will provide a nominal anchor to the economy (Guillermo, 2008). Accordingly, this is necessary to guarantee the central bank pursuing long-run policies and resisting the problem or the political

pressures to consider short-run expansionary strategies that are not consistent with the objective of long-run price stability.

It was observed that in the past, some countries practiced currency pegs where they link their currency to another low-inflation country. This practice was disadvantaged because it was not able to respond to economic disturbances. Inflation targeting became an alternative nominal anchor. However, for inflation targeting to be a success the central bank is required to be, to reasonably extent, independent from government intrusion and free to choose appropriate instrument that enables the achievement of the rate of inflation that the government considers suitable for their fiscal operations (Debelle et al, 1998).

More argument in favour of the implementation of inflation targeting reveals that inflation targeting supports macroprudential policy roles in mitigating financial stability risks (Belkhir et al, 2023). The study which examined the impact of macroprudential policy on bank systemic risk and the part which inflation targeting plays in such scenario, analyzed data of forty-five countries using bank-level data. These economies cover different monetary and exchange rate regimes. However, findings also show that complementary relationship existed between monetary and macroprudential guidelines. In addition, this was made effective by the tightening of most macroprudential tools which included debt service-to-income and loan-to-value limits, and capital requirements.

As threshold, the gross domestic product (GDP) measures the total market value of all goods and services produced within a country during a specific period. It is used as measure the totality of the periodic productivity and its economic wellness and provides an economic snapshot, allowing people to estimate the size of the economy and its growth rate (Kramer, 2024). GDP provides very necessary perceptions into an economy's total well-being and guides stakeholders on taking investment decision at various instances.

What is the interaction between GDP and inflation and hence the impact of inflation targeting? Monetary policy (MP) has a direct impact on the regulation of the economy in the framework of the transition to inflation targeting (Eshov & Benetti, 2023). The investigation which employed four criteria from the 4×2 matrix model on a ten-year period Uzbekistan data show that increases in bank capital adequacy, loan portfolio,

money supply, and lower inflation help maintain macroeconomic stability. Further analysis revealed a significant positive relationship exists between the MP instruments economic growth in the long-run. On the other hand, negative relationship existed between country's economic growth relative to the increase of the bank deposits and depreciation of local currency.

The link between inflation and economic growth is indeed complex and can be considered in two distinct dimensions. When economic growth is high, leading to increased demand for goods and services prices go up if supply does not match demand and also leads to higher costs of production which businesses transfer higher prices to consumers. The debate goes on discussing the interlinkages between macroeconomic indicators in various economies and different geographical regions. Outcomes differ depending on their various combinations.

Baltar 2015, investigated Inflation and economic growth in an open developing country. The study explored the case of Brazil considering GDP growth, domestic inflation and nominal exchange rate, tradable sector Inflation and non-tradable sector Inflation. It was found that direct correlation between the growth rate of the nominal exchange rate and inflation, inverse correlation between inflation and GDP growth and direct correlation between GDP growth and the relative price of non-tradeable goods. Furthermore, it concluded that reduction in international price proxied in domestic price reduction causes reduction in domestic inflation. Lower nominal exchange rate growth reduces the growth rate of exported and imported inputs price, decreasing the cost growth rate of tradeable and non-tradeable goods production. In Nigeria situation, the question is, what will be appropriate inflation rate that would drive the growth. Hence the study, how much is too much for economic growth in Nigeria (Fabayo & Ajilore, 2006)? This is about inflation rate. The work concluded that existence of Inflation threshold of 6% was confirmed for the period under the study and that positive relationship exists at that level and levels below. At the point inflation goes above the threshold, growth is impeded. At situation as this, it is simply natural to recommend maintaining the threshold within the revealed range. Hence the monetary authority should be at work to committing their appropriate tool to operation. Hence, the study suggested that single digit inflation should be the goal of macroeconomic management in Nigeria taking 6 percent as optimal target. How practically could this be implemented

successfully. Nigeria has been in episodes of inflation rise for over a substantial period (Mohammed, 2021). The study stated that inflation in Nigeria may spring up from various sources including money supply, interest rate, and oil, however, the monetary authorities have been using monetary targeting framework to implement the mandate of price stability that enables growth to thrive. Still question begging for response remains, how effective has the frame been? A well implemented policy could mitigate the negative impact of inflation. This is supported Ewurum, Kalu, Nwankwo, 2017 stating that all the countries that implemented inflation targeting recorded a reduction in the inflation levels. Another study which considered inflation targeting and economic growth in Nigeria supports that inflation targeting is marginally effective to cause growth (Saidu, 2017). Musa, Magaji & Salisu, 2021 employed SVAR analysis to examine the impact of monetary policy in the context of Inflation Targeting on economic growth in Nigeria and found out that monetary policy has a positive shock on economic growth. Though the impact was minimal, however, it was confirmed that the monetary policy rate has a positive relationship with growth at three percent. From analysis a conclusion that inflation targeting framework, in addition to other supplementary instruments can be appropriate in mitigating constraints of economic growth. With the peculiar situation in the Nigeria economy where inflation is already above 30 percent mark, interest rate increased consecutively by monetary authority and the floating of the exchange, as the apex bank in the country plans to adopt inflation targeting, what impact will this have on the productive capacity of the economy and its interlinkage effects on the various macroeconomic indicators to give a guide to direction to relevant authorities. This is the focus of this work.

3. Data and Methodology

3.1 Data

A balanced data set including the nominal gross domestic product (gdp), inflation rate, exchange rates and trade balance covering the period 1991 to 2022 for the selected countries, namely: Brazil, Colombia, Chile, Hungary, Poland, South Africa, New Zealand, Korea People's Republic, Turkey, Thailand, Peru, Philippines, Mexico and Nigeria. For uniformity, the exchange rates data represents the nominal exchange rates of the respective domestic currencies per United States Dollar (US$). For uniformity, all data are sourced from the World Bank Development Indicators.

The selected countries adopted inflation targeting for over 2 decades and have varied experiences and outcomes. Each country's specific context and policy decisions influenced their inflation targeting approach during these economic tempest times, 2007-2009 and the COVID-19 period. The emergence of inflation during the pandemic was different from previous global recessions. In 2020, global inflation declined due to collapsing demand and falling oil prices. However, since May 2020, inflation has rebounded rapidly, primarily driven by increased global demand. More than fifty percent of inflation-targeting emerging economies, developing and some advanced economies, the increases in inflation during

Table 1. Selected Countries that practised Inflation Targeting (IT) Policy Framework.

Country	Inflation Targeting Adoption Year	Annual Average Inflation Rate (2007-2009)	Annual Average Inflation Rate (2019-2022)
New Zealand	1991	2.8	3.6
Poland	1998	3.5	6.3
Korea People's Republic	1998	3.3	2.1
Chile	1999	5.6	4.9
Brazil	1999	4.7	6.1
Colombia	1999	5.6	4.9
South Africa	2000	7.8	4.7
Thailand	2000	2.3	1.8
Hungary	2001	6.1	6.6
Mexico	2001	4.8	5.2
Turkey	2006	8.5	29.8
Peru	2002	3.5	4.2
Philippines	2002	5.1	3.6
Nigeria	Non-IT	9.8	15.1

the pandemic were still found around their target bounds. If inflation expectations remain anchored, it may not require immediate monetary policy adjustments. However, central banks in some of those countries needed to observe inflation expectations to prevent them from becoming unanchored.

3.1.1 Descriptive Statistics

a. Nigerian Data

The summary statistics of some Nigeria macroeconomic variables used in the study are presented in Table 2. The country's average inflation rate is approximately 18.4 percent

from 1991 to 2022, with maximum and minimum rates of 72.8 percent and 5.4 percent, respectively. The GDP growth during the same period averaged 4.1 percent, ranging from a maximum of 15.3 percent to a minimum of -2.0 percent.

Table 2. Summary Statistics of Variables (Nigeria Specific)

Variables	Mean	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis	Probability
GDP	320.96	535.34	153.73	142.23	0.16	1.40	0.17
G	4.05	15.33	(2.04)	3.78	0.48	3.79	0.35
P	18.42	72.84	5.39	16.25	2.16	6.62	0.00
TB	16.84	63.70	(5.01)	17.98	1.08	3.30	0.04
FX	150.88	425.98	9.91	115.78	0.83	2.92	0.16

Note: GDP is Gross domestic product (USDBillion); G is GDP growth rate; P is Inflation rate; TB is Trade balance (USD Billion); and Fx is Nominal exchange rates in all the work.

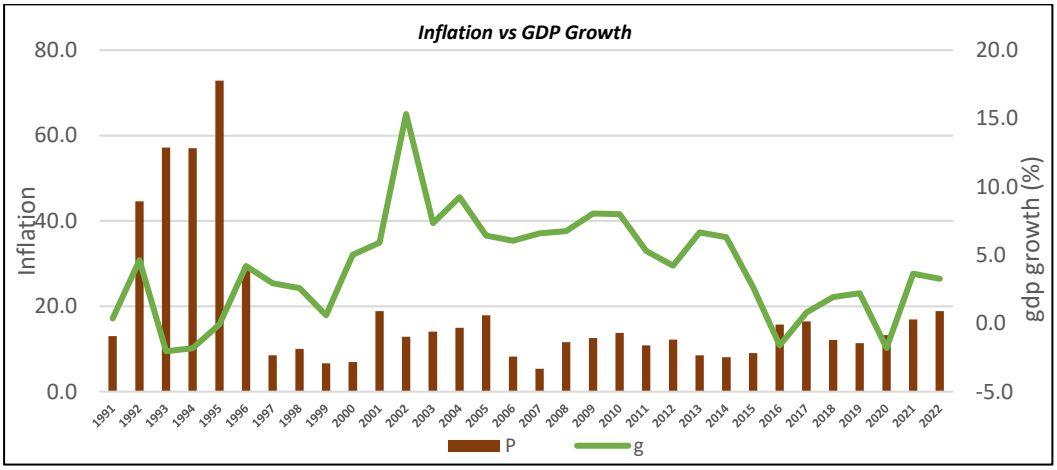


Fig. 2 Nigeria data Inflation rate and gdp growth Rate.

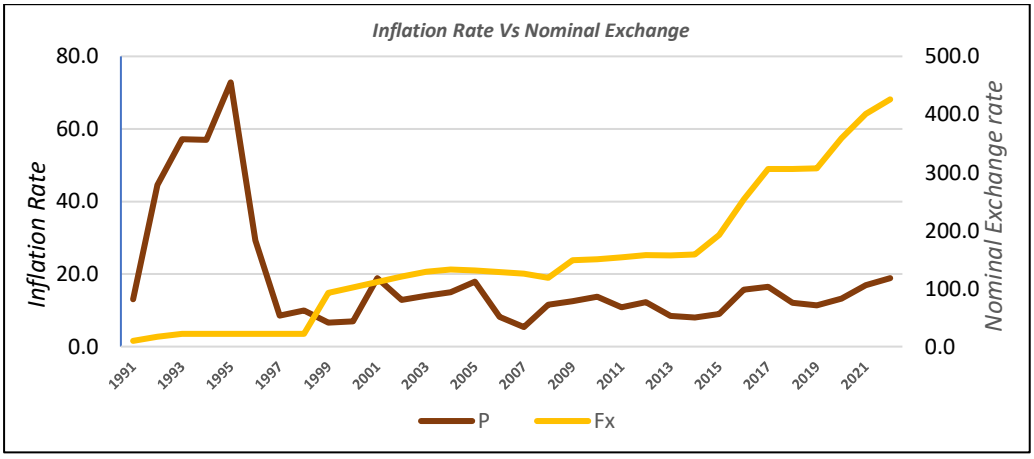


Fig. 3. Nigeria Inflation rate and Foreign Exchange Rate.

b. Panel Data

Table 3 shows the summary statistics of the selected macroeconomic variables of the selected sample of countries used in the study. The mean GDP for the period under invested posited US$ 464. 53 Billion, with maximum and minimum values of USD1,901.46 Billion and USD60.58 Billion, respectively. The average inflation rate is approximately 22.8 percent during the period. The averagely, negative trade balance was recorded for the period at (USD3.25), the lowest being (USD109.52). This shows that most the selected countries were import dependent. The average exchange rate per USD was 416.99 and GDP growth averaged throughout the period, 3.5 percent cross-sectionally.

Table 3. Summary Statistics of Variables (The Panel Data)

Variables	Mean	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis	Probability
GDP	464.5269	1,901.46	60.5855	443.9246	1.5506	4.4588	0.0000
G	3.5182	15.33	(10.8696)	3.4722	(0.6912)	5.0887	0.0000
P	22.8264	2,075.89	(0.9004)	143.4336	12.5165	168.0342	0.0000
TB	(3.2552)	95.22	(109.5190)	22.6014	(0.3038)	8.4784	0.0000
FX	416.9930	4,256.19	0.0002	833.9082	2.3558	8.0588	0.0000

c. Panel Unit root Test

We applied the Levin, Lin & Chu (2002), Im, Pesaran and Shin W-stat, (2003), ADF-Fisher Chi-square (1981) and PP-Fisher Chi-square panel unit root test on all series thereby putting into consideration the cases of single unit roots of our panel data. From the test result, it is confirmed gdp growth rate and inflation rate were stationary at level, though Levin, Lin and Chu has not confirmed this (Table 4). The rest of gdp absolute, nominal exchange rates and trade balance were found to be stationary at first difference with a p-values of zero (0.0000).

Table 4. Panel Unit Root Result

Macroeconomic Variable	Levin, Lin & Chu t*	Im, Pesaran and Shin W-stat	ADF - Fisher Chi-square	PP - Fisher Chi-square
GDP -I(1)	-8.5414 (0.0000)	-9.0240 (0.0000)	134.770 (0.0000)	251.582 (0.00000
G - I(0)	-6.7788 (0.0000)	-8.2888 (0.0000)	122.114 (0.0000)	223.763 (0.0000)
P - I(0)	2.7324 (0.00000)	-7.7463 (0.00000)	96.3179 (0.0000)	142.562)
FX - I(1)	-4.5871 (0.0000)	-6.2485 (0.0000)	115.277 (0.0000)	148.611 (0.0000)
TB - (I(1)	-8.6887 (0.0000)	-11.0741 (0.0000)	172.568 (0.0000)	189.262 (0.0000)

3.2 Methodology

The main objective of this study is to test the null hypothesis that the inflation targeting framework is an appropriate monetary policy tool that boosts the economic performance in terms of the inflation rate and output growth in the IT countries especially, during episodes of economic shocks. The panel ordinary least square approach is used to provide the evaluation of inflation targeting impacts on the dependent variable.

Hence, for this purpose, both the Fixed Effects Model, also known as the least squares dummy variable (LSDV) model, and the Random Effects Model was utilized. A balanced panel data set which has equal number of observations for each country will be used in the cross-sectional analysis while fixed effects model hypothesis testing, random effects model versus fixed effects model, Hausman specification test will all be applied for model selection (see Gujarati, 2003 & Hsiao, 2003).

The model and basic framework adopted in this study is similar to that of Greene (2010) and Akbar et al (2011). The later studied the indicators that determine economic growth in Asian countries using panel data. Most of the discussion will centre on regression model of the form in equation (1) amongst a variety of different models applied for panel data in this research paper.

$$y_{it} = x'_{it}\beta + Z'_i\alpha + \varepsilon_{it}$$

$$y_{it} = x'_{it}\beta + c_i + \varepsilon_{it} \quad (1)$$

There are N regressors in x_{it}, without a constant term. The heterogeneity, or country effect is $Z'_i\alpha$ where Z_i has a constant term and a set of country specific variables, that could be considered, such as gross domestic (GDP), inflation rates (P), trade balance (Tb) and nominal exchange rates (Fx) or unobserved, such as regional specific characteristics, and country's heterogeneity in culture or preferences, such that they are regarded as constant over time t. If Z_i is observed for all countries, then the complete model can be taken as an ordinary linear model and fit by least squares.

$$\beta = \partial E[y_{it}|\frac{x_{it}}{\sigma x_{it}}] \quad (2)$$

The main objective in equation (3.2) is to have a consistent and efficient estimation of the partial effects which largely depends on the assumptions of the unobserved effects. However, it is best to first begin with operation of the strict exogeneity assumption for the independent variables as indicated below in equation (3.3).

$$E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots,] = 0| \quad (3)$$

This demonstrates that the existing restraints should not have any correlation with the free variables during the entirety of the period t. The heterogeneity of the input data is the most important part of the equation.

$$E[c_i|x_{i1}, x_{i2}, \dots,] = \alpha \quad (4)$$

A more convenient assumption would be the mean independence in equation (3.4), which states that if the missing variable (or variables) are uncorrelated with the variables that are included in the model, then it implies that those variables may be represented in the model residual. On the other hand, this assumption is what is known as a strong assumption, and it is what underpins the random effects model, which we are going to investigate later.

$$E\ [c_i \mid x_{i1}, x_{i2},\dots,] = h\left(x_{i1,}, x_{i2},\dots,\right) = h(X_i). \quad (5)$$

As a more general alternative to equation (3.4), equation (3.5) has been presented. This equation is more complex due to the possibility of requiring even more assumptions about the function's nature.

Assuming that coefficients are constant across time and persons, assuming that there is no major country impact or substantial temporal effect, and pooling all of the data together, we may run an ordinary least squares (OLS) regression model on all of the data.

In the work, fixed effect model will be estimated as indicated in equation (6) below in order to allow for cross-section specific intercepts (β_i). We capture country-specific effects only through cross-section specific intercepts and try to save degrees of freedom by modeling common slope parameters (β_0, β_1, β_2, and β_3) thus estimating eq. (6) for each of the fourteen countries:

Here is what the model would look like:

$$G_{it} = \beta_0 + \beta_1 P_{it} + \beta_2 logFx_{it} + \beta_3 logTb_{it} + \mathcal{U}_{it} \qquad (6)$$

Where G_{it} is the gross domestic product (gdp) growth, P_{it} is inflation rates, logFx_{it} is the natural log of nominal exchange rates, logTb_{it} is the natural log of trade balance of the *i*th country and at a time period, *t*, respectively.

4. Empirical Results

4.1 Pool Mean Group (PMG) /Auto-Regressive Distributed Lag (ARDL) Model Analysis Result

The model considered lag length structure of 4, with Akaike info criterion (AIC). The specified model includes the three macroeconomic variables as stated earlier including inflation rate, log of nominal exchange rates and the log of trade balance. From the results significant impact of all the independent variables were observed. The long-run impact of inflation was positive as expected. A unit change in inflations leads to 3.1 percent positive change in the dependent variable, the gdp. While and inverse relationship was revealed of trade balance and gdp in this model, a positive relationship was discovered between the dependent variable and nominal exchange rate in the long run. (Table 4).

Table 5. Long-Run PMG/ARDL Model Result (1)

Long Run Equation				
Variable	Coefficient	Std. Error	t-Statistic	Prob.*
INF	3.096	0.1328	23.3117	0.0000
LOGFX	0.5479	0.1373	3.9911	0.0001
LOGTB	-0.4134	0.0571	-7.2411	0.0000

***Source*:** Author's Computation from EViews11

In the short-run model, the error correction was seen to be significant showing positive impact of some independent variables on the gross domestic product. At base, one unit increase in inflation will result in a significant 2.9 percent impact on the dependent variable. In the second and third lag periods, it was observed to cause 1.5 percent increase, respectively. However, the impact in lag 3 was not significant. The nominal exchange rates showed an inverse relationship with the gross domestic product. It is noted that the foreign exchange rate impact was insignificant in the short-run also is the trade balance (Table 5)..

Table 5. Short-Run PMG/ARDL Model Result

Short Run Equation				
Variable	Coefficient	Std. Error	t-Statistic	Prob.*
COINTEQ01	-0.46767	0.18613	-2.51265	0.012801702
D(LOGGDP(-1))	-0.41662	0.18395	-2.2649	0.02463017
D(LOGGDP(-2))	-0.47219	0.20212	-2.33617	0.02050905
D(INF)	2.92388	1.17451	2.48944	0.01363998
D(INF(-1))	1.47984	0.72454	2.04245	0.042466743
D(INF(-2))	1.47581	0.70585	2.09083	0.037852382
D(INF(-3))	0.17299	0.22966	0.75327	0.452205825
D(LOGFX)	-0.92556	0.85697	-1.08004	0.281473259
D(LOGFX(-1))	0.55157	1.80942	0.30483	0.760820596
D(LOGFX(-2))	-1.3594	0.87176	-1.55937	0.120547406
D(LOGFX(-3))	0.40032	0.35145	1.13906	0.256090121
D(LOGTB)	0.05818	0.09681	0.60095	0.548580455
D(LOGTB(-1))	0.07563	0.08652	0.87412	0.383137562
D(LOGTB(-2))	0.1085	0.07839	1.38417	0.167903852
D(LOGTB(-3))	0.03249	0.08952	0.36295	0.717040913
C	-1.92268	2.39469	-0.8029	0.423022746
@TREND	0.1827	0.14669	1.2455	0.214458596

5. Conclusion

The study considered panel annual data from fifteen countries. From this group, fourteen of them have adopted inflation targeting. Nigeria is the only one amongst the group is just planning to adopt the monetary policy of targeting inflation. However, the inflation in the economy has been on rapid increase for the past three years. The study gathered the cross-sectional data the selected countries for the range of period 1991 to 2022. It employed pooled mean group auto-regressive distributed lag (PMG/ARDL) model to investigate the relationship between the regressants and the regressors. An apriori relationship between the inflation and the dependent variables was confirmed as its increase impact proportional positive increase on the gdp in both short-run and long-run. The nominal exchange rates showed an inverse relationship with the gross domestic product. It is noted that the foreign exchange rate impact was insignificant in the short-run but a positive relationship in the long-run. However, the trade balance posited an inverse relationship with dependent variable in the long-run but positive effect in the short-run. Should Nigeria adopt the inflation targeting policy and they plan? This calls for a more in-depth study to unravel the range of inflation rates that trigger productivity in the economy. This would inform the appropriate policy measure to adopt.

References

Belkhir, M. L., Ben Naceur, S., Candelon, B., Choi, W. G., & Mugrabi, F. (2023). *Macroprudential policy and bank systemic risk: Does inflation targeting matter?* SSRN Electronic Journal.

Debelle, G., Masson, P., & Savastano, M. (1998). *Inflation targeting as a framework for monetary policy.* International Monetary Fund.

Eshov, M. P., & Benetti, K. (2023). *Impact of monetary policy sustainability indicators on economic growth in transition to inflation targeting.* In *Liberec Economic Forum 2023*.

Ewurum, N. C., Kalu, C. U., & Nwankwo, D. J. (2017). Inflation targeting and economic growth nexus in Nigeria: Implications for monetary policy. *The International Journal of Academic Research in Business and Social Sciences, 7*(12), 12–27.

Fabayo, J. A., & Ajilore, O. T. (2006). Inflation: How much is too much for economic growth in Nigeria. *Indian Economic Review, 41*(2), 129–144.

Mohammed, D. (2021). Oil price, exchange rate, money supply and inflation in Nigeria: Further analysis using the VECM approach. *West African Journal of Monetary and Economic Integration*.

Musa, I., Magaji, S., & Salisu, A. Y. (2022). Monetary policy shocks and economic growth: Evidence from SVAR modelling. *International Journal of Indonesian Business Review*.

Saidu, B. Z. (2017). *Inflation targeting and economic growth in Nigeria* [Unpublished manuscript].

Truhlar, Z. (2023). *Impact of inflation targeting on a selected indicator of the real economy*. Doctoral Student Conference at the University of Finance and Administration.

On the relevance of post-UTME screening examination in Nigeria: A case study of the University of Benin, Benin City

C. O. Odijie and N. Ekhosuehi

Department of Statistics, University of Benin, Benin City

**Corresponding Author; E-mail: cyril.odijie@uniben.edu*

Abstract

To assess the relevance, or otherwise, of the so-called post-UTME screening examinations conducted by majority of the reputable universities in Nigeria over the years, this paper investigates the relationship between scores of candidates in Unified Tertiary Matriculation Examination (UTME) conducted yearly by the Joint Admissions and Matriculation Board (JAMB) and their scores in post-UTME screening examination of the University of Benin for three academic sessions. The well-known traditional statistical techniques of correlation and regression were employed to three data sets using the statistical software, SPSS and spreadsheets for the analysis. Results of the analysis showed correlation coefficients of 0.483, 0.585, and 0.255 between the UTME scores and post-UTME scores for the 2021/2022, 2022/2023 and 2023/2024 academic sessions, respectively, with respective R^2 *values of 0.234, 0.343 and 0.064 for the three sessions under regression. Also, the regression coefficients were statistically significant at 5% level (*$t = 10.774$, $p = 0.000$; $t = 14.203$, $p = 0.000$ *and* $t = 24.312$, $p = 0.000$*, respectively). Since all the correlations between UTME scores and post-UTME scores were weak or just moderate as in one of the datasets, and although the regression coefficients were statistically significant, there was no strong explanatory power of UTME scores on post-UTME screening scores for the three sessions/datasets considered as can be concluded from the* R^2 *values, the study affirms the relevance of post-UTME examination in Nigerian universities to a large extent using the Universisty of Benin as a case study.*

Keywords: Correlation, regression, UTME, post-UTME, significance

1. Introduction

Over the years, since its inception, there have been arguments for and against the relevance and continuity of post-UTME examinations (e.g. Umoru, 2017; Adedigba, 2018; and Lawal, 2021). A comprehensive review of these arguments can be seen in Ikoghode (2015). The fact that the number of prospective candidates for admission who meet the pass mark of 200 (out of a total of 400) set by JAMB is far greater than the number of students the institutions can adequately admit and the claim that the process of the UTME has been compromised to a large extent,

especially before the advent of the Computer Based Testing (CBT) for UTME in 2013 have also existed over the years. For example, examination malpractices, which have often led to withheld results or outright cancellation of results from some UTME centres, have been a widespread issue in the country prior to CBT (Tijani, 2015). This has also been observed by Umo and Ezeudu (2007). Even after the introduction of CBT by JAMB, there have still been cases of malpractices leading to withheld results in some centres or outright cancellation in recent years (e.g. Suleiman, 2023; Erunke, 2023; Babalola, 2024; Tolu-Kolawole, 2024). In the light of these reasons, among others, the Minister of Education in 2005 introduced a screening exercise to be done after UTME – a screening examination that was later adopted by many tertiary institutions in the country and came to be known as post-UTME screening examination/test. The tertiary institutions which conduct this screening examination or test today, base their final admission criterion on an aggregate score computed from the UTME scores and post-UTME screening scores of the candidates in relation to the cut-off marks set for the different courses of interest offered by the institutions. While the aggregate score is usually scaled to 100% across these institutions, the method of aggregation varies among the participating institutions. Some universities even include a weighted average of the grades obtained by candidates from their secondary school leaving eximination. Despite the differences in the components that make up the aggregate, they are all a linear combination of the scores from the respective constituent examination scores or grades that contribute to the aggregate score. For instance, UNIBEN, the case study of this paper, uses an aggregate of UTME score and post-UTME score such that both exams contribute exactly 50 points to the total aggregate of 100 points. That is, the aggregate score (say, A) is a linear function of the form

$$A = w_1 u + w_2 p \tag{1}$$

where, w_1 and w_2 are weights assigned to the UTME score, u and the post-UTME score, p, respectively, such that

$$w_1 u = w_2 p = 50 \tag{2}$$

Now, the total score obtainable in UTME (u) is 400, resulting from the four major courses required for entry into a specific course of study offered by the admitting institution and the total score obtainable in the post-UTME screening exam of UNIBEN (our case study) is 100.

The direct implication, from equation (2), is that $w_1 = 0.125$ or $\frac{1}{8}$ and $w_2 = 0.5$ or $\frac{1}{2}$. In other words, 12.5% of the UTME score of a candidate and 50% of the post-UTME score of the same candidate go into the aggregate score computed over a total of 100. So, for instance, a candidate who scores 345 (out of 400) in UTME and 75 (out of 100) in post-UTME will have an aggregate score of 80.625 (or 81 approximately).

Some proponents of the notion that post-UTME screening exams or tests are irrelevant and should be discontinued have seen it as an unnecessary duplication of the same process. This will imply that candidates performance in UTME will be in same direction as their performance in post-UTME screening exam (or the reverse case) and so there will no need to for the latter. If this claim or hypothesis is true, from a statistical point of view, this would suggest a high positive correlation between UTME scores and post-UTME scores of candidates or, put differently, a high predictive power of post-UTME scores by UTME scores under simple linear regression since passing UTME is a prerequisite for taking post-UTME exam in the first place. This forms the motivation for, and objective of, this paper which seeks to examine the nature of possible relationship between UTME and post-UTME scores with UNIBEN as a case study, and thereby revealing the relevance, or otherwise, of the post-UTME screening exam or test.

The remaining sections are outlined as follows: section 2 reviews some relevant literature on the study, section 3 presents the methods applied, results and discussions are presented in section 4 and the concluding remarks follow in section 5.

2. Literature Review

When two or more quantitative or numerical variables are taken on the same subject, correlation is usually suspected between at least a pair of variables involved. From simple height and weight data of individuals to other intricate variables such as DNA signatures, learning outcomes, body mass indexes, etc., studies have been carried out to examine the existence and nature of correlation in simple two- and multi-dimensional cases (e.g. Bewik et al., 2003; Crawford, 2006; O'Brien and Scott, 2012; Bezuidenhout and Domleo, 2013, among others). Correlation indicates the degree of association of two variables and it is a prerequisite for causality, since the latter can only exist usually when the former does. However, the converse is not true. To establish causality when association exists, regression comes into play, where one variable is expressed as linear combination of the other(s) in a statistical model. This

model can then be analyzed with prescribed techniques for overall fit of the data while seeking to establish the explanatory power of the independent variable(s) on the dependent variable.

Some studies involving the key variable of this study have been carried out in the literature. For example, Umo and Ezeudu (2007) in a conference paper, showed that UTME and post-UTME scores of students in the University of Nigeria, Nsukka (UNN) were poorly correlated and they recommended that screening exercises after UTME should be sustained. However, the study was simply based on Pearson product moment correlation coefficient measured on data taken across faculties in the institution without any recourse to testing for actual explanatory power of UTME scores on the screening scores obtained by the students. Imasuen and Ebuwa (2020) examined the predictive power of UTME and post-UTME on final grades of students from randomly selected departments in the university of Benin and reported that both exams had no significant predictive power on the final grades of students in the university. They recommended, among others, that every tertiary institution should be allowed to conduct their own screening examinations. Other related studies can be found in Omodara (2010), Eze (2014), Aina (2017) Afu and Ukofia (2017), among others.

3. Methodology

Data on 2020/2021, 2021/2022 and 2022/2023 UTME scores and post-UTME scores of candidates who applied to the University of Benin for admission and sat for the two exams, were collected from the Central Records Processing Unit (CRPU) in the ICT Unit of the University of Benin. The datasets consist of candidates' registration number, UTME score (recorded as "*jamb_score*"), post-UTME screening score (recorded as "*screening_score*") and aggregate score (recorded as "aggregate") in Comma Separated Values (CSV) format, among other variables. Since the data in our case study are bivariate data collected on each subject for some number of subjects under study, correlation and regression analysis form a good combination of statistical analysis techniques to examine the sort of relationship that may exist between the two variables of interest: UTME scores and post-UTME scores of prospective students of the University of Benin for the three academic sessions. A random sample of 383, 389 and 389 scores, respectively from the 2021/2022, 2022/2023 and 2023/2024, were randomly selected from the relevant columns of the datasets (*jamb_score* and *screening_score*) based on Taro Yamane's formula for selecting adequate sample from a finite population (Yamane, 1967). The population sizes of the raw dataset were 11110, 19812, and 17372 for the 2021/2022, 2022/2023 and 2023/2024 datasets, respectively. They reduced to 8688, 13106

and 13735, respectively, after cleaning incomplete records since there were candidates with *jamb_score* recorded but no screening_score recorded. More than one sessional dataset was considered so as to check for consistency of results, or otherwise. For each sessional dataset, summary statistics were first computed, then the Pearson correlation coefficient (r) was computed and finaly, regression anlysis was performed using Microsoft Excel spreadsheets and Statistical Package for the Social Sciences, SPSS.

The simple regression model for each dataset is of the form

$$scre\_score = \beta_0 + \beta_1(jamb\_score) + \varepsilon \quad (3)$$

where β_0 and β_1 are the regression parameters to be estimated, $\varepsilon \sim N(0, \sigma^2)$ is the random error and $scre\_score$ is the abbreviated form of the post-UTME screening score recorded originally as *screening_score*. While β_1 is the main parameter of interest, which measures how much change in *scre_score* will be brought about by a unit change in *jamb_score*, β_0 has been included to avoid loss of generality. In this case, it does not make a numerical sense, since there would not be any *scre_score* if *jamb_score* is zero (0), as its usual meaning would have suggested.

The results of the analysis and brief discussion of each of them are presented in the next section.

Results and Discussion

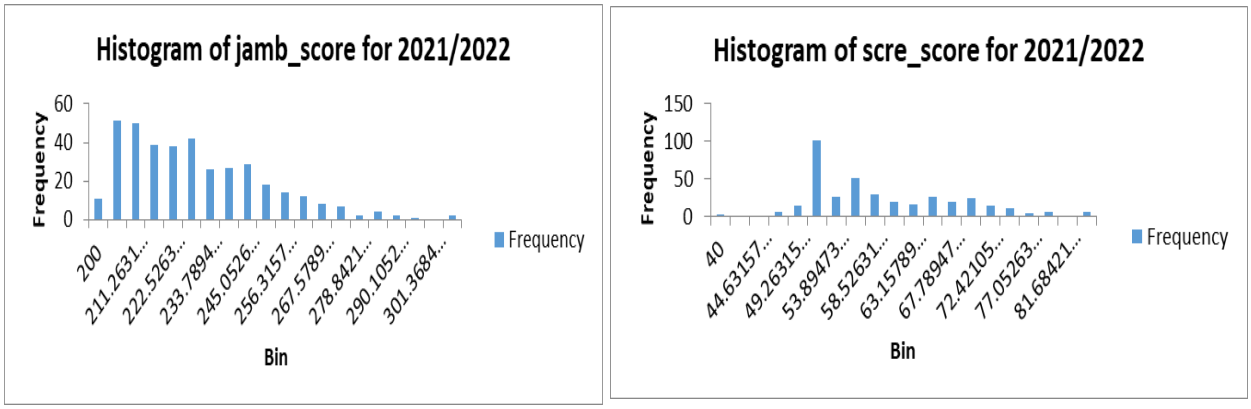
Table 1: Descriptive statistics of the random sample selected from the datasets for the three sessions.

Academic Session	Variable	Mean	Median	Mode	Standard Deviation	Min.	Max.	Skew.	Kurt.
2021/2022	*jamb_score*	226.64	223	211	20.96	200	307	0.95	0.67
	scre_score	57.93	56	50	8.97	40	84	0.71	-0.25
2022/2023	*jamb_score*	232.58	227	204	25.78	200	318	1.04	0.55
	scre_score	55.82	52	50	8.32	20	92	1.03	2.81
2023/2024	*jamb_score*	235.65	227	209	30.50	200	333	1.09	0.51
	scre_score	54.13	53.34	50	8.34	16.68	90	0.22	3.57

It can be seen clearly from Table 1 that the random sample of 383 candidates' scores drawn from the 2021/2022 dataset had a mean *jamb_score* (UTME score) of 226.64 and a mean *scre_score* (post-UTME screening score) of 57.93. The UTME scores were fairly positively skewed toward lower scores around the UNIBEN set minimum of 200 for qualification for

post-UTME screening exam. The median and modal *jamb_score*s are, respectively, 223 and 211 from the sample. The modal value and the moderate right skewness of the sample can also be noticed easily in Figure 1(a).

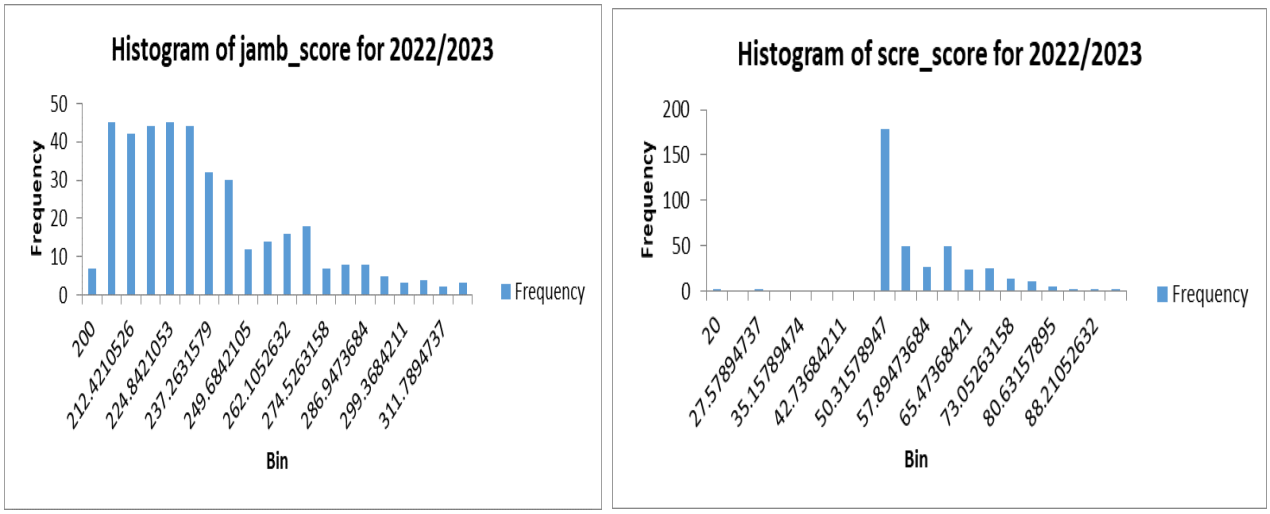
In a similar manner, the *scre_scores* were also moderately positively skewed (see Figure 2(b)).



(a) **(b)**

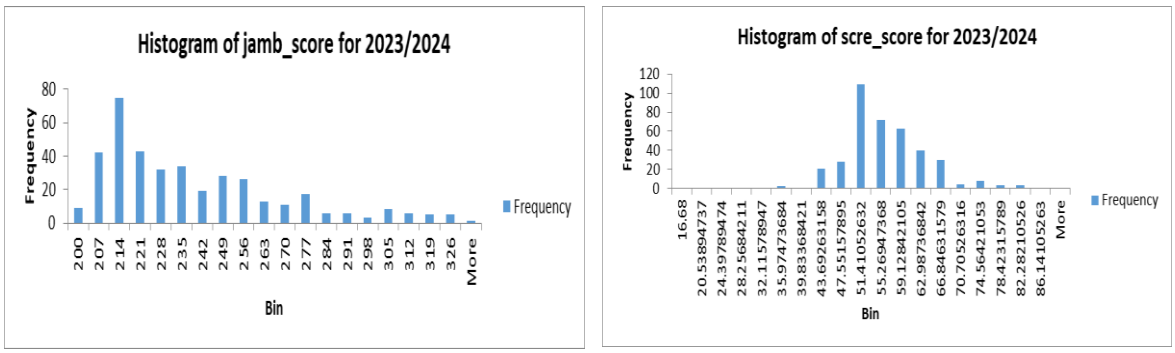
Figure 1: Histograms of the random sample selected from 2023/2024

The samples taken from the 2022/2023 and 2023/2024 sessions/datasets exhibited more positive skewness except for the *scre_scores* of 2023/2024 that had the closest coefficient of skewness to zero (i.e. 0.22). The mean *jamb_score*s were 232.58 and 235.65 in the two sessions, respectively, while the median *jamb_score*s were the same (i.e. 227) for both sessions. The modal *scre_score* stood at 50 across all the three samples drawn randomly from the three sessional datasets. Refer to Table 1 again and Figure 3(b).



(a) (b)

Figure 2: Histograms of the random sample selected from 2023/2024 dataset.



(a) **(b)**

Figure 3: Histograms of the random sample selected from 2023/2024 dataset.

Table 2: Regression model summary of the datasets for the three academic sessions.

Regression statistic	2021/2022 dataset	2022/2023 dataset	2023/2024 dataset
R	0.483	0.585	0.255
R^2	0.234	0.343	0.065
Adjusted R^2	0.232	0.341	0.063
Std. Error of the Estimate	7.864	6.751	8.070
Sample size (n)	383	389	389

The summaries of the regression analysis displayed in Table 2 immediately reveal the poor explanatory or predictive power of *jamb_score* on *scre_score* since, in each of the three sessions, the coefficient of determination (R^2) was small, with the highest among them just being 0.341 for 2022/2023 dataset. This clearly implies that in all the three sessions considered, the highest explanation of the variability in post-UTME scores of candidates in the University of Benin by their UTME scores constitutes merely 34%, leaving the remaining 66% of the variations to other unknown sources and chance. Also the correlations coefficients (R) were small in each case, indicating a weak linear relationship between UTME and post-UTME scores of the candidates under study.

Table 3: Regression coefficients and their statistics for the datasets in the three academic sessions.

Academic Session	Regression Coefficient	Estimated Value	Std. Error	t-Stat.	*p*-value	95% CI for β	
						Lower	Upper
2021/2022	β_0	11.050	4.369	2.529	0.012	2.460	19.641
	β_1 *	0.207	0.012	10.774	0.000	0.169	0.245
2022/2023	β_0	11.909	3.111	3.829	0.000	5.793	18.024
	β_1 *	0.189	0.013	14.203	0.000	0.163	0.215
2023/2024	β_0	37.721	3.192	11.817	0.000	31.445	43.997
	β_1 *	0.063	0.003	24.312	0.000	0.043	0.096

* β_1 (the slope in each case) is the main parameter of interest in this study.

All the coefficients (β's) of the regression models were significant, indicating the existence of linear relationship between UTME scores and post-UTME scores which has been seen to be weak, however, from the results in Table 2. The significance of the overall fit is also alluded by the analysis of variance whose results are shown in Table 4. Again, this only shows that the linear regression model is a good fit for the data and that there is a significant linear relationship between the datasets under study. The main indicator of explanatory strength still remains the coefficient of determination, R^2, which has a consistently small value over the three datasets, as already seen. The choice of linear regression as good model for the datasets is also revealed by the residual plots which shows no pattern but a random in the variability of the residuals random along the horizontal axis as expected for a linear regression and the estimated residuals exhibiting the desirable property of zero mean value and constant variance (Figures 4, shown only for 2021/2022 dataset. Others are similar).

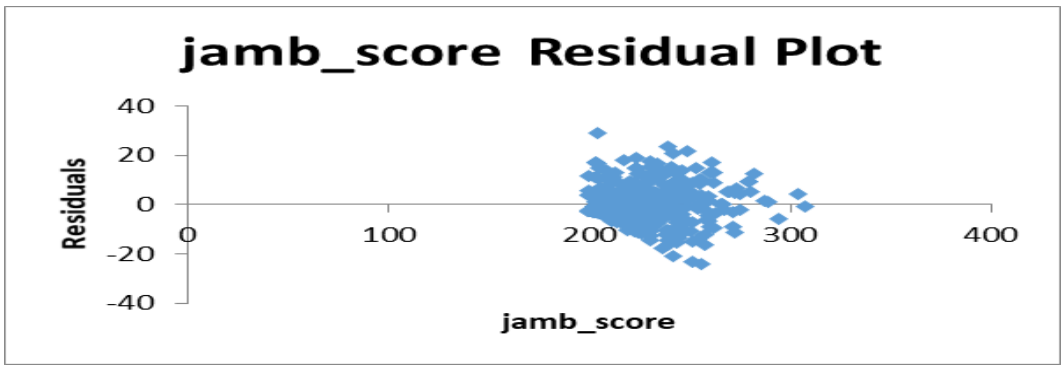


Figure 4: Residual plot of jamb_score showing random variability along the horizontal axis (mean = 0)

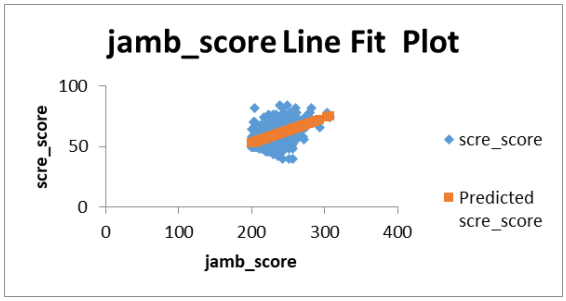
From Table 3, we can write the fitted regression models (lines) for each of the three datasets as follows.

For 2021/2022: $scre\_score = 11.05 + 0.207(jamb\_score)$ (4)

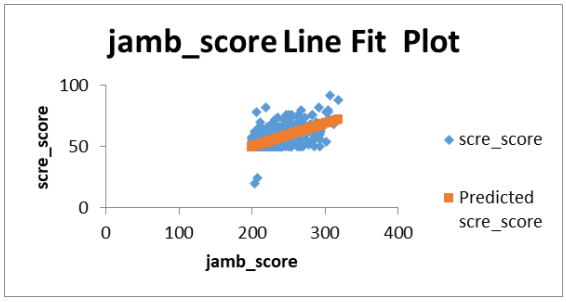
For 2022/2023: $scre\_score = 11.909 + 0.189(jamb\_score)$ (5)

For 2023/2024: $scre\_score = 37.721 + 0.063(jamb\_score)$ (6)

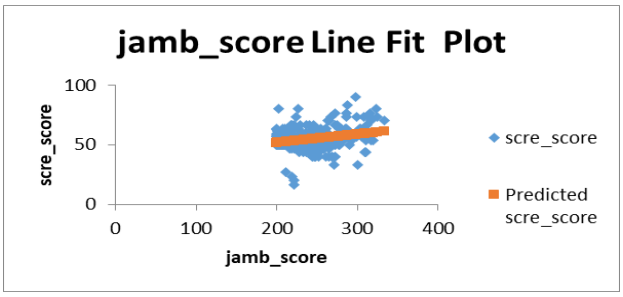
The fitted lines are in shown in red in Figure 5 (a), (b) and (c).



(a)



(b)



(c)

Figure 5: Scatterplots of the samples from three datasets with fitted line embedded

Table 4: ANOVA of the datasets for the three academic sessions.

Academic Session	Source of Variation	df	SS	MS	F	*p-value*
2021/2022	Regression	1	7179.86	7179.86	116.10	0.00
	Residual	381	23562.09	61.84		
	Total	382	30741.95			
2022/2023	Regression	1	9193.95	9193.95	201.73	0.00
	Residual	387	17637.45	45.57		
	Total	388	26831.40			
2023/2024	Regression	1	1750.74	1750.74	26.88	0.00
	Residual	387	25204.51	65.13		
	Total	388	26955.26			

If the study had observed that UTME (*jamb_score*) has a high or good predictive power for post-UTME (*scre_score*), then equations (4), (5) and (6) would have served as tools for

predicting the post-UTME score of any candidate, given their UTME score. The prediction will be less precise in this case of poor predictive power of UTME scores on post-UTME scores if the equations are used in this case.

4. Summary and Conclusion

This paper set out to examine the relevance, or otherwise, of post-UTME screening exam or test by assessing the nature of relationship between the screening exam and the prerequisite exam, UTME. Three datasets were obtained, cleansed and a sample was taken from each population for the analysis. The results have shown that UTME scores have no strong predictive capability for post-UTME scores, although there exists a significant linear relationship between them as should be expected. This immediately signals the fact that post-UTME exams are relevant in the final assessment of candidates for admission suitability. Since, by the results of the analysis, a candidate who performs well in UTME may not necessarily perform well in post-UTME and a student who performs poorly in UTME may end up performing well in post-UTME (both cases for whatever reasons), it therefore makes enough sense to continue both exams side by side and assess candidates for admission on aggregate performances in the two exams, as is the status quo.

References

Adedigba, A. (2018, March 3). Post-UTME's are illegal, stop now, court rules. Premium Times. https://www.premiumtimesng.com/news/top-news/260505-post-utmes-illegal-stop-now-court-rules.html?tztc=1

Aina, B. B (2017). UTME and post-UTME scores as predictor of academic performance of university undergraduates of Adekunle Ajasin University, Akungba-Akoko, Ondo State, Nigeria. *International Journal of Education and Evaluation*. 3(1), 37-43.

Babalola, O. (2024, June 1). Revisiting concerns over 2024 UTME result. The Guardian. https://www.guardian.ng/news/revisiting-concerns-over-2024-utme-result/

Bewick, V., Cheek, L. & Ball, J. (2003). Statistics Review 7: Correlation and regression. *Crit Care.* 7(451), 451-459. https://doi.org/10.1186/cc2401

Bezuidenhout, C. N. and Domleo, R. R. (2013). A demonstration of correlation graphs to human body
Dimensions. *Scientific Research and Essays*. 8(27), 1273-1281. DOI 10.5897/SRE2013.5398

Crawford S. L. (2006). Correlation and Regression. *Circulation. Volume 114, number 19. https://doi.org/101161/CIRCULATIONAHA.105.586495*

Erunke, J (2023, July 3). JAMB raises alarm over candidates parading fake scores. Vanguard. https://www.vanguardngr.com/2023/07/jamb-raises-alarm-over-candidates-parading-fake-scores/

Eze, E. C. (2014). University Matriculation Examination as a predictor of students' final grades in the Faculty of Health Sciences and Technology of University of Nigeria, Nsukka. Master's Thesis, University of Nigeria, Nsukka.

Ikoghode, A. (2015). Post-UTME screening in Nigrian universities: how relevant today?. *International Journal of Education and Research.* 3(8), 101-116

Imasuen, K. and Ebuwa, S. O. (2020). Unified Tertiary Matriculation Examination (UTME) and the Post Unified Tertiary Matriculation Examination (PUTME) as predictors of undergraduate students' final grades. *International Journal of Interdisciplinary Research Methods.* 7(3), 48-60

Lawal, I. (2022, July 22). Stakeholders discuss relevance-continuity of post-UTME. The Guardian. https://www.guardian.ng/stakeholders-discuss-relevance-continuity-of-post-utme/

O'Bien, D. and Scott, P. S. (2012). Correlation and Regression. In Approaches to Quantitative Research – A Guide for Dissertation Students, Ed, Chen, H, Oak Tree Press.

Omodara, F. M. (2010). Assessment of the predictive validity of UME scores on academic performances of science university undergraduates. *Multidisciplinary Journal of Research Development.* 14(1), 66-71

Suleiman, Q. (2023, July 5). FLASHBACK: Five times JAMB caught candidates who falsified their UTME results. Premium Times. https://www.premiumtimesng.com/news/top-news/608129-flashback-five-times-jamb-caught-candidates-who-falsified-their-utme-results.html?tztc=1

Tijani, M (2015, March 9). How prepared is JAMB for full-fledged CBT? The Cable. https://www.thecable.ng/prepared-jamb-full-fledged-cbt/amp/

Tolu-Kolawole, D. (2024, April 24). 2024 UTME: JAMB withholds results of 64,624 candidates. Punch. https://punchng.com/2024-utme-jamb-withholds-results-of-64624-candidates/

Umo, U. and Ezeudu, S. A. (2007). Relationship between University Matriculation Examination scores and the screening scores at the University of Nigeria, Nsukka (UNN). *IAEA 33rd annual conference paper (online at* https://iaea.info/documents/relationship-between-university-matriculation-examination-scores-and-the-screening-scores-at-the-university-of-nigeria-nsukka-unn/)

Umoru, H. (2017, October 4). Senate moves to scap post-UTME. Vanguard. https://www.vanguardngr.com/2017/10/senate-moves-to-scrap-post-utme/

Yamane, T. (1967). Statistics, An Introductory Analysis, 2nd Ed., New York: Harper and Row.

A new lifetime distribution derived from the less popular paralogistic distribution via transformed-transformer method: Statistical properties and applications

Sunday A. Osagie[1] and Joseph E. Osemwenkhae[1]

[1]Department of Statistics, University of Benin, Benin City, Nigeria

**Corresponding Author E-mail:* sunday.osagie@uniben.edu

Abstract

The paper introduces a novel, flexible probability distribution developed through the Transformed-Transformer method. By deploying the lesser-known paralogistic distribution, the potential for innovative modeling in distribution theory is demonstrated. The proposed distribution exhibits desirable properties, including flexibility and uniqueness, offering a compelling alternative to existing models. Important statistical properties of the proposed distribution are derived and a Monte-Carlo simulation study is carried out to evaluate the performance of the distribution. The results of the application of the proposed distribution in comparison with some existing distributions in literature illustrate excellent fits and predictive capabilities in modeling real-world phenomena, particularly in lifetime modeling. They highlight the usefulness and flexibility of the new distribution, positioning it as a valuable addition to the existing lifetime distributions.

Keywords: Transformed-Transformer method, Paralogistic distribution, Simulation study, Distribution theory, Lifetime modeling.

1. Introduction

In the last three decades, classical lifetime distributions have played a crucial role in developing new models to analyse the time-to-failure of complex components, systems, and living organisms. While popular distributions like the exponential, normal, Weibull, loglogistic and Burr XII distributions are widely used to model monotonic failure data or generate more flexible distributions to handle the nonmonotonic nature of the failures, there exist other less popular distributions that can better capture the underlying mechanisms of certain phenomena. One less popular distribution is the paralogistic distribution, which has garnered limited attention in lifetime modeling due to its one shape parameter. The limited popularity of the distribution and lack of extensive research have hindered its widespread acceptance.

The paralogistic distribution was introduced as a submodel of the generalized gamma family (McDonald, 1984). The cumulative distribution and density functions of the one-parameter paralogistic distribution are given as

$$R(t) = 1 - (1 + t^{\theta})^{-\theta},\ t > 0, \theta > 0 \quad (1)$$

and

$$r(t) = \theta^2 (1 + t^{\theta})^{-\theta-1}. \quad (2)$$

Despite the limitation and limited literature of the paralogistic distribution as a statistical model, it can be applied in various fields to analyze the lifetime of components and systems with complex failure mechanisms in reliability engineering, the survival times of living organisms with non-linear hazard rates in survival analysis and extreme events and heavy-tailed behavior in financial data.

This paper aims to develop a new probability distribution by applying the transformed-transformer (T-X) method to the paralogistic distribution, enabling more accurate and flexible modeling of real-world phenomena than the paralogistic and Burr XII distributions. The properties of the new distribution such as the probability density function, cumulative distribution function, moments, entropy measures, and mean deviations are investigated. The flexibility of the new distribution will be compared with existing distributions, including the paralogistic distribution, in terms of goodness-of-fit and accuracy, and potential applications to real data sets will be explored. Ultimately, the new distribution addresses the limitations of the paralogistic and Burr XII distributions, and provide a more robust and versatile model for real-life phenomena than the two distributions.

The motivation behind the development of the new distribution from the paralogistic distribution via the T-X method is based on the following reasons.

i. The development of the new distribution can lead to new insights into reliability and survival analysis.
ii. The need to combine the features of the paralogistic and Burr XII distributions to develop a more flexible distribution.
iii. Show the flexibility and usefulness of the new distribution in lifetime analysis through application to real-life phenomena.

The organisation of the paper is presented as follows. Section 2 considered the review of the literature of the paralogistic and Burr XII distributions. Section 3 presents the construction of the Paralogistic-X family and Paralogistic-Burr XII (PBXII) distribution is introduced as a submodel. In Section 4, statistical properties of the new distribution are derived. Also, the maximum likelihood estimation and simulation study for the new distribution are presented. In section 5, application of the PBurr XII distribution is presented to illustrate its usefulness and flexibility, by comparison with some lifetime distributions. Section 6 presents the conclusion

of the paper.

2. Literature Review

Recently, Idemudia and Ekhosuehi (2019) introduced the three-parameter paralogistic distribution by adding a location parameter to the two-parameter distribution. They derived the properties of the distribution and applied it to some data sets. Pizon and Arcede (2022) developed the inverse paralogistic distribution based on the Farla-Gumbel-Morgenstern copula and derived the moment properties of the distribution. Wasinat and Choopradit (2023) introduced the paralogistic-negative binomial distribution. Choopradit and Wasinat (2023) developed the Poisson-Paralogistic distribution. Marasigan (2023) extended the paralogistic distribution by developing the three-parameter gamma inverse paralogistic distribution. Anghel and Ilinca (2023) used the paralogistic distribution as one of the four distributions to predict future flood risks in climate change. They concluded that the paralogistic distribution is a suitable model for frequency analysis.

The Burr XII distribution (Burr, 1942) in its classical and modified forms has been prominently applied in lifetime modeling for two decades. Bhatti et al. (2021) introduced the three-parameter unit generalized log Burr XII distribution with flexible hazard rates. They developed the bivariate and multivariate versions for the proposed distribution. They compared the distribution to other existing distributions. The results from the application showed that the proposed distribution outperformed the other distributions. Guerra et al. (2017) proposed the four-parameter Zografos-Balakrishnan Burr XII (ZBXII) distribution. They showed the flexibility of the proposed distribution by applying to some datasets in comparison to some competing distributions. It was seen that the ZBXII distribution produced better fits to the other distributions. Yari and Tondpour (2018) introduced the Burr XII-Exponential distribution and investigated the shapes of the hazard function using MLE and some estimation methods at different sets of parameter values. Gad et al. (2019) developed the Burr XII-Burr XII distribution and derived some properties of the proposed distribution. They considered the characterization study and application of the distribution. It was concluded that the distribution that the new distribution was more flexible than some compared distributions. Aslam et al. (2021) introduced the generalized Burr XII family of distributions and derived the statistical properties. They investigated the performance of the generalized Burr XII distribution by comparing it with some existing distributions. It was concluded that the proposed distribution performed better the the competing distributions. Afzal et al. (2023) introduced the modified

Burr XII distribution as a more flexible distribution than the Burr XII distribution and applied it to some data sets in literature. The results obtained should that the proposed distribution gave better fits to the datasets than the other competing models.

3. Methodology

Alzaatreh et al. (2013) expressed the cumulative distribution function (cdf) and probalility density function (pdf) of a T −X family as

$$G_{T-X}(x;\theta,\varepsilon) = \int_{c}^{W[F(x;\varepsilon)]} r(t;\theta)dt = R(W[F(x;\varepsilon)];\theta),\ x>0 \qquad (3)$$

where $\varepsilon > 0$ is the parameter vector of the baseline distribution, $F(x;\varepsilon)$ is the cdf of the baseline distribution and $W[F(x;\varepsilon)]$ is a function of $F(x;\varepsilon)$.

Let $W[F(x;\varepsilon)] = \dfrac{F(x;\varepsilon)}{1-F(x;\varepsilon)}$ and r(t) is the pdf of the one-parameter paralogistic distribution in (2), then

$$G_{T-X}(x;\theta,\varepsilon) = \int_{c}^{\frac{F(x;\varepsilon)}{1-F(x;\varepsilon)}} \theta^2(1+t^\theta)^{-\theta}dt = 1-\left(1+\left(\frac{F(x;\varepsilon)}{1-F(x;\varepsilon)}\right)^\theta\right)^{-\theta},\ x>0 \qquad (4)$$

defines the cdf of the new T-X generator, which will be known as the Paralogistic-*X* family of distributions.

Suppose $F(x;\varepsilon)$ is the cdf of the two-parameter Burr XII distribution, which is defined as

$$F(x;\alpha,\lambda) = 1-(1+x^\alpha)^{-\lambda},\ x>0,\ \alpha,\lambda>0$$

is substituted into the generator in (4), the cdf of the proposed Paralogistic-Burr XII (PBXII) distribution is given as

$$G(x;\theta,\alpha,\lambda) = 1-(1+((1+x^\alpha)^\lambda-1)^\theta)^{-\theta},\ x>0,\ \theta,\alpha,\lambda>0$$

(5)

where $\theta,\alpha,\lambda > 0$ are shape parameters.

The corresponding pdf of (5) is given as

$$g(x;\theta,\alpha,\lambda) = \theta^2\alpha\lambda x^{\alpha-1}(1+x^\alpha)^{\lambda-1}((1+x^\alpha)^\lambda-1)^{\theta-1}(1+((1+x^\alpha)^\lambda-1)^\theta)^{-\theta-1}.$$

(6)

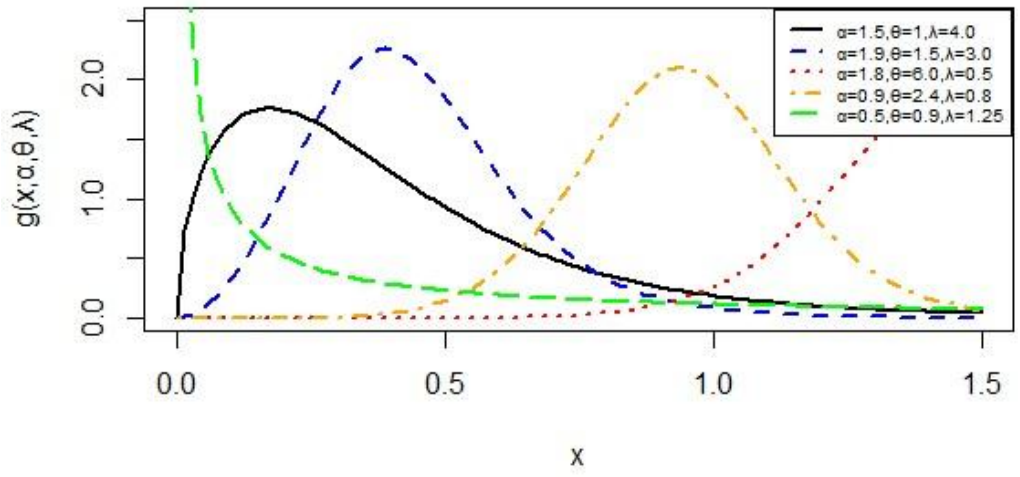


Figure 1: The shapes of the pdf of the PBXII distribution

The survival and hazard functions of (5) are given as

$$\overline{G}(x;\theta,\alpha,\lambda) = (1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{-\theta}$$

and

$$hz(x;\theta,\alpha,\lambda) = \theta^2\alpha\lambda x^{\alpha-1}(1+x^{\alpha})^{\lambda-1}((1+x^{\alpha})^{\lambda}-1)^{\theta-1}(1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{-1}.$$

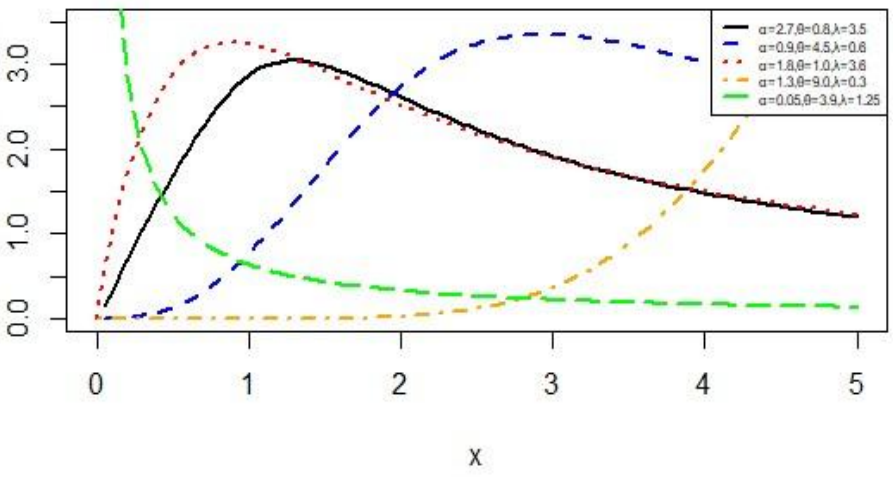


Figure 2: The shapes of the hazard function of the PBXII distribution

3.1 Series expansion of the density function of the PBXII distribution

For a non-integer $m > 0$, the following power series

$$(1+x)^m = \sum_{i=0}^{\infty} \binom{m}{i} x^i \quad \text{and} \quad (1+x)^{-m} = \sum_{i=0}^{\infty} \binom{m+i-1}{i} (-1)^i x^i$$

(7)

will be useful for expanding the pdf of the PBXII distribution. The series expansion of the density function becomes necessary in the less rigorous derivations of some statistical properties. The pdf of the PBXII distribution is given as

$$g(x;\theta,\alpha,\lambda) = \theta^2 \alpha\lambda x^{\alpha-1}(1+x^{\alpha})^{\lambda-1}((1+x^{\alpha})^{\lambda}-1)^{\theta-1}(1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{-\theta-1}.$$

Employing the power series in (7),

$$g(x;\theta,\alpha,\lambda) = \theta^2 \alpha\lambda \sum_{i=0}^{\infty} \binom{\theta+i}{i} (-1)^i x^{\alpha-1}(1+x^{\alpha})^{\lambda-1}((1+x^{\alpha})^{\lambda}-1)^{\theta(i+1)-1}.$$

Further simplification gives

$$g(x;\theta,\alpha,\lambda) = \theta^2 \alpha\lambda \sum_{i,j=0}^{\infty} \binom{\theta+i}{i} \binom{\theta(i+1)-1}{j} (-1)^{i+j} x^{\alpha-1}(1+x^{\alpha})^{-\lambda(j-\theta(i+1))-1}. \quad (8)$$

The series expansion in (8) will be employed to obtain statistical properties such as the raw and incomplete moments, moment generating function and mean deviations for the PBXII distribution.

3.2 Statistical properties of the Paralogistic-Burr XII distribution

This section considers the derivations of the statistical properties of the PBXII distribution such as quantile function, noncentral (raw) moments, incomplete moments, moment generating function, probability-weighted moments, order statistic, entropies and mean deviations.

3.2.1 Quantile function

Quantile function is used to obtain random numbers which are used in simulation study of real-life phenomena. For $0 < p < 1$, the quantile function for a random variable X following the PBXII distribution is obtained by inverting (6) as

$$Q(p) = G^{-1}(p) = p$$

where G^{-1} is the inverse of the cdf of the lifetime distribution. It follows that

$$1-(1+((1+x_p^{\alpha})^{\lambda}-1)^{\theta})^{-\theta} = p$$

Further manipulation gives the quantile function of the PBXII distribution as

$$Q(p) = \left(\left(\left((1-p)^{-1/\theta}-1\right)^{1/\theta}+1\right)^{1/\lambda}\right)^{1/\alpha}. \quad (9)$$

Table 1 gives the values for the quantile function of the PBXII distribution at different sets of parameter values.

Table 1: Values of quantile function of the PBXII distribution for $(\theta, \alpha, \lambda)$

	(0.7,1.3,2.5)	**(6,0.4,0.5)**	**(3,0.9,2)**	**(6.3,1.5,4)**	**(0.9,0.8,0.6)**
0.1	0.06596041	1.860260	0.1241864	0.2308399	0.1087760
0.2	0.15986078	2.732630	0.1617449	0.2472415	0.3580352
0.3	0.28322388	3.496387	0.1910303	0.2580454	0.8224087
0.4	0.44360319	4.240802	0.2173581	0.2667026	1.6888075
0.5	0.65442763	5.015284	0.2429874	0.2742690	3.4027257
0.6	0.94140230	5.868864	0.2697455	0.2814729	7.1817756
0.7	1.36045550	6.874721	0.2996818	0.2888040	17.1790398
0.8	2.06232307	8.183665	0.3368126	0.2969640	54.0035476
0.9	3.67746949	10.266158	0.3929832	0.3076941	348.2520829

For $p = 0.5$ in (9), the median of the PBXII distribution is given as

$$Q(0.5) = \left(\left(\left(2^{1/\theta}-1\right)^{1/\theta}+1\right)^{1/\lambda}\right)^{1/\alpha}.$$

3.2.2 Raw, incomplete and probability weighted moments

Moments are the summary measures of the distribution of a random variable describing the lifetimes of systems. The measures are essential in statistics and probability distribution theory, allowing in-depth analysis and understanding of the distribution.

3.2.2.1 Raw moments

Let X be a random variable following the pdf in (6), the r^{th} (r=1(1)m, m is a positive integer) raw moments of the PBXII distribution is defined as

$$\mu_r = E(X^r) = \int_0^{\infty} x^r g(x;\theta,\alpha,\lambda)dx. \tag{10}$$

(see Afzal et al., 2023).

Substituting the series expansion in (8) into (10),

$$\mu_r = \theta^2\alpha\lambda \sum_{i,j=0}^{\infty} \binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\int_0^{\infty} x^{r+\alpha-1}(1+x^{\alpha})^{-\lambda(j-\theta(i+1))-1}dx$$
$$= \theta^2\alpha\lambda \sum_{i,j=0}^{\infty} \binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}\frac{(-1)^{i+j}}{\alpha}\mathrm{B}\left(1+\frac{r}{\alpha}, \lambda(j-\theta(i+1))-\frac{r}{\alpha}\right),$$

where $\frac{1}{q}\mathrm{B}\left(\frac{u}{q}, V-\frac{u}{q}\right) = \int_0^{\infty} y^{u-1}(1+y^q)^{-v}dy$ (Gradshteyn and Ryzhik, 2007).

Hence, the r^{th} raw moments of the PBXII distribution is given as

$$\mu_r = \theta^2\lambda \sum_{i,j=0}^{\infty} \binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\mathrm{B}\left(1+\frac{r}{\alpha}, \lambda(j-\theta(i+1))-\frac{r}{\alpha}\right). \tag{11}$$

Table 2 presents the the r^{th} (r=1(1)4) raw moments and other related quantities such as the standard deviation, coefficient of variation, coefficient of skewness and coefficient of kurtosis.

Table 2: Raw moments and other related quantities at values of (θ,α,λ)

	(3.0,6.0,9.0)	(1.8,0.3,1.5)	(0.2,0.5,0.7)	(0.9,3.0,0.6)	(4.0,0.8,0.05)
E(X)	0.39086906	0.06021071	0.014682027	0.2165798	0.03752323
E(X^2)	0.15331514	0.03454306	0.007431809	0.1619939	0.02911993
E(X^3)	0.06033520	0.02419869	0.004972769	0.1287180	0.02366398
E(X^4)	0.02381818	0.01862272	0.003736120	0.1064824	0.01986468
SD	0.02316287	0.17583438	0.084948499	0.3392449	0.16646904
CV	0.05925993	2.92031714	5.785883591	1.5663740	4.43642641
CS	-0.83396291	3.38379693	7.588393825	1.1213943	4.44197272
CK	4.40590641	14.12961717	66.319789033	2.5641682	21.55466214

3.2.2.2 Incomplete moments

The r^{th} incomplete moments of a random variable X following the pdf in (6) is defined as

$$\mu_{r,t} = E(X^r / X > t) = \frac{1}{\overline{G}(t;\theta,\alpha,\lambda)} \int_t^{\infty} x^r g(x;\theta,\alpha,\lambda) dx. \qquad (12)$$

Substituting (5) and (8) into (12)

$$\mu_{r,t} = \theta^2 \alpha\lambda (1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{\theta} \sum_{i,j=0}^{\infty} \binom{\theta+i}{i} \binom{\theta(i+1)-1}{j} (-1)^{i+j} \int_{t.}^{\infty} x^{r+\alpha-1} (1+x^{\alpha})^{-\lambda(j-\theta(i+1))-1} dx.$$

Letting $y = (1+x^{\alpha})^{-1}$,

$$\mu_{r,t} = \theta^2 \alpha\lambda (1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{\theta} \sum_{i,j=0}^{\infty} \binom{\theta+i}{i} \binom{\theta(i+1)-1}{j} (-1)^{i+j} \int_{0.}^{(1+t^{\alpha})^{-1}} y^{\lambda(j-\theta(i+1))-r/\alpha-1} (1-y)^{r/\alpha} dy.$$

If $\mathrm{B}_t(u,v) = \int_0^t y^{u-1}(1-y)^{v-1} dy$, therefore

$$\mu_{r,t} = \theta^2 \alpha\lambda (1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{\theta} \sum_{i,j=0}^{\infty} \binom{\theta+i}{i} \binom{\theta(i+1)-1}{j} (-1)^{i+j} \mathrm{B}_{(1+t^{\alpha})^{-1}} \left(1+\frac{r}{\alpha}, \lambda(j-\theta(i+1))-\frac{r}{\alpha}\right)$$

(13)

defines the r^{th} incomplete moments of the PBXII distribution.

3.2.2.2 Probability weighted moments

The probability weighted moments, $PWM_{r,l,m}(X)$, for a random variable X following the pdf in (6) is defined as

$$PWM_{r,l,m} = E(X^r G^l(X) \overline{G}^m(X)) = \int_0^{\infty} x^r G^l(x) \overline{G}^m(x) g(x;\theta,\alpha,\lambda) dx, \qquad (14)$$

where r, l, m>0 are positive integers.

Substituting (5), (6) and the survival function into (14), the $PWM_{r,l,m}(X)$ of the PBXII distribution is obtained as

$$PWM_{r,l,m} = \theta^2 \alpha\lambda \int_0^{\infty} x^{r+\alpha-1} (1+x^{\alpha})^{\lambda-1} ((1+x^{\alpha})^{\lambda}-1)^{\theta-1} (1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{-\theta(m+1)-1}$$
$$(1-(1+((1+x^{\alpha})^{\lambda}-1)^{\theta})^{-\theta})^{l} dx \quad .$$

Utilizing (7),

$$PWM_{r,l,m} = \theta^2 \alpha\lambda \sum_{i=0}^{l} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \binom{l}{k} \binom{\theta(m+k+1)+i}{i} \binom{\theta(i+1)-1}{j} (-1)^{i+j} \int_0^{\infty} x^{r+\alpha-1} (1+x^{\alpha})^{-\lambda(j-\theta(i+1))-1} dx$$

.

Hence, the $PWM_{r,l,m}(X)$ of the PBXII distribution is given as

$$PWM_{r,l,m} = \theta^2\lambda\sum_{i=0}^{l}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{l}{k}\binom{\theta(m+k+1)+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\mathrm{B}\left(1+\frac{r}{\alpha},\lambda(j-\theta(i+1))-\frac{r}{\alpha}\right).$$

(15)

Remarks

(1) If $l = m = 0$, then

$$PWM_{r,0,0} = E(X^r) = \theta^2\lambda\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\mathrm{B}\left(1+\frac{r}{\alpha},\lambda(j-\theta(i+1))-\frac{r}{\alpha}\right).$$

(2) If $m = 0$, then

$$PWM_{r,l,0} = E(X^r G^l(X)) = \theta^2\lambda\sum_{i=0}^{l}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{l}{k}\binom{\theta(k+1)+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}$$
$$\mathrm{B}\left(1+\frac{r}{\alpha},\lambda(j-\theta(i+1))-\frac{r}{\alpha}\right).$$

3.2.3 Moment generating function

The moment generating function ($M_X(t)$) for a random variable *X* following the pdf in (6) is obtained as

$$M_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} g(x;\theta,\alpha,\lambda)dx,\ t > 0. \qquad (16)$$

Substituting the expansions of e^{tx} and (8) into (16) gives

$$M_X(t) = \theta^2\alpha\lambda\sum_{k=0}^{\infty}\frac{t^k}{k!}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\int_0^{\infty}x^{k+\alpha-1}(1+x^{\alpha})^{-\lambda(j-\theta(i+1))-1}dx\right]$$
$$= \theta^2\lambda\sum_{k=0}^{\infty}\frac{t^k}{k!}\left[\sum_{i,j=0}^{\infty}\binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\mathrm{B}\left(1+\frac{k}{\alpha},\lambda(j-\theta(i+1))-\frac{k}{\alpha}\right)\right].$$

(17)

Equation (17) defines the moment generating function of the PBXII distribution.

3.2.4 Order statistics

Let $X_{(m)}, m = 1(1)n$ be the m^{th} order statistics of $X_1, X_2, \cdots, X_m$ of m variables, then the pdf of the m^{th} order statistics for the PBXII distribution can be defined as

$$g_{i:m}(x;\theta,\alpha,\lambda)=\frac{1}{\mathrm{B}(i,m-i+1)}[G(x;\theta,\alpha,\lambda)]^{i-1}[1-G(x;\theta,\alpha,\lambda)]^{m-1}g(x;\theta,\alpha,\lambda)$$

(18)

Replacing $G(x;\theta,\alpha,\lambda)$ with $1-\overline{G}(x;\theta,\alpha,\lambda)$ in (18) gives

$$g_{i:m}(x;\theta,\alpha,\lambda)=\frac{1}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\binom{i-1}{k}(-1)^k[\overline{G}(x;\theta,\alpha,\lambda)]^{m+k-i}g(x;\theta,\alpha,\lambda)$$

Substituting (6) and the survival function into the right-hand side of $g_{i:m}(x;\theta,\alpha,\lambda)$ gives

$$\begin{aligned}g_{i:m}(x;\theta,\alpha,\lambda)&=\frac{\theta^2\alpha\lambda}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\binom{i-1}{k}(-1)^k x^{\alpha-1}(1+x^\alpha)^{\lambda-1}((1+x^\alpha)^\lambda-1)^{\theta-1}\\&\quad(1+((1+x^\alpha)^\lambda-1)^\theta)^{-\theta(m+k-i+1)-1}\\&=\frac{\theta^2\alpha\lambda}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\sum_{j=0}^{\infty}\binom{i-1}{k}\binom{\theta(m+k-i+1)+j}{j}(-1)^{k+j}x^{\alpha-1}(1+x^\alpha)^{\lambda-1}\\&\quad((1+x^\alpha)^\lambda-1)^{\theta(j+1)-1}.\end{aligned}$$

After some algebraic simplifications, the pdf of the m^{th} order statistics of the PBXII distribution is given as

$$\begin{aligned}g_{i:m}(x;\theta,\alpha,\lambda)&=\frac{\theta^2\alpha\lambda}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\binom{i-1}{k}\binom{\theta(m+k-i+1)+j}{j}\binom{\theta(j+1)-1}{l}(-1)^{k+j+l}x^{\alpha-1}\\&\quad(1+x^\alpha)^{\lambda(\theta(j+1)-l)-1}.\end{aligned}$$

(19)

The r^{th} raw moments of the m^{th} order statistics of the PBXII distribution for *r=1(1)n* is given as

$$\begin{aligned}\mu_{r,i:m}=E(X_{i:m}^r)&=\frac{\theta^2\alpha\lambda}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\binom{i-1}{k}\binom{\theta(m+k-i+1)+j}{j}\binom{\theta(j+1)-1}{l}(-1)^{k+j+l}\\&\quad\int_0^\infty x^{r+\alpha-1}(1+x^\alpha)^{\lambda(\theta(j+1)-l)-1}dx.\end{aligned}$$

Hence, the r^{th} raw moments of the m^{th} order statistics of the PBXII distribution is defined as

$$\begin{aligned}\mu_{r,i:m}=E(X_{i:m}^r)&=\frac{\theta^2\lambda}{\mathrm{B}(i,m-i+1)}\sum_{k=0}^{i-1}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\binom{i-1}{k}\binom{\theta(m+k-i+1)+j}{j}\binom{\theta(j+1)-1}{l}(-1)^{k+j+l}\\&\quad\mathrm{B}\left(1+\frac{r}{\alpha},\lambda(l-\theta(j+1))-\frac{r}{\alpha}\right).\end{aligned}$$

3.2.5 Entropy

Entropy is an important measure of information of randomness associated with a random

variable following a probability distribution. The measure has con-tributed to many fields of science where lifetime analysis is important. This section will derive the expression of the Rényi entropy of the PBXII distribution.

The Rényi entropy (Rényi, 1961) of the PBXII distribution is obtained using

$$I_R(a) = \frac{1}{(1-a)} \log\left(\int_0^\infty g^a(x;\theta,\alpha,\lambda)dx \right), a > 0, a \neq 1. \tag{20}$$

Substituting (6) into (20),

$$I_R(a) = \frac{1}{(1-a)} \log\left((\theta^2\alpha\lambda)^a \int_0^\infty x^{a(\alpha-1)}(1+x^\alpha)^{a(\lambda-1)}((1+x^\alpha)^\lambda - 1)^{a(\theta-1)}(1+((1+x^\alpha)^\lambda - 1)^\theta)^{-a(\theta+1)} dx \right).$$

After some expansions,

$$I_R(a) = \frac{1}{(1-a)} \log\left((\theta^2\alpha\lambda)^a \sum_{i=0}^\infty \sum_{j=0}^\infty \binom{a(\theta+1)+i-1}{i} \binom{\theta(a+i)-a}{j} (-1)^{i+j} \int_0^\infty x^{a(\alpha-1)}(1+x^\alpha)^{-((a+\lambda j)-\lambda\theta(a+i))} dx \right).$$

Hence, the Rényi entropy of the PBXII distribution can be expressed as

$$I_R(a) = \frac{a}{(1-a)} \log(\theta^2\alpha\lambda) + \frac{1}{(1-a)} \log\left[\sum_{i=0}^\infty \sum_{j=0}^\infty \binom{a(\theta+1)+i-1}{i} \binom{\theta(a+i)-a}{j} (-1)^{i+j} \frac{\mathrm{B}\left(\frac{a(\alpha-1)+1}{\alpha}, ((a+\lambda j)-\lambda\theta(a+i)) - \frac{a(\alpha-1)+1}{\alpha} \right)}{\alpha} \right]. \tag{21}$$

3.2.6 Mean deviations

Mean deviations (MD) are measures of the variations in a value in a data set from the mean and median. The mean deviations about the mean and median for any random variable X following the pdf defined in (6) are given as

$$MD_\mu = E(|X-\mu|) = 2\mu G(\mu) - 2\int_0^\mu x g(x;\theta,\alpha,\lambda)dx = 2\mu G(\mu) - 2\left(\mu - \int_\mu^\infty x g(x;\theta,\alpha,\lambda)dx \right)$$

and

$$MD_{Q(0.5)} = E(|X - Q(0.5)|) = \int_0^{Q(0.5)} |x - Q(0.5)| g(x;\theta,\alpha,\lambda)dx = -\mu + 2\int_{Q(0.5)}^\infty x g(x;\theta,\alpha,\lambda)dx.$$

Employing (8), the mean deviations of the PBXII distribution are defined as

$$MD_{\mu} = 2\mu\left(-(1+((1+\mu^{\alpha})^{\lambda}-1)^{\theta})^{-\theta} - \frac{\theta^2\lambda}{\mu}\sum_{i,j=0}^{\infty}\binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j} \right.$$
$$\left. \mathrm{B}_{(1+\mu^{\alpha})^{-1}}\left(1+\frac{1}{\alpha}, \lambda(j-\theta(i+1))-\frac{1}{\alpha}\right)\right)$$

and

$$MD_{Q(0.5)} = -\mu + 2\theta^2\lambda\sum_{i,j=0}^{\infty}\binom{\theta+i}{i}\binom{\theta(i+1)-1}{j}(-1)^{i+j}\mathrm{B}_{(1+Q(0.5)^{\alpha})^{-1}}\left(1+\frac{1}{\alpha}, \lambda(j-\theta(i+1))-\frac{1}{\alpha}\right)$$

respectively.

3.2.6 Simulation Study and Parameter Estimation

In this study, a simulation study is carried out to investigate the sensitivity of the values of the parameters of the PBXII distribution. This is to ascertain the potentiality of using the new model in real-life events. The maximum likelihood estimation method is considered to obtain parameter estimates of the PBXII distribution.

3.2.6.1 Simulation Study

In the simulation study of the PBXII distribution, the MLEs, bias and mean square errors (MSEs) are computed as performance metrics at different sets of values of the parameters of the PXII distribution. For a large N of 1000 samples, sample sizes of n = 25, 50, 100, 200, 400, 800 are generated using the quantile function of the PBXII distribution. The choice of a large N value is to achieve a true sampling distribution of randomly generated data based on the MLE estimates of the parameters of the distribution. The sets of parameter values considered for the simulation study are $(\theta, \alpha, \lambda) = \{(1.5, 0.08, 2.0), (2.5, 1.5, 0.2)\}$. The performance metrics for the sets of parameter values are presented in Table 3.

Table 3: Simulation study for PBXII distribution

Λ	**N**	$\theta = 1.5, \alpha = 0.08, \lambda = 2.0$			$\theta = 2.5, \alpha = 1.5, \lambda = 0.2$		
		MLE	**Bias**	**MSE**	**Mean**	**Bias**	**MSE**
θ	25	1.39644	-0.10356	0.09825	2.68516	0.18516	2.11368
	50	1.25145	-0.24855	0.15762	2.74960	0.24960	0.97348
	100	1.34245	-0.15755	0.09612	2.44683	-0.05317	0.67160
	200	1.38875	-0.11125	0.07318	2.64157	0.14157	0.48917
	400	1.43047	-0.07953	0.05031	2.45827	0.03233	0.09104
	800	1.50774	0.00774	0.00572	2.47285	-0.02715	0.04444

α	25	0.09611	0.01611	8.77438e-02	1.96150	0.46150	1.31192
	50	0.09669	0.01669	8.03264e-02	1.52856	0.02856	0.06298
	100	0.09318	0.01318	6.03915e-02	1.60672	0.10672	0.04904
	200	0.08915	0.00915	3.11984e-02	1.51825	0.01825	0.01720
	400	0.08535	0.00535	2.39644e-02	1.50651	0.00651	0.00386
	800	0.08007	0.00007	1.52667e-02	1.50106	0.00106	0.00191
λ	25	2.78342	0.78342	2.23254	0.23363	0.03363	0.01401
	50	3.26034	1.26034	3.56497	0.19585	-0.00415	0.00212
	100	2.73843	0.73843	2.00557	0.21508	0.01508	0.00288
	200	2.53371	0.53371	1.21666	0.19608	-0.00392	0.00155
	400	2.37856	0.37856	0.79718	0.19770	-0.00230	0.00036
	800	2.01089	0.01089	0.02520	0.20452	0.00452	0.00024

The results presented in Table show that the bias and MSEs for the sets of parameter values decrease towards zero as the sample sizes increase. It implies that the MLE method is an appropriate method to estimate the parameters of the PBXII distribution.

3.2.6.2 Maximum likelihood estimation (MLE) method

Let random variable $X \sim$ PXII distribution having $F = (\theta, \alpha, \lambda)$ as parameter vector. For a random sample of complete observations given as $x_1, x_2, \cdots, x_n$, the likelihood function of the random sample is given as;

$$L_i(F) = \prod_{i=1}^{n} g(x_i; \theta, \alpha, \lambda) = \prod_{i=1}^{n} (\theta^2 \alpha \lambda x_i^{\alpha-1} (1 + x_i^{\alpha})^{\lambda-1} ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta-1} (1 + ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta})^{-\theta-1})$$

(22)

The corresponding total log-likelihood function of (22) is

$$L_n = n \log(\theta^2 \alpha \lambda) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) + (\lambda - 1) \sum_{i=1}^{n} \log(1 + x_i^{\alpha}) + (\theta - 1) \sum_{i=1}^{n} \log((1 + x_i^{\alpha})^{\lambda} - 1)$$
$$- (\theta + 1) \sum_{i=1}^{n} \log(1 + ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta}).$$

Equating the first partial derivatives of L_n with respect to $F = (\theta, \alpha, \lambda)$ to zero will give the following system of nonlinear equations, which is

$$\frac{\partial L_n}{\partial \theta} = \frac{2n}{\theta} + \sum_{i=1}^{n} \log((1 + x_i^{\alpha})^{\lambda} - 1) - (\theta + 1) \sum_{i=1}^{n} \frac{((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta} \log((1 + x_i^{\alpha})^{\lambda} - 1)}{(1 + ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta})} = 0,$$

$$\frac{\partial L_n}{\partial \alpha} = \frac{2n}{\alpha} + \sum_{i=1}^{n} \log(x_i) + (\lambda - 1)\sum_{i=1}^{n} \frac{x_i^{\alpha} \log(x_i)}{(1 + x_i^{\alpha})} + \lambda(\theta - 1)\sum_{i=1}^{n} \frac{x_i^{\alpha}(1 + x_i^{\alpha})^{\lambda - 1} \log(x_i)}{(1 + x_i^{\alpha})}$$
$$- \theta\lambda(\theta + 1)\sum_{i=1}^{n} \frac{x_i^{\alpha}(1 + x_i^{\alpha})^{\lambda - 1}((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta - 1} \log(x_i)}{(1 + ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta})} = 0$$

and

$$\frac{\partial L_n}{\partial \lambda} = \frac{n}{\lambda} + \sum_{i=1}^{n} \log(1 + x_i^{\alpha}) + (\theta - 1)\sum_{i=1}^{n} \frac{(1 + x_i^{\alpha})^{\lambda} \log(1 + x_i^{\alpha})}{((1 + x_i^{\alpha})^{\lambda} - 1)}$$
$$- \theta(\theta + 1)\sum_{i=1}^{n} \frac{(1 + x_i^{\alpha})^{\lambda}((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta - 1} \log(1 + x_i^{\alpha})}{(1 + ((1 + x_i^{\alpha})^{\lambda} - 1)^{\theta})} = 0.$$

The resulting system of equations is solved iteratively with an iterative method to obtain the parameter estimates of the PBXII distribution. For the computation of the parameter estimates, the *AdequacyModel* package in R programming language will be used.

For statistical inference for $F = (\theta, \alpha, \lambda)$ of the PBXII distribution to be achieved, the asymptotic confidence intervals are derived for the parameter estimates using the Fisher's information matrix. These statistics obtained from the matrix are used to obtain approximated confidence intervals for the parameters in F. The total observed information matrix, $I_n(F)$ is given as

$$I_n(F) = -\begin{pmatrix} L_{n(\theta\theta)} & L_{n(\theta\alpha)} & L_{n(\theta\lambda)} \\ - & L_{n(\alpha\alpha)} & L_{n(\alpha\lambda)} \\ - & - & L_{n(\lambda\lambda)} \end{pmatrix}.$$

Hence, $\hat{\theta} \pm Z_{\Psi/2}\sqrt{\hat{L}_{n(\theta\theta)}}$, $\hat{\alpha} \pm Z_{\Psi/2}\sqrt{\hat{L}_{n(\alpha\alpha)}}$ and $\hat{\lambda} \pm Z_{\Psi/2}\sqrt{\hat{L}_{n(\lambda\lambda)}}$ give the $100(1-\Omega)$ confidence intervals for $F = (\theta, \alpha, \lambda)$, where $Z_{\Psi/2}$ is the standard normal upper percentile.

4. Results and Discussion

In this section, two datasets are presented to demonstrate the capability and versatility of the PBXII distribution in lifetime modeling in comparison with the logistic Burr XII (LBXII)(Guerra et al., 2023), Marshall-Olkin Burr XII (MOBXII)(Afify et al., 2020), Marshall-Olkin Lomax (MOLx)(Afify et al., 2020), Burr XII (BXII) and paralogistic (Paral) distributions. The discrepancy criteria to compare these distributions are the log-likelihood function, Akaike information criterion (AIC), Corrected Akaike information

criterion (CAIC), Bayesian information criterion (BIC), Hann-Quin information criterion (HQIC) and Kolmogorov-Smirnov test. The best-fitted fitted distribution exhibits the lowest values of the discrepancy criteria.

The two data sets are secondary data sets are presented as follow. The first data set is the confirmed cases of Lassa fever for 67 weeks in Edo State from 30th December, 2019 to 7th June, 2020 (Chinagorom et al., 2021). The data is given as;

9, 3, 31, 24, 20, 10, 6, 11, 20, 23, 8, 4, 6, 2, 1, 2, 3, 2, 2, 2, 2, 4, 2, 4, 1, 5, 4, 4, 2, 7, 2, 2, 3, 3, 3, 7, 5, 2, 6, 2, 5, 2, 5, 3, 2, 1, 7, 18, 33, 34, 36, 39, 41, 22, 31, 25, 11, 8, 3, 5, 6, 1, 2, 3, 2, 2, 2.

The second data set contains the active repair time (in hrs.) for the airborne communication transceivers given in Jorgensen (2012). The data is given as;

0.50, 0.60, 0.60, 0.70, 0.70, 0.70, 0.80, 0.80, 1.00, 1.00, 1.00, 1.00, 1.10, 1.30, 1.50, 1.50,1.50, 1.50, 2.00, 2.00, 2.20, 2.50, 2.70, 3.00, 3.00, 3.30, 4.00, 4.00, 4.50, 4.70, 5.00, 5.40, 5.40, 7.00, 7.50, 8.80, 9.00, 10.20, 22.00, 24.50.

Tables 3 and 4 show the parameter estimates of the competing distributions and their corresponding values for the discrepancy criteria for the data sets.

Table 4: Parameter estimates and discrepancy criteria for first data set

Distributions	α	θ	λ	**-2ll**	**AIC**	**CAIC**	**BIC**	**HQIC**	**KS (p-value)**
PBXII	4.36273 (2.00748)	1.61153 (0.15771)	0.08274 (0.03667)	**406.5613**	412.5613	412.9423	419.1754	415.1785	**0.14312 (0.1285)**
LBXII	2.04662 (1.26207)	0.34290 (0.18754)	2.71109 (0.55503)	411.2943	417.2943	417.6752	423.9084	419.9115	0.15589 (0.0771)
MOBXII	8.67964 (8.29385)	0.14863 (0.14569)	5.60117 (2.31943)	406.8733	412.8733	413.2542	419.4874	415.4905	0.14293 (0.1294)
MOLx	5.05733 (33.68824)	0.35381 (2.35678)	22.44266 (9.97216)	421.6852	427.6852	428.0661	434.2992	430.3024	0.15513 (0.0795)
BXII	1.00116e02 (51.36523)	6.13168e-02 (0.00309)	-	422.9856	428.9856	427.1731	431.3950	428.7304	0.28686 (3.252e-06)
Paral	0.58817 (0.04885)	-	-	441.3382	443.3382	443.3998	445.5429	444.2106	0.70744 (<2.2e-16)

Table 5: Parameter estimates and discrepancy criteria for the second data set

Distributions	α	θ	λ	-2ll	AIC	CAIC	BIC	HQIC	KS (p-value)
PBXII	7.86188 (3.75213)	0.45732 (0.17389)	0.58777 (0.26397)	**180.6650**	186.6650	187.3317	191.7316	188.4969	**0.09417 (0.8701)**
LBXII	0.54238 (0.65016)	1.08633 (0.33713)	4.89628 (5.15441)	182.9347	188.9347	189.6013	194.0013	190.7666	0.10283 (0.7915)
MOBXII	2.70284 (0.80883)	0.44273 (0.23259)	1.49162 (1.01285)	181.6261	187.6261	188.2927	192.6927	189.4580	0.10593 (0.7605)
MOLx	28.84711 (20.16304)	0.07204 (0.29996)	11.27250 (6.14503)	186.7040	192.7040	193.3706	197.7706	194.5359	0.11363 (0.6799)
BXII	2.97281 (0.67766)	0.33827 (0.08707)	-	181.9732	185.9732	186.2975	189.3510	187.1945	0.10329 (0.7869)
Paral	0.70965 (0.08117)	-	-	208.1291	210.1291	210.2343	211.8179	210.7397	0.19512 (9.54e-05)

From Tables 4 and 5, the PBXII distribution exhibits better fits over the competing distributions because it exhibits the lowest values of the discrepancy criteria. Hence, the PBXII distribution is more flexible than the competing distributions.

5. Summary and Recommendation

The PBXII distribution is introduced as a new distribution in lifetime modeling. The statistical properties of the new distributions are derived, while the simulation study and maximum likelihood estimation for parameters of the distribution are presented. The usefulness and versatility of the proposed distribution are illustrated by the application to two real-life data sets. It is shown that the new distribution is a flexible model for lifetime data analysis.

It is recommended that the PBXII distribution be considered as an alternative model for the analysis of skewed or heavy-tailed failure data found in reliability engineering, survival analysis and lifetime data modeling.

Acknowledgements

The authors wish to thank the Editor and the anonymous reviewers for their constructive comments, which have greatly improved the contents of the paper.

Declaration of interest: None

References

Afify, A. & Abdellatif, A. (2020). The Extended Burr XII Distribution: Properties and Applications. J. of Nonlinear Sci. and Appl, 13, 133-146.

Afzal, F., Kausar, S., Qamar, M. S. & Akbar, S. (2023). Modified Burr XII Distribution: Properties and Applications. Asian Journal of Probability and Statistics, 21(2), 1-15.

Alzaatreh, A., Lee, C. & Famoye, F. (2013). A New Method for Generating Families of Continuous Distributions. Metron, 71(1), 63-79.

Anghel, C. G. & Ilinca, C. (2023). Predicting Future Flood Risks in the Face of Climate Change: A Frequency Analysis Perspective. Water, 15(22), 3883.

Aslam, M., Usman, R. M. & Raqab, M. Z. (2021). A New Generalized Burr XII distribution with Real Life Applications. Journal of Statistics and Management Systems, 24(3), 521-543.

Bhatti, F. A., Ali, A., Hamedani, G., Korkmaz, M. C. & Ahmad, M. (2021). The Unit Generalized Log-Burr XII distribution: Properties and application. AIMS Mathematics.

Burr, I. W. (1942). Cumulative frequency functions. Annals of Mathematical Statistics, 13,
215-232.

Chinagorom, N., Chekwube, B. & Adachukwu, E. (2021). Statistical Distribution of Lassa Fever in Edo State, Nigeria. Asian Journal of Probability and Statistics, 15(4), 88-96.

Choopradit, B. & Wasinrat, S. (2023). The Poisson-Paralogistic Distribution: Properties and Applications. Journal of Applied Science and Emerging Technology, 22(1), e249656-e249656.

Gad, A. M., Hamedani, G. G., Salehabadi, S. M. & Yousof, H. M. (2019). The Burr XII-Burr XII Distribution: Mathematical Properties and Characterizations. Pakistan Journal of Statistics.

Gradshteyn, I. S. & Ryzhik, I.M. (2007). Table of Integrals, Series, and Products. Elsevier.

Guerra, R. R., Pena-Ramırez, F. A. & Cordeiro, G. M. (2017). The Gamma Burr XII Distributions: Theory and Applications. Journal of Data Science, 15, 467-494.

Guerra, R. R., Pena-Ramırez, F. A. & Cordeiro, G. M. (2023). The Logistic Burr XII Distribution: Properties and Applications to Income Data. Stats, 6(4), 1260-1279.

Idemudia, R. & Ekhosuehi, N. (2019). A New Three-Parameter Paralogistic distribution: its properties and application. Journal of Data Science, 17(2), 239-257.

Jorgensen, B. (2012). Statistical Properties of the Generalized Inverse Gaussian distribution (Vol. 9). Springer Science and Business Media.

Marasigan, A. E. (2023). A New Extension of the Inverse Paralogistic Distribution using Gamma Generator with Application. Mindanao Journal of Science and Technology, 21(1).

McDonald, J. B. (1984). Some Generalized Functions for the Size Distribution of Income. In Modeling Income Distributions and Lorenz curves (pp. 37-55). New York, NY: Springer New York.

Pizon, M. & Arcede, J. P. (2022). Product and Quotient of Inverse Paralogistic distribution generated based on Farlie-Gumbel-Morgenstern (FGM) Copula. Innovative Journal of Mathematics (IJM), 1(3), 1-14.

Rényi, A. (1961). On Measures of Entropy and Information. In Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (Vol. 4, pp. 547-562). University of California Press.

Wasinrat, S. and Choopradit, B. (2023). The Paralogistic-Negative Binomial Distribution and Its Application. The 17th International Days of Statistics, September 7-9 2023, 540-549.

Yari, G. & Tondpour, Z. (2018). Estimation of the Burr XII-Exponential Distribution Parameters. Applications and Applied Mathematics. An International Journal (AAM), 13(1), 4.