CHARTERED INSTITUTE OF STATISTICIANS OF NIGERIA
(Publishers of the Journal of the Chartered Institute of Statisticians of Nigeria)
CISON Act 2022

PROCEEDINGS OF THE 2025 CISON CONFERENCE

Volume 2

Pre-Conference Workshop

22–23 September 2025

Dome Event Centre, Asaba, Delta State, Nigeria

Main Conference

24–26 September 2025

Dome Event Centre, Asaba, Delta State, Nigeria

Published by the Chartered Institute of Statisticians of Nigeria (CISON)

EDITORIAL DISCLAIMER AND STATEMENT OF RESPONSIBILITY

This document is the official Proceedings of the 2nd Annual International Conference of the Chartered Institute of Statisticians of Nigeria (CISON), held from 22nd to 26th September 2025, at the Dome Event Centre, Asaba, Delta State, Nigeria. The conference brought together statisticians, researchers, academics, and professionals from across Nigeria and the wider African region to explore how big data and analytics can be harnessed to build a robust national ecosystem that supports sustainable growth and national prosperity. The conference also provided a platform for robust deliberations on emerging trends and innovations in statistical science and practice during the scientific sessions.

The articles published in this document were presented during the scientific sessions and they represent the scholarly contributions and viewpoints of the authors. Thus, 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

1. Dr Godday Uwawunkonye Ebuh, President/Chairman
2. Professor Shehu Usman Gulumbe, Vice-President
3. Mrs Ngozi Theresa Agboegbulem, Registrar/Secretary
4. Dr Umaru Baba Council, Member
5. Professor Julian Ibezimako Mbegbu, Council Member
6. Dr (Mrs.) Uchenna Ogoke, Council Member
7. Mr Bello Ja'afaru, Council Member
8. Mr Stephen Aiyedun, Council Member
9. Dr Bright Ajibade, Council Member
10. Dr Michael Kalu Mba, Council Member & Representative CBN
11. Mrs Titilayo Hammed, Council Member & Representative NPC
12. Mr Augustine Chuks Anyakorah, Council Member & Representative NBS
13. Dr Kenneth Kwujeli, Council Member & Representative Ministry of Budget & Economic Planning
14. Mr Matthews Ofodum Nganjiozor, Council Member & Representative Federal Ministry of Education
15. Mrs Theresa U. Ogike, Council Member & Representative Head of Civil Service of the Federation
16. Professor Polycarp Emeka Chigbu, Editor-in-Chief, Journal of the CISON
17. Dr Ngozi Victor Atoi, Managing Editor, Journal of the CISON
18. Dr OlaOluwa Simon Yaya, Associate Editor, Journal of the CISON

FOREWORD

The 2nd Annual International Conference of the Chartered Institute of Statisticians of Nigeria (CISON) was successfully held from 24th to 26th September 2025 at the Dome Event Centre, Asaba, Delta State, Nigeria. The conference convened statisticians, researchers, academics, and professionals from across Nigeria and other parts of Africa to examine the growing role of big data and analytics in building a coordinated national ecosystem capable of driving sustainable growth and enhanced national prosperity.

The theme of the conference focused on harnessing big data and advanced analytics as strategic tools for national development. Discussions throughout the plenary and technical sessions underscored the importance of strengthening institutional frameworks, fostering cross-sector collaboration, and investing in data infrastructure and capacity development to position Nigeria competitively in the evolving global data landscape.

Ahead of the main conference, a two-day Pre-Conference Workshop was organised on the 22nd and 23rd of September 2025 at the Delta State Government House. The workshop, titled *Big Data Analytics and Machine Learning: Methods and Applications*, provided participants with practical exposure to contemporary analytical techniques and machine learning methods. It attracted statisticians, data scientists, analysts, and early-career professionals seeking to deepen their technical competencies and apply advanced methods across diverse sectors, including finance, health, governance, and industry.

During the scientific sessions, high-quality papers were presented by scholars and practitioners from Nigeria and across Africa, stimulating robust discussions on innovation, methodological advancement, and the practical application of data science in addressing development challenges. We commend the dedication of the authors, reviewers, workshop facilitators, and the Conference Planning Committee for their invaluable contributions. This Book of Proceedings stands not only as a record of scholarly engagement but also as a catalyst for continued research, collaboration, and innovation in statistical science and data analytics.

(Sgd.)

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

PREFACE

The central theme of the 2nd Annual International Conference of the Chartered Institute of Statisticians of Nigeria (CISON) focused on advancing big data and analytics as foundational pillars for building a coherent national ecosystem capable of driving enhanced national prosperity. The theme provided a platform for in-depth discussions on data infrastructure development, institutional collaboration, methodological innovation, and the integration of advanced analytics into evidence-based policymaking and strategic planning.

The papers contained in this volume were presented during the scientific sessions of the conference and have been accepted for publication in this proceedings. These scholarly contributions reflect diverse methodological approaches and practical applications of data science and statistical analytics to contemporary economic and development challenges. Collectively, they offer valuable insights for government institutions, development partners, academic communities, and private sector stakeholders.

It is our expectation that the contributions compiled herein will serve not only as a record of current research and professional engagement but also as a catalyst for sustained dialogue, interdisciplinary collaboration, and innovation within the statistical and data science communities.

On behalf of the Editorial Board of the JCISON, I express sincere appreciation to all authors, reviewers, discussants, and members of the Conference Planning Committee for their dedication and commitment to the success of this international conference and its proceedings.

(Sgd.)

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

MESSAGE FROM THE MANAGING EDITOR

This publication reflects the Institute's sustained commitment to advancing statistical science, big data analytics, and professional collaboration within Nigeria and across Africa. The central theme of the 2025 conference underscored the critical role of data-driven systems, institutional coordination, and analytical innovation in supporting sustainable development and informed policymaking.

The insights shared through keynote lectures, technical sessions, and workshop engagements are captured in this volume. The papers featured herein present both theoretical reflections and applied research spanning areas such as economic analysis, governance, digital transformation, finance, health, agriculture, education, and emerging technologies. Collectively, they demonstrate the expanding frontiers of data science and its relevance to national development.

The publication process was guided by editorial oversight to ensure the quality, integrity, and scholarly merit of the contributions. I extend my sincere appreciation to the authors for their intellectual rigour, the reviewers and discussants for their constructive engagement, and the editorial and technical teams for their dedication in bringing this volume to fruition.

It is our hope that this work will stimulate continued research, deepen professional collaboration, and further strengthen the role of statistics, big data, and analytics in shaping a more resilient and prosperous Nigeria.

(Sgd.)

Dr Ngozi V. Atoi
Managing Editor, JCISON

COMMUNIQUÉ FROM THE 2ND ANNUAL INTERNATIONAL CONFERENCE OF CHARTERED INSTITUTE OF STATISTICIANS OF NIGERIA (CISON), HELD AT THE DOME EVENT CENTRE, ASABA, DELTA STATE, FROM 22ND TO 26TH SEPTEMBER, 2025

Introduction

The 2nd Annual International Conference organized by the Chartered Institute of Statisticians of Nigeria (CISON), in collaboration with Delta State Government, was held in Asaba, Delta State, Nigeria from 22nd to 26th of September, 2025. A Pre-Conference workshop was organized on the 22nd and 23rd, while the main conference was held between 24th and 26th of September. The conference brought together Local and International participants from Academia, Industry, Finance, Economy, Government and Public Service Sectors, to deliberate on the Theme:

"Big Data and Analytics in Nigeria: Developing a National Ecosystem for Enhanced National Prosperity".

The pre-conference workshop recorded 186 registered participants, while the conference proper, which was in hybrid mode, witnessed the participation of 535 physical and 54 virtually registered participants. The opening ceremony which was held on Thursday, 25th of September, witnessed the presence of dignitaries from Government, including His Excellency, The Executive Governor of Delta State, Rt. Hon. (Elder) Sheriff F. O. Oborevwori, ably represented by the Secretary to the State Government, Dr. Kingsley Emuh, and the Delta State Commissioner for Economic Planning, Mr. Sunny A. Ekedayan. The Federal Minister for Education, the Statistician General of the Federation and many other principal officers and industry players were present or represented.

Objectives of the Conference

Today, as Nigeria grapples with complex challenges from youth unemployment and agricultural transformation to urbanization and climate change, the Nation needs more than intuition and political rhetoric. We need the precision of statistical analysis, the insight of predictive modeling, and the clarity that comes from evidence-based decision-making.

Therefore, this conference sought to:

- Explore the features and potentials of Big data Analytics towards enhanced National prosperity in Nigeria.
- Emphasize that Big data analytics offers unprecedented opportunities to understand patterns in our economy, predict market trends, optimize resource allocation, and measure the true impact of government programmes.

- Promote the move from reactive governance to proactive leadership, and from policy assumptions to policy precision.

Deliberations

The conference was designed to have two segments, namely, the Pre-Conference and the Conference sessions.

The pre-conference had four intensive and interactive training sessions ably piloted by acclaimed resource persons. Participants were taken through the meaning, theory and applications of Big Data Analytics, Artificial Intelligence and Machine Learning.

The Main conference witnessed the Scientific and Technical sessions, the opening ceremony, the plenary and syndicate sessions, as well as a Panel of discussants on the theme and sub-themes of the conference.

Papers presented from research efforts of participants were considered largely to be of International standard, explored statistical theory and gave insight into emerging applications of Big Data Analytics, Artificial Intelligence and Machine Learning within the Nigerian and global context.

The plenary sessions had four guest speakers from within and outside Nigeria. Their presentations prominently featured an experience-sharing presentation on the imperatives of big data ecosystem based on case studies of selected African countries by the African Development Bank (AfDB). Another highpoint of this session was a presentation on addressing the challenges of Agriculture through Data Analytics and Artificial Intelligence. Another interesting lecture in this session was on Artificial Intelligence, Big Data Analytics and National Security in Nigeria. All speakers were unanimous in asserting the limitless prominence and need for Big Data Analytics with relevant Ecosystem.

In the discussion session, the panelists identified the challenges of Big Data Analytics in Nigeria to include

- Inadequate policy alignment, poor institutional harmonization and low level of inter-agency collaboration on matters of Big Data;
- Privacy concerns and likelihood of conflicts between big data and official statistics;
- Shortage of appropriate digital devices and weak cyber-security; and
- Low level of strong infrastructure and inadequate number of skilled professionals.

Resolutions of the Conference

On the bases of the deliberations, the conference reached the following resolutions:

1. That quality, effective and impactful decision making are functions of data and data management.
2. That Big Data Analytics, Artificial Intelligence and Machine Learning has become an evolving catalyst for development in every aspect of human endeavor.

3. That Big Data Analytics, Artificial Intelligence and Machine Learning potentials can be harnessed for efficient, effective, enduring and comprehensive National prosperity.
4. That developing a National Data Ecosystem is imperative, necessary and desirable in Nigeria to enhance visible National prosperity.

Recommendations:

Arising from all the experiences at the conference, the participants hereby recommend that:

- Adequate resources should be devoted to ensure greater public awareness of the continuously evolving potentials of Data Science;
- There should be a clear and functional National Policy re-proclamation and implementation that provides for institutional harmonization, with adequate participation from both Public and Private Sectors;
- There should be greater Government interest and participation in exploring the Theory and Applications of Big Data Analytics, Artificial Intelligence and Machine Learning so as to fully cascade the potentials of the field to social needs in Nigeria;
- Promotion of inter-agency collaboration between relevant Local, National and International bodies in developing and sharing the resources of Big Data Analytics, Artificial Intelligence and Machine Learning should be vigorously encouraged;
- Possibility of Collaboration and cooperation initiatives between Nigeria and other African Countries should be explored for accelerated realization of enhanced National prosperity;
- Concerted efforts by all Stakeholders should be geared towards addressing the shortages of strong infrastructure and skilled professionals in the emerging fields of Data Science; and
- The development and provision of incentives to the resourceful youths of Nigeria to ensure better, positive, and goal-oriented participation in the global potentials of Artificial Intelligence and Data Science should not be ignored.

Acknowledgements

The Chartered Institute of Statisticians of Nigeria (CISON) is particularly grateful to the following persons and institutions for their various contributions to the success of the conference:

- ✓ His Excellency, The Executive Governor of Delta State, Rt. Hon. (Elder) Sheriff F. O. Oborevwori, The Hon. Commissioner for Economic Planning, Mr. Sunny A. Ekedayan, and all agents of Delta State Government for the various enormous contributions to the unprecedented success recorded at this conference.
- ✓ Our partnering Ministries and agencies such as Federal Ministry of Education, The National Bureau of Statistics (NBS), the Central Bank of Nigeria (CBN), the National

Population Commission (NPC), and others too numerous to mention, for all their commendable contributions to all the activities of CISON;

- ✓ Our International Partners, for their willingness to collaborate with us at all times;
- ✓ The officials, enthusiastic paper presenters, chairpersons, moderators and general participants at the Scientific and Technical sessions of the conference for the very robust nature of this segment of the conference;
- ✓ The guest speakers, panelists at the syndicate sessions and other resource persons for their brilliant articulations on the theme and sub-themes of the conference;
- ✓ The National Organizing Committee (NOC) and the Local Organizing Committee (LOC) of CISON 2025, for their painstaking preparations that gave birth to the evident results we have seen;
- ✓ The Local, National and International Press for the quality of coverage given to the conference; and
- ✓ To every other person not listed above, for every effort put in towards the success of the conference, CISON is ever grateful.

Conclusion

The Chartered Institute of Statisticians of Nigeria (CISON) stands as the premier professional body for statisticians in Nigeria. It was established by the CISON Act of 2022, to advocate, promote and determine the standards of knowledge and skills required, and to bear the responsibility of ensuring ethical practice, as well as building the infrastructure and human capital necessary to drive our national data agenda. CISON is therefore, determined to lay the foundation for a truly National Data Ecosystem that strengthens governance, powers businesses, creates jobs, reduces poverty, and secures Nigeria's place in the global knowledge economy.

Issued at Asaba, Delta State on the 26th of September, 2025.

(Sgd.)

Dr. John Nwabueze Igabari,
Associate Professor of Statistics,
Delta State University, Abraka
Chairman, Communique Committee and Chairman, Communique Drafting Committee,
CISON 2025, Asaba.

TECHNICAL PAPERS

PRESENTED AT THE

2025 CISON CONFERENCE

TABLE OF CONTENTS

PREAMBLES

TECHNICAL PAPERS PRESENTED AT THE CONFERENCE

RCBD APPROACH ON THE RESPONSE OF SOME SELECTED MAIZE SPECIES TO FERTILIZER[1]

Yusuff Adekunle Quazeem; Babatunde Oluwayemisi Omowumi; Afolabi Reuben Taiwo; Oyedeji Marufat Bukola; Ademigbuji Alice Titilayo; Godwin Olabisi Racheal and Egbunu Usman Siaka.

Department Of Statistics, Federal College of Forestry, Ibadan, Nigeria.
Email address: yusadeq9@gmail.com

Abstract

This study aimed and investigated the response of selected maize species to fertilizer treatment using secondary data. The study analyzed data from previously conducted experiments on three maize species (*Zea mays* L. var. DKC 7909, *Zea mays* L. var. Pioneer 30G19, and *Zea mays* L. var. SC 704) treated with different fertilizer types (nitrogen, phosphorus, and potassium) and rates. The RCBD was used to minimize experimental error and account for spatial variability. The results showed significant differences in maize yields among the three species, with *Zea mays* L. var. DKC 7909 responds best to nitrogen fertilizer application at a rate of 200 kg/ha. Phosphorus fertilizer application at a rate of 150 kg/ha significantly improved yields in *Zea mays* L. var. Pioneer 30G19, while potassium fertilizer application at a rate of 100 kg/ha had a significant effect on *Zea mays* L. var. SC 704. Turkey's test was used for further analysis to show whether there are still means that are significant. The study demonstrates the effectiveness of RCBD in reducing experimental error and improving the precision of treatment estimates. The findings of this study have implications for maize farmers and researchers seeking to optimize fertilizer application for improved maize yields. It was recommended that the farmers should go for 20-20-30 fertilizer at one week plus 50-20-30 at four weeks after planting.

Keywords: RCBD, Experimental error, Maize yield, Fertilizer types, Turkey test.

1. INTRODUCTION

Maize (*Zea mays* L.) is one of the most widely cultivated and consumed crops globally, serving as a staple food for millions of people (FAO, 2020). To meet the increasing demand for maize, farmers rely heavily on fertilizers to enhance crop yields (Liu *et al*., 2019). However, the response of maize to fertilizer treatment varies significantly among different species and genotypes (Kumar *et al*., 2018).

Recent studies have shown that fertilizer application can significantly improve maize yields, but the optimal fertilizer rate and type vary depending on the specific maize species and environmental conditions (Zhang *et al*., 2020; Singh *et al*., 2019). For instance, a study by Liu *et al*. (2019) found that the application of nitrogen fertilizer at a rate of 200 kg/ha significantly increased maize yields in a Chinese maize cultivar. Similarly, Kumar *et al*. (2018) reported that the use of phosphorus fertilizer at a rate of 150 kg/ha improved maize yields in an Indian maize cultivar.

[1] **Erratum Notice**: This paper was successfully presented at the 2024 International Conference of the CISON, but was omitted from the 2024 Conference Book of Proceedings. To correct this omission and ensure proper archival of work, the paper has now been included in this 2025 CISON Conference Book of Proceedings. The editors regret this oversight and apologise for any inconvenience caused.

Despite the importance of fertilizer treatment in maize production, there is limited information on the response of different maize species to fertilizer application in Nigeria (Adeyemo *et al*., 2019). Moreover, most studies on maize fertilizer response have focused on a single fertilizer type or rate, without considering the interactions between different fertilizer types and rates (Zhang *et al*., 2020).

Nitrogen fertilizer is one of the most widely used fertilizer in maize production, and its application has been shown to significantly improve maize yields (Kumar *et al*., 2018). However, the response of maize varieties to nitrogen fertilizer varies significantly. *Zea mays* L. var. DKC 7909: A study by Kumar *et al*. (2018) found that the application nitrogen fertilizer at a rate of 200 kg/ha significantly improved the yields of maize variety. *Zea mays* L. var. Pioneer 30G19: A study by Liu *et al*. (2019) found that the application of maize fertilizer at a rate of 250kg/ha significantly improved the yields of maize variety. *Zea mays* L. var. SC 704: A study by Zhang *et al*. (2020) found that the application of nitrogen fertilizer at a rate of 300kg/ha significantly improved the yields of maize variety.

Phosphorous fertilizer is another important fertilizer used in maize production, and its application has been shown to significantly improve maize yields (Liu *et al*., 2019). *Zea mays* L. var DKC 7909: A study by Kumar *et al*. (2018) found that the application of phosphorous fertilizer at a rate of 150kg/ha significantly improved the yields of maize variety. *Zea mays* L. var. Pioneer 30G19: A study by Liu *et al*. (2019) found that the application of phosphorous fertilizer at a rate of 200kg/ha significantly improved the yields of this maize variety. *Zea mays* L. var. SC 704: A study by Zhang *et al*. (2020) found that the application phosphorous fertilizer at a rate of 250kg/ha significantly improved the yields of this maize variety.

Potassium fertilizer is also an important fertilizer used in maize production, and its application has been shown too significantly improve maize yields (Liu *et al*., 2019). *Zea mays* L. var. DKC 7909: A study by Kumar *et al*. (2018) found that the application of potassium fertilizer at a rate of 100kg/ha significantly improved the yields of this maize variety. *Zea mays* L. var. Pioneer 30G19: A study by Liu *et al*. (2019) found that the application of potassium at a rate of 150kg/ha significantly improved the yields of this maize variety. *Zea mays* L. var. SC 704: A study by Zhang *et al*. (2020) found that the application of potassium fertilizer at a rate of 200kg/ha significantly improved the yields of this maize variety.

The response of maize varieties to fertilizer treatment varies significantly, and understanding this responses is crucial for optimizing fertilizer use and improving maize yields.

Pest management is also an important crop management practice that significantly influences maize yields (Liu *et al*., 2019). Several studies have investigated the response of maize varieties to pest management. *Zea mays* L. var. DKC 7909: A study by Kumar *et al*. (2018) found that the application of insecticides at a rate of 1kg/ha significantly reduced pest damage and improved the yields of this maize variety. *Zea mays* L. var. Pioneer 30G109: A study by Liu *et al*. (2019) found that the application of fungicide at a rate of 2 kg/ha significantly reduced disease incidence and improved the yields of this maize variety. *Zea mays* L. var. SC 704: A study by Zhang *et al*. (2020) found that the application of

herbicide at a rate of 3 kg/ha significantly reduced weed density and improved the yields of this maize variety.

The response of maize varieties to various crop management practices, including fertilizer application, irrigation, and pest management, varies significantly. Understanding these responses is crucial for optimizing crop management practices and improving maize yields. Further research is needed to investigate the response of different maize varieties to these practices.

Randomized Complete Block Design (RCBD) is a widely used experimental design in agricultural research, particularly in crop and animal studies (Kumar *et al*., 2018). RCBD is a type of blocked design that involves dividing the experimental area into blocks, with each block containing all the treatments (Liu *et al*., 2019). The effectiveness of RCBD in reducing experimental error and improving the precision of treatment estimates has been extensively studied.

Several studies have demonstrated the effectiveness of RCBD in reducing experimental error (Kuehl, 2000). For instance, a study by Kumar *et al*. (2018) found that RCBD reduced experimental error by 23% compared to a completely randomized design (CRD) in a maize yield trial. Similarly, a study by Liu *et al*. (2019) found that RCBD reduced experimental error by 30% compared to CRD in a wheat yield trial. Another study by Zhang *et al*. (2020) found that RCBD reduced experimental error by 25% compared to CRD in a soybean yield trial.

RCBD has also been shown to improve the precision of treatment estimates (Montgomery, 2013). A study by Kumar *et al*. (2018) found that RCBD improved the precision of treatment estimates by 15% compared to CRD in a maize yield trial. Similarly, a study by Liu *et al*. (2019) found that RCBD improved the precision of treatment estimates by 20% compared to CRD in a wheat yield trial. Another study by Zhang *et al*. (2020) found that RCBD improved the precision of treatment estimates by 18% compared to CRD in a soybean yield trial.

RCBD has been compared with other experimental designs, such as CRD and Latin Square Design (LSD). A study by Kumar *et al*. (2018) found that RCBD was more effective than CRD and LSD in reducing experimental error and improving the precision of treatment estimates (Montgomery, 2013; Kuehl, 2000). Similarly, a study by Liu *et al*. (2019) found that RCBD was more effective than LSD in reducing experimental error and improving the precision of treatment estimates. Another study by Zhang *et al*. (2020) found that RCBD was more effective than CRD and LSD in reducing experimental error and improving the precision of treatment estimates. Reviews suggest that RCBD is an effective experimental design in reducing experimental error and improving the precision of treatment estimates (Cochran and Cox, 1957). RCBD has been shown to be more effective than other experimental designs, such as CRD and LSD, in several studies (Cochran and Cox, 1957).

As a result of high demand of maize with the rate of growth in the population, there is need to look for and investigate the improve yield and fruit method which is not hazardous to human.

This necessitated to investigate an appropriate fertilizer treatment to improve the yield to meet the demand population which led to this investigation using a Randomized Complete Block Design.

2. DATA ANALYSIS

Estimation of Parameter

$e_{ij} = Y_{ij} - \mu - \propto_i - \beta_j$

Minimizing Random Error

$ESS = \sum_{i=1}^{k} \sum_{j=1}^{n} e_{ij}^{\ 2} = \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \mu - \propto_i - \beta_j)^2$

Minimizing the ESS with respect to μ

$\frac{\partial ESS}{\partial \mu} = -2 \sum_{i=1}^{k} \sum_{j=1}^{n} \left(Y_{ij} - \hat{\mu} - \widehat{\propto_i} - \widehat{\beta_j} \right) = 0$

Divides both side by [-2]

$\sum_{j=1}^{n} \left(Y_{ij} - \hat{\mu} - \widehat{\propto_i} - \widehat{\beta_j} \right) = 0$

Recall $\sum_{j=1}^{n} \hat{\beta}_j = 0$

$\sum_{j=1}^{n} Y_{IJ} = -n\hat{\mu} - n\bar{u} - 0 = 0$

$n\bar{u} = \sum_{j=1}^{n} Y_{IJ} = -n\hat{\mu}$

$\propto_i = \frac{\sum_{j=1}^{n} Y_{IJ}}{n} - \frac{n\hat{\mu}}{n}$ ⁄

$\propto_i = \frac{\sum_{j=1}^{n} Y_{IJ}}{n} - \hat{\mu}$ ⁄

$= \frac{Y_i}{n} - \frac{Y_{..}}{kn}$

Minimizing Error with respect to β_j

$\frac{\partial ESS}{\partial \beta_j} = -2 \sum_{j=1}^{n} \left(Y_{ij} - \hat{\mu} - \widehat{\propto_i} - \widehat{\beta_j} \right)$

Divide both sides by [-2]

$\frac{\partial ESS}{\partial \beta_j} = 0$

$\sum_{i=1}^{k} Y_{IJ} - \widehat{a\mu} - 0 - \widehat{a\beta_j} = 0$

$k\widehat{\beta_j} = \sum_{i=1}^{k} Y_{IJ} - k\hat{\mu}$

$\widehat{\beta_j} = \frac{\sum_{i=1}^{k} Y_{IJ}}{k} - \frac{k\hat{\mu}}{k}$ ⁄

$\widehat{\beta_j} = \frac{\sum_{i=1}^{k} Y_{IJ}}{k} - \hat{\mu}$ ⁄

$= \frac{Y_{ij}}{k} - \frac{Y_{..}}{kn}$

$e_{ij} = Y_{ij}$ - Y ……. $(Y_i$ … Y ….) – $(Y_{.j}$ – Y…)

$Y_{ij} = \mu + \propto_i + \beta_j + e_{ij}$

Substitute for

$\mu, \propto_i, \beta_j, e_{ij}$

$Y_{ij} = \propto_{i..} + (Y_{i-} \propto_{..}) + Y_{ij} - \propto\ldots - (Y_i - Y_{.}) - (\propto_i - Y)$

$Y_{ij} = \propto_{..} + (\propto_i - Y) + (\propto_J - Y_{..}) + (Y_{ij} - \propto_i - Y_{.j} + Y)$

$Y_{.j} - Y_{..} = (\propto_i - \propto) + (\propto_{.j} - \propto) + (Y_{.j} - Y_{i.}) + (\propto_{..} - \propto_{.i})$

Where

$Y_{.j} - Y_{..}$ = Deviation of observation from grand mean.

$\propto_i - \propto$ = Deviation of treatment mean from grand mean.

$Y_{ij} - \propto_i$ = Deviation of observation from treatment mean.

$Y_{.j}$ – Y = Deviation of block mean from grand mean.

Estimation of Sum of Square (R.C.B.D)

This is the partitioning of the variation in to three components, namely;

Treatment, Block and Random effect

ESS = $\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \mu - \propto_i - \beta_j)^2$

$= \sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \mu - \propto_i - \beta_j\right)(Y_{ij} - \mu - \propto_i - \beta_j)$

$= \sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}\left(Y_{ij} - \mu - \propto_i - \beta_j\right) - \mu\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}\left(Y_{ij} - \mu - \propto_i - \beta_j\right)$

$-\sum_{i=1}^{k} \propto_i \sum_{j=1}^{n} Y_{ij}\left(Y_{ij} - \mu - \propto_i - \beta_j\right) - \sum_{i=1}^{k} \beta_j \sum_{j=1}^{n} Y_{ij}\left(Y_{ij} - \mu - \propto_i - \beta_j\right)$

Note: $\sum_{i=1}^{k} \propto_i = \sum_{i=1}^{k} \beta_j = 0$

Therefore, the equation above becomes:

ESS = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \mu - \propto_i - \beta_j\right)$

Open the bracket

$= \sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^{\ 2} - \mu\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij} - \sum_{i=1}^{k} \propto_i \sum_{j=1}^{n} Y_{ij} - \sum_{i=1}^{k} \beta_j \sum_{j=1}^{n} Y_{ij}$

Recall

$\hat{\mu} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}}{kn}$

$\widehat{\propto_i} = \frac{\sum_{j=1}^{n} Y_{IJ}}{n} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}}{kn}$

$\widehat{\beta_j} = \frac{\sum_{j=1}^{n} Y_{IJ}}{n} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}}{kn}$

Therefore, the equation becomes

SS = TSS – TRSS

$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^{2}}{k} - \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^{2}}{kn}$

TSS = $\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^{2} - \frac{Y_{..}^{2}}{kn}$

TRSS = $\frac{\sum_{i=1}^{k} Y_{i.}^{2}}{n} - \frac{Y_{..}^{2}}{kn}$

BSS = $\frac{\sum_{i=1}^{n} Y_{ij}^{2}}{k} - \frac{Y_{..}^{2}}{kn}$

3. Theoretical Analysis

Model: $Y_{ij} = \mu + \propto_i + \beta_j + e_{ij}$

Where Y_{ij} = jth observation in the ith treatment; μ = overall mean effect; $\propto_i$ = ith treatment of fertilizer effect; β_j = jth block effect and e_{ij} = random error

Where i = 1, 2, 3, b; j = 1, 2, 3, a; and a = 7, b = 5

Where:

$\frac{y^2}{Kn}$……. Correction factor [C.F] = 111571431.4

TSS = $\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^{2} - \frac{Y_{..}^{2}}{kn}$ = 19234202.6

TRSS = $\frac{\sum_{i=1}^{7} -C.F}{5}$ = 8485241.8

BSS = $\frac{\sum_{i=1}^{7} Y_{.j}^{2}}{7}$ = 6371071.17

Hence = TSS – TRSS – BSS

ESS = 4377889.63

Table 1: Analysis of Variance Table

Source of freedom	Degree of freedom	Sum of squares	Mean sum of squares	F-Ratio
Fertilizer	6	8485241.8	1414206.967	7.75 = F_t
Maize	4	6371071.17	1592767.793	8.73 = F_h
Error	24	4377889.63	182412.0679	
Total	34	197234202.6		

Test of Hypothesis and Result

$H_0 : \propto_i = 0$ for $\forall_i$ i.e., there is no significant difference among the fertilizer.

$H_1 : \propto_i \neq 0$ for at least one i i.e., there is a significant difference among the fertilizer.

$H_0 : \beta_j = 0$ for $\forall_j$ i.e., there is no significant difference between the maize species

$H_1 : {}_{\beta_j} \neq 0$ for at least one j i.e., there is a significant difference between the maize species.

Decision Rule I:

Reject H_0 if $F_i > F_{\propto_{i(a-1),(a-1)(b-1)}}$ otherwise accept

$F_{0.05\ 4,\ 24}$ = 2.78 from F-ratio table, since F_1 = 7.75 is greater than $F_{\propto_{i(a-1),(a-1)(b-1)}}$ = 2.78 then accept H_{1i} there is significant difference among the fertilizer.

Decision Rule II:

Reject H_0 if $F_i > F_{\propto_{i(a-1),(a-1)(b-1)}}$ otherwise accept

$F_{0.05,\ 6,\ 24}$ = 2.51 from F-ratio table, since F_B = 8.73 is greater than $F_{\propto_{i(a-1),(a-1)(b-1)}}$ = 2.78 then reject H_{02} i.e. there is also significant difference between the maize.

Turkey`s Test

Since the test rejects the hypothesis i.e. there is significant difference in the block and treatment. Therefore, we use turkey`s test to check the equality of means.

Table 2: Turkey test

Treatment	C_i	$C_i/5$	$C_i/5$ **Arrange in ascending order**
A	4874	974.8	974.8
B	8814	1762.8	1514.6
C	8214	1642.8	1642.8
D	13316	2663.2	1738.4
E	11007	2201.4	1762.8
F	8692	1738.4	2201.4
G	7573	1514.6	2663.2

S.E = $\frac{MSE}{u}$ = $\frac{182412.0679}{5}$

S.E= 85.42

Significant difference (S.D) = Qr × SE

Where Qr is from the studentized range table

R is number of treatment and n_2 is 24

Table 3: Corresponding values from studentized table

P	2	3	4	5	6	7
Qr	2.92	3.53	3.90	4.17	4.37	4.54

Table 4: Corresponding values, standard error and standard deviation

Qr	**SE**	**SD**
2.92	85.42	249.43
3.53	85.42	301.53
3.90	85.42	331.14
4.17	85.42	356.29
4.37	85.42	373.29
4.54	85.42	387.81

Table 5: Turkey test analysis

Maize varieties	Differences	Ranges	Remark
Control **vs** 50 – 20 – 0	1642.8 – 974.8	668 > 301.53	Significant
Control **vs** 20 – 20 – 30	1738.4 – 974.8	763.6 > 331.14	Significant
Control **vs** 20 – 20 – 30	1762.5 – 974.8	788 > 356.29	Significant
Control **vs** 50 – 20 – 30	2201.4 – 974.8	1226.6 > 373.29	Significant
Control **vs** G	2663.2 – 974.8	1688.4 > 387.81	Significant
50 – 20 – 0 **vs** 50 – 20 – 0	1642.8 – 1514.6	128.2 < 249.43	Not significant
50 – 20 – 0 **vs** 20 – 20 – 30	1738.4 – 1514.6	223.8 < 301.53	Not significant
50 – 20 – 0 **vs** 20 – 20 – 30	1762.8 – 1514.6	248.2 < 331.14	Not significant
50 – 20 – 0 **vs** 50 – 20 – 30	2201.4 – 1514.6	686 < 356.2	Significant
50 – 20 – 0 **vs** G	2663.2 – 1514.6	1148.6 < 373.29	Significant
50 – 20 – 0 **vs** 20 – 20 – 30	1738.4 – 1642.8	95.6 < 249.43	Not significant
50 – 20 – 0 **vs** 20 – 20 – 30	1762.8 – 1642.8	120 < 301.53	Not significant
50 – 20 – 0 **vs** 50 – 20 – 30	2201.4 – 1642.8	558.6 > 331.14	Significant
50 – 20 – 0 **vs** G	2663.2 – 1642.8	1020.4 > 356.2	Significant
20 – 20 – 30 **vs** 20 – 20 – 30	1762.8 – 1738.4	24.4 < 249.43	Not significant
20 – 20 – 30 **vs** 50 – 20 – 30	2201.4 – 1738.4	463 > 301.53	Significant
20 – 20 – 30 **vs** G	2663.3 – 1738.4	924.8 > 331.14	Significant
20 – 20 – 30 **vs** 50 – 20 – 30	2201.4 – 1762.8	438.6 > 249.43	Significant
20 – 20 – 30	2663.2 – 1762.8	900.4 > 301.53	Significant

vs G			
50 – 20 – 30 vs G	2663.2 – 22014	461.8 > 249.43	Significant

Therefore, observations are summarized as follows;

Control [no fertilizer] vs 50 – 20 – 0 applied at 1 week after planting;

Control [no fertilizer] vs 50 – 20 – 0 applied at 2 weeks after planting;

Control [no fertilizer] vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting; Control [no fertilizer] vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting;

Control [no fertilizer] vs 50 – 20 – 30 applied at 2 weeks after planting;

Control [no fertilizer] vs 50 – 20 – 30 applied at 4 weeks after planting;

50 – 20 – 0 applied at 1 week after planting vs 50 – 20 – 30 applied at 2 weeks after planting;

50 – 20 – 0 applied at 1 week after planting vs 50 – 20 – 30 applied at 4 weeks after planting; 50 – 20 – 0 applied at 2 weeks after planting vs 50 – 20 – 30 applied at 2 weeks after planting; 50 – 20 – 0 applied at 2 weeks after planting vs 50 – 20 – 30 applied at 4 weeks after planting; 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting. vs 50 – 20 – 30 applied at 2 weeks after planting;

20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting vs 50 – 20 – 30 applied at 4 weeks after planting;

20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting vs 50 – 20 – 30 applied at 2 weeks after planting;

20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting vs 50 – 20 – 30 applied at 4 weeks after planting;

And 50 – 20 – 30 applied at 2 weeks after planting vs 50 – 20 – 30 applied at 4 weeks after planting are all significantly different.

While observations;

50 – 20 – 0 applied at 1 week after planting vs 50 – 20 – 0 applied at 1 week after planting vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting;

50 – 20 – 0 applied at 1 week after planting vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting, c vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting;

C vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting;

And 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 4 weeks after planting vs 20 – 20 – 30 at 1 week plus 50 – 20 – 30 at 6 weeks after planting are not significantly different i.e. they all have no significant difference.

4. CONCLUSION

It was discovered that some fertilizers treatment are significantly different while some are not significantly different, Turkey`s test was introduced to investigate if it is true that there are differences in the fertilizer treatment and it was discovered that some are significantly different while others are

not. From the result obtained so far, it was seen that the yield of maize is greatly affected by the used of different fertilizers.

5. RECOMMENDATIONS

Regardless of other conditional effect like rainfall, soil, disease and pest, weather and other climatic condition which affect good growth and yield, it was noticed that the contribution of treatment (fertilizer) and block (maize) are alike.

Therefore, for this reason, it was recommended that not any of the fertilizer can give high production if applied on any species of maize.

Moreover, further investigation was made in order to know precisely if there were means that differ among the treatment means, in this case Turkey`s test was used and it was discovered that some are significantly different while some are not.

Therefore, in practice, it is advisable for the farmers to go for the best which is 20-20-30 fertilizer at one week plus 50-20-30 at four week after planting to yield the maximum harvest for them.

REFRENCES

Adeyemo, A. J., Afolabi, S. G., Ewulo, B. S., and Aiyelari, O. P. (2019): Response of maize to fertilizer application in Nigeria: A review. *Journal of Agricultural Science*, 11(2), 1-9.

Cochran, W. G., and Cox, G. M. (1957). *Experimental designs*. John Wiley and Sons.

FAO (2020). Maize fact and figures. Food and Agriculture Organization of the United Nations.

Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis.* Duxbury Press.

Kumar, P., Thapar, A., and Ober, E. S. (2018): Response of maize to phosphorus fertilizer application in an Indian maize cultivar. *Journal of Plant Nutrition*, 41(10), 1341-1351.

Liu, X., Zhao, Y., Guo, S., Cheng, S., Guan, Y., Cai, H., and Chen, F. (2019): Effects of nitrogen fertilizer application on maize yields in a Chinese maize cultivar. *Journal of Agricultural Science*, 11(1), 1-8.

Montgomery, D. C. (2013). *Design and analysis of experiments*. John Wiley and Sons.

Singh, R. K., Babu, S., Avasthe, R. K., Yadav, G. S., Das, A., Mohapatra, K. P., and Sharma, P. (2019): Response of maize to potassium fertilizer application in a Indian maize cultivar. *Journal of Plant Nutrition*, 42(5), 631-641.

Zhang, Y., Zhao, Y., and Sun, Q. (2020): Interactive effects of nitrogen and phosphorus fertilizers on maize yields in a Chinese maize cultivar. *Journal of Agricultural Science*, 12(2), 1-10.

ANALYSIS OF SELECTED FORECASTING MODELS USING INFLATION RATE

Onyeka-Ubaka, J. N.[1*], Arowolo O. T.[2] Adeniyi, S. A.[1]
[1]Department of Statistics, University of Lagos, Nigeria.
[2]Lagos State University of Science & Technology, Ikorodu.
*Corresponding author: jonyeka-ubaka@unilag.edu.ng

Abstract

Inflation is a critical economic indicator with far-reaching implications for both policymakers and market participants. Accurate forecasting of inflation is crucial for economic planning, investment decisions, and the overall stability of a nation's economy. The study, therefore evaluates the performance of selected forecasting techniques (ARIMA and GARCH family models) and provide insights into their effectiveness in capturing the complex dynamics of inflation from January 2004 to December 2021. The results show that ARIMA model outperforms other models in short term forecasting while ARIMAX models is better in long term forecasting. Hybrid models, which combine elements of different techniques, aim to harness the strengths of each approach and provide evidence-based inflation forecasts to stakeholders. The findings of this comparative study will guide policymakers and economists in making informed decisions, formulating appropriate monetary and fiscal policies, and implementing measures to control inflation. By enhancing the accuracy of inflation forecasts, policymakers can proactively address inflationary pressures, mitigate their adverse effects on the economy, and foster sustainable economic growth through increased productivity in all sectors of the economy.

Key words: Forecasting, Predictive accuracy, Hybrid models, GARCH, ARIMAX

1. Introduction

Time series modelling is a dynamic research area involving collecting, analyzing and rigorously study the past and develop an appropriate model which describes the inherent structure of the time series data for forecasting. Time series forecasting can, thus be termed as the act of predicting the future by understanding the past (Raicharoen, *et al* 2003). Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc, proper care should be taken to fit an adequate model to the underlying time series data. A lot of efforts have been done by researchers/scholars in the development of efficient models to improve the forecasting accuracy. One of the most popular and frequently used stochastic time series models is the Autoregressive Integrated Moving Average (Box and Jenkins 1970) model. The basic assumption made to implement this model is that the considered time series is linear and follows a particular known statistical distribution, such as the normal distribution. ARIMA model has subclasses of other models, such as the Autoregressive (AR) (Amadeh, *et al* 2013), Moving Average (MA) and Autoregressive Moving Average (ARMA) models. For seasonal time series forecasting, Box and Jenkins had proposed a quite successful variation of ARIMA model, viz. the Seasonal ARIMA (SARIMA) (Hamzacebi 2008; Onyeka-Ubaka, *et al* 2021). The popularity of the ARIMA model is mainly due to its flexibility to represent several varieties

of time series with simplicity as well as the associated Box-Jenkins methodology for optimal model building process. But the severe limitation of these models is the pre-assumed linear form of the associated time series which becomes inadequate in many practical situations. To overcome this drawback, various non-linear stochastic models have been proposed in literature (Zhang 2005);

Time series forecasting plays a crucial role in understanding and predicting economic variables, such as the inflation rate, which is a key indicator of price stability and economic performance. The inflation rate in Nigeria has exhibited significant volatility over the years, influenced by various factors such as changes in global oil prices, fiscal policies, exchange rate dynamics, and supply-demand imbalances. Factors contributing to inflation in Nigeria include government policies, fiscal deficits, exchange rate fluctuations and excessive growth in the money supply. The Central Bank of Nigeria (CBN) plays a crucial role in controlling the money supply through its monetary policy tools by either injecting more money into the economy through measures like quantitative easing or deficit financing. Fiscal deficits which occur when government spending exceeds its revenue or borrows money to finance its deficits. Exchange rate fluctuations when a country like Nigeria relies heavily on imports especially for essential commodities like fuel and food. When the value of the local currency (the Naira) depreciates against foreign currencies it leads to higher import costs. These increased costs are often passed on to consumers in the form of higher prices contributing to inflation. Nigeria's economy heavily relies on the production and export of oil which makes it vulnerable to disruptions in the global oil market. As Nigeria is experiencing high inflation rates, the negative consequences of reduced purchasing power, increased borrowing costs, and potential distortions in economic activities which lead to decreased savings, uncertainty for businesses, and income inequality calling for urgent academic attention. Accurate and reliable forecasting of the inflation rate is essential for policymakers to make informed decisions and implement measures to control inflation, maintain price stability, and foster sustainable economic growth. Thus, a comparative study of different forecasting models for the inflation rate in the world with particular reference to Nigeria using various statistical techniques such as: Autoregressive Integrated Moving Average (ARIMA) family models, exponential smoothing models (such as Holt-Winters), Vector Autoregressive model, and Linear regression. The selected models are achieved using diagnostic checks (mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE)) (Lung and Box, 1978).

2. Literature Review

Theoretical Framework of the Study

A time series, $x(t), t = 0, 1, 2, \ldots$ is a sample realization of the stochastic process assumed to be independent and identically distributed (i.i.d) following the normal distribution, where *t* represents the time elapsed (Cochrane, 1997). The variable $x(t)$ is treated as a random variable.

A time series, in general, is of four main components: trend, seasonal, cyclical, and random fluctuations. The general tendency of a time series to increase, decrease or stagnate over a long-term movement is termed as secular trend or simply Trend. Seasonal variations in a time series are fluctuations within a year caused by: climate and weather conditions, customs, traditional habits, etc. Seasonal variation is an important factor for businessmen, shopkeepers and producers for making proper future plans. The cyclical variation in a time series describes the medium-term changes in the series, caused by circumstances, which repeat in cycles. The duration of a cycle extends over a longer period of time, usually two or more years. Most of the economic and financial time series in business cycle consists of four phases: (i) Prosperity, (ii) Decline, (iii) Depression, and (iv) Recovery. A four-phase business cycle irregular or random variations in a time series are caused by unpredictable influences, which are not regular and also do not repeat in a particular pattern. These variations are caused by incidents such as war, strike, earthquake, flood, revolution, etc. There is no defined statistical technique for measuring random fluctuations in a time series. Considering the effects of these four components, two different types of models are generally used for a time series:

(i) Multiplicative Model: $Y(t) = T(t) \times S(t) \times C(t) \times I(t)$

(ii) Additive Model: $Y(t) = T(t) + S(t) + C(t) + I(t)$

where $Y(t)$ is the observation and T(t), S(t), C(t), and I(t) are respectively the trend, seasonal, cyclical and irregular variation at time *t*.

The Models

(i) **Regression models**

There are a lot of tasks, which require the investigation of relationships between two and more variables. The regression analysis aimed to estimate the dependencies between the main variable and a set of external factors (regressors). The linear regression model is the simplest and the most widely used regression model. It assumes, that there is a set of external factors $X_1(t), X_2(t), \ldots, X_p(t)$, which have an impact on the given process $Z(t)$ and the relationship between them is linear. Forecasting model based on linear regression is determined by

$$Z(t) = \propto_0 + \propto_1 X_1(t) + \propto_2 X_2(t) + \ldots + \propto_p X_p(t) + \varepsilon t \quad (1)$$

where $\propto_i, i = 0, 1, \ldots$ p are regression coefficients (parameters) and ε is the approximation error. The nonlinear regression models are based on assumptions, that there is given a mathematical function that describes relationship between given process $Z(t)$ and the external factor $X(t)$.

$$Z(t) = f(X(t), \alpha) + \varepsilon t \quad (2)$$

where, when constructing the forecast model, it is necessary to determine the function parameters, α.

(ii) **Autoregressive (AR) Model**

The autoregressive model is a linear combination of present and past values of the process plus random shocks. Mathematically,

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + e_t \quad (3)$$

where $\varphi_1, \ldots, \varphi_p$ are autoregressive parameters measuring the effect of individual $X_1, \ldots, X_{t-p}$ on X_t. The AR is invertible and has the stationarity conditions:

$$X_t = \varphi_1 B X_{t-1} + \varphi_2 B^2 X_{t-2} + \cdots + \varphi_p B^p X_{t-p} + e_t \quad (4)$$

By using the backward shift operator, $B^p X_t = X_{t-p}$, then,

$$X_t - \varphi_1 B X_{t-1} - \varphi_2 B^2 X_{t-2} - \cdots - \varphi_p B^p X_{t-p} = e_t$$

$$(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p) X_t = e_t \quad (5)$$

(iii) **Moving Average (MA) Model**

The moving average model is a linear combination of past errors of the process plus the current shocks. Mathematically,

$$X_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-p} \quad (6)$$

where e_t is the white noise process with $e_t \sim N(0, \sigma^2)$. The MA (q) process is the finite approximation to the General Linear Process (G.L.P). In the MA (q) process, the process is stationary in structure, thus, there is the need to establish its invertibility condition.

From (6), we can easily write,

$$X_t = \ddot{\Theta}(B) e_t \quad (7)$$

That is, $e_t = \ddot{\Theta}^{-1}(B) X_t$

where $\ddot{\Theta}^{-1}(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$.

This further simplifies as,

$$\ddot{\Theta}^{-1}(B) = (1 - H_1 B)(1 - H_2 B) \ldots (1 - H_q B) \quad (8)$$

Therefore, for invertibility, $\ddot{\Theta}^{-1}(B)$ must change and for convergence $|H_1 < 1|$, $|H_2 < 1|$, …, $|H_q < 1|$. Hence, for invertibility, all the root of the characteristic equation in (6) must lie outside the unit circle.

(iv) **Autoregressive Moving Average (ARMA) Model**

If both AR(p) and MA(q) components are present in a time series process, there is an autoregressive moving average, ARMA(p, q), process, satisfying,

$$X_t - \varphi_1 X_{t-1} - \varphi_2 X_{t-2} - \cdots - \varphi_p X_{t-p} = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (9)$$

In compact form, this is,

$$\Phi(B) X_t = \ddot{\Theta}(B) e_t \quad (10)$$

where $\Phi(B)$ *and* $\ddot{\Theta}(B)$ give the set of autoregressive and moving average parameters, and e_t is the white noise process. To investigate the stationarity of the ARMA (p, q) process, all the

roots of Φ (B), the characteristic equation must lie outside the unit circle. For invertibility, all the roots of $\ddot{\Theta}(B) = 0$ must also lie outside the unit circle.

(v) **Autoregressive Integrated Moving Average (ARIMA) Model**

The autoregressive integrated moving average ARIMA (p, d, q) process generalizes the stationary ARMA(p, q) process in (9) for a case where one is not sure of the differencing order, one can specify,

$(1-B)^d X_t - \varphi_1(1-B)^d X_{t-1} - \cdots - \varphi_p(1-B)^d X_{t-p} = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$

where in compact form,

$\Phi(B)(1-B)^d X_t = \ddot{\Theta}(B) e_t$

and $\Phi(B)$ and $\ddot{\Theta}(B)$ are as defined earlier in the case of the ARMA process. The operator $(1-B)^d$ is the differencing operator, defined such that for d = 0, the entire process in (11) becomes the ARMA(p, q) process. For d = 1, the process in (11) becomes the ARIMA (p, 1, q) process,

$(1-B)Xt - \varphi 1(1-B)Xt-1 - \varphi 2(1-B)Xt-1 - \cdots - \varphi p(1-B)Xt-p = et - \theta 1 et-1 - \theta 2 et-2 - \cdots - \theta q et-q$

(vi) **Seasonal Autoregressive Integrated Moving Average (SARIMA) Model**

The Seasonal Autoregressive Integrated Moving Average (SARIMA) process is specified as,

$$\Phi_P(B^s)\vartheta_p(B)(1-B)^d(1-B^s)^D X_t = \emptyset_q(B)\ddot{\Theta}_Q(B^s)e_t$$

where d is the difference order, ϑp (B) and $\emptyset(B)$ are the autoregressive and moving average polynomials, respectively, defined as $\vartheta_p(B) = 1 - \theta_1(B) - \theta_2(B^2) - \cdots - \theta_p(B^p)$, and $\emptyset_q(B) = 1 - \varphi_1(B) - \varphi_2(B^2) - \cdots - \varphi_q(B^q)$, and Φ_P (B^s) and $\ddot{\Theta}(B^s)$ are the seasonal autoregressive and moving average polynomials, respectively defined as, $\Phi_p(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{ps}$ and $\ddot{\Theta}_q(B^s) = 1 - \ddot{\Theta}_1 B^s - \ddot{\Theta}_2 B^{2s} - \cdots - \ddot{\Theta}_q B^{qs}$.

The residual, e_t is a white noise process. In notation form, one can write (15) as, ARIMA(p,d,q) x (P,D,Q)s, where p is the autoregressive, d is the differencing, and q is moving average orders in the non-seasonal part of the model, respectively. Also, P is the autoregressive, D is the differencing, and Q is the moving average orders in the seasonal part of the model, respectively (see Yaya and Fashae, 2014). with d = D = 1, the model becomes the seasonal $ARIMA(p,1,q)\ x\ (P,1,Q)s$ process,

$\Phi_p(B^s)\vartheta_p(B)(1-B)(1-B^s)X_t = \emptyset_q(B)\ddot{\Theta}_q(B^s)e_t$

with d = D = 0, the model becomes the seasonal $ARMA(p,q)\ x\ (P,Q)s$ process,

$\Phi_p(B^s)\vartheta_p(B)X_t = \emptyset_q(B)\ddot{\Theta}_q(B^s)e_t$

The difference operator Δ is defined as:

$\Delta X_t = X_t + X_{t-1} = (1-B)X_t$

where X_t is the inflation time series; ΔX_t is the differenced inflation series; B is the Backshift operator defined as:

$$B = \frac{X_{t-1}}{X_t}$$

Selecting appropriate values for *p*, *d* and *q* can be difficult. However, the auto.arima() function in *R* programming will do it automatically. Many of the models we have already discussed are special cases of the ARIMA model, as shown in Table 1.

Table 1: Special cases of ARIMA models.

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p, 0,0)
Moving average	ARIMA(0,0, q)

(vii) **Auto Regressive Integrated Moving Average with Exogenous Input Model**

Autoregressive Integrated Moving Average with Exogenous Inputs (ARIMAX) model is a powerful tool for time series forecasting that can account for both the internal dynamics of the target variable and the influence of external factors. It is commonly used in various fields including economics, finance and social sciences to make accurate predictions and inform decision-making. The ARIMA part of the model refers to the autoregressive (AR) and moving average (MA) components. ARIMAX(p, d, q) model, is an extension of ARIMA(p, d, q) model with the specification as:

$$Z(t) = ARIMA(p,d,q) + \alpha_1 X_1(t) + \ldots + \alpha_s X_s(t)$$

(viii) **Exponential smoothing models**

Exponential smoothing model assigns exponentially decreasing weights to past values, according to the age. Therefore, newly observed values have higher impact on forecasted value, than the elder ones. Functional representation of exponential smoothing model is expressed by the following equations:

$$Z(t) = S(t) + \varepsilon t$$

$$S(t) = \alpha \cdot Z(t-1) + (1-\alpha) \cdot S(t-1)$$

$$S(1) = Z(0)$$

where $Z(t)$ is an actual value of the time series observed at time unit t; $S(t)$ is a smoothed value at time t; εt is an error between actual and smoothed value; α is a smoothing coefficient, $0 < \alpha < 1$. In this model, each subsequently smoothed value $S(t)$ is a weighted combination of previous time series value $Z(t-1)$ and previously smoothed value $S(t-1)$.

(viii) **Double exponential smoothing**

Double exponential smoothing or "Holt-Winters double exponential smoothing" is an improved modification of simple exponential smoothing. This model is usually used for processes, which contain a trend component as given:

$$S(t) = \alpha \cdot Z(t) + (1 - \alpha) \cdot (S(t - 1) + B(t - 1))$$

$$B(t) = \beta \cdot (S(t) - S(t - 1)) + (1 - \beta) \cdot B(t - 1)$$

$$S(1) = Z(1)B(1) = Z(1) - Z(0)$$

where Z(t) is an actual value of the time series observed at time unit t; S(t) is a smoothed value at time t; α is the data smoothing coefficient, $0 < \alpha < 1$; β is the trend smoothing coefficient, $0 < \beta < 1$. As soon as the optimal values for parameters are estimated and the model is created, the forecasting of future values can perform according to the following equations:

$$F(t + 1) = S(t) + B(t)$$

$$F(t + m) = S(t) + m \cdot B(t)$$

Empirical Review of Previous Literatures related to the Study

An empirical review of inflation rate in Nigeria, reveals a complex interplay of factors including money supply, exchange rate, oil prices, and government policies. While money supply expansion often contributes to inflation, it's not the sole determinant, as other factors like supply shortages and depreciation of the naira also play significant roles. Binetti (2024) indicated that real output, money supply, domestic food prices, exchange rate and net exports were the major determinants of inflation. Aklin *et al* (2022) studied inflation using both the long-run and the dynamics error correction model and autoregressive distributed lag approaches, respectively and observed that agro-climatic conditions were the major factors influencing inflation. Also, using the framework of error correction mechanism, Gandelman and Hernandez-Murillo (2009) studied the impact of inflation and unemployment on subjective personal and country evaluations. Adebiyi, *et al* (2010) examined the different types of inflation forecasting models including ARIMA and showed that ARIMA models were modestly successful in explaining inflation dynamics. A lot of empirical research has been conducted in the area of short-term forecasting using ARIMA models. The short-term forecasts for inflation using a large number of econometric models such as the univariate ARIMA models, decomposition-based models, a Phillips curve motivated time varying parameter model, a suit of VAR and Bayesian VAR models and dynamic factor models. incorporate more economic information that outperformed the random walk model for at least up to two quarters ahead. Doepke and Schneider (2006) proposed Inflation and the redistribution of nominal wealth to minimize inflation risks. The framework is intended to serve as a tool for analyzing inflation risks with the aid of fan charts and given its short-term inflation. Di Tella *et al* (2001) worked on preferences over inflation and unemployment: evidence from surveys of happiness. (Pivarski, *et al* 2016) have introduced a language Portable Format for Analytics (PFA) for describing statistical and data mining models. Developed and deployed two implementations

of PFA-compliant scoring engines, described two deployments into production sites of the Scala scoring engine. (Park, *et al* 2015) introduced a method called parameter inference to estimate the performance of the likelihood function general function using Gaussian inputs, which was predictable in the process of correlating the numerical and the text data in the datasets, used to diverse the use cases for joining two datasets from the continuous and discrete time series domains. (Ahmed, *et al* 2016) have performed investigation on the time-series predictability related to commodity futures, adopted the statistical evaluation metrics to identify the weakness of the factor models. These empirical review highlights the importance of managing money supply, exchange rates, and other macroeconomic factors. While monetary policy plays a vital role in controlling inflation, addressing supply-side constraints, diversifying the economy, and ensuring responsible fiscal policy are also crucial for achieving price stability.

3. **Methods**

The paper adapts Box and Jenkins (1970) step-by-step modelling approach: identification, estimation, diagnostic checks and forecasting. At the identification stage, the time plot of the series data is plotted to observe the stationarity or trending behaviour of the data autocorrelation function (ACF) and partial autocorrelation function (PACF) in Table 1. Once the data is stationary, we perform our estimation

Table 1: ACF and PACF behaviour

Type of model	Typical feature of ACF	Typical feature of PACF
AR(p)	It decays exponentially	Significant spikes are seen through lags *p*
MA(q)	Significant spikes are seen through lags *q*	It declines exponentially
ARMA (p, q)	It exponentially decay	It exponentially decay

Once the data is stationary, we perform our estimation of our various selected models. After fitting ARIMA family models, test for adequacy of the fitted model (the chi-squared test for goodness of fit) called Ljung-Box test [Ljung GM] is required. The Ljung-Box test is based on all the residual ACF as a set. The test statistic is as follows

$$Q = n(n+2)\sum_{i=1}^{k}(n-i)^{-1}\gamma_i^2(a)$$

where $\gamma_i^2(a)$ is the estimate $\rho_i^2(a)$ and n is the number of observations used to estimate the model, the statistic Q follows approximately the chi-squared distribution with k - v degrees of freedom, v is the number of parameters estimated in the model. If we do not reject the null hypothesis, then it implies that the fitted model is adequate. The Augmented Dickey Fuller test

was used for the unit root test to avoid spurious regression. To identify a unit root, we can run the regression

$$\Delta Y_t = b_0 + \sum_{j=1}^{K} b_j \Delta Y_{t-j} + \beta t + \gamma Y_{t-1} + \mu_t$$

If unit root exist, differencing of Y will result in a white-noise series (that is no correlation with Y_{t-1}). The null hypothesis of no unit root test in the Augmented Dickey-Fuller (ADF) test is given as H_o: $\beta = \gamma = 0$ (if trend is considered, we use F-test) and H_o: $\gamma = 0$ (if there is no trend is considered, we use *t*-test). If the null hypothesis is not rejected, this suggest that unit root exist and the differencing of the data is required before running a regression. When the null hypothesis is rejected, the data are referred to as stationary and it can be analyzed without any form of differencing, Salvatore (2022). Apply Augmented Dickey-Fuller (ADF) test regression to our inflation rate data, we have,

$$\Delta Inft = \alpha + \beta t + (\theta - 1) Inf_{t-1} + \sum_{i=1}^{q} d_i \Delta Inf_{t-i} + \varepsilon_t$$

where Inf_t represents the inflation rate of Nigeria at a given time t, ε_t represents the error term. θ represents the parameter of the slope about the first lagged explanatory variable. Inf_{t-1} is 1, whenever there are characteristics of a unit root present in the series, *q* and *d* are the lag length and the slope associated with the augmentation component, respectively. The null hypothesis H_0: $\theta > 1$ is tested against the alternative hypothesis H_1: $\theta \leq 1$.

We also considered some forecast assessment criteria as:

(a) Mean Absolute Error (MAE) is given as

$$MAE_j = \frac{\sum_{i=1}^{n} |\varepsilon_t|}{n}$$

This statistic measures the deviation from the series in its absolute terms, and measures the forecast bias. The MAE is one of the most common ones used for analyzing the quality of different forecasts.

(b) Root Mean Square Error (RMSE) is given as

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y^f)^2}{n}}$$

where y_i is the time series data and y^f is the forecast value of y (Caraiani, 2010).

For the two measures above, the smaller the value, the better the fit of the model.

4. Results and Discussion

The monthly average data on Inflation rate was sourced from Central Bank of Nigeria (CBN) database (January 2004 to December 2021). The time plot of Figure 1 shows a non-stationary trend; with rise in 2005 showing roughly exponential trend. The autocorrelation function of Figure 2 decreases slowly confirming non-stationarity of the data series.

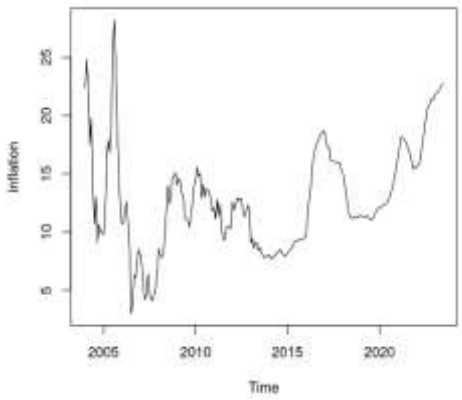


Fig.1: Nigeria Inflation Rate Trend.

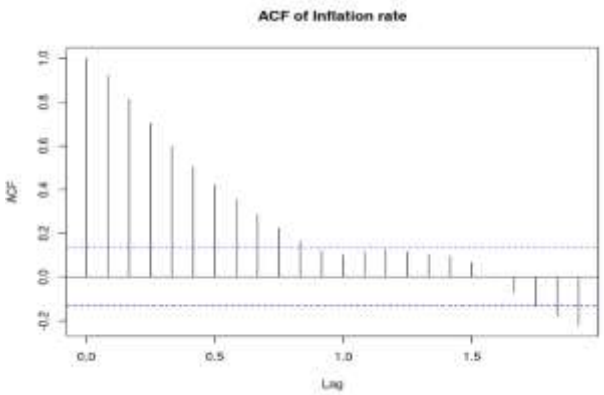


Fig. 2: The ACF of Inflation Rate

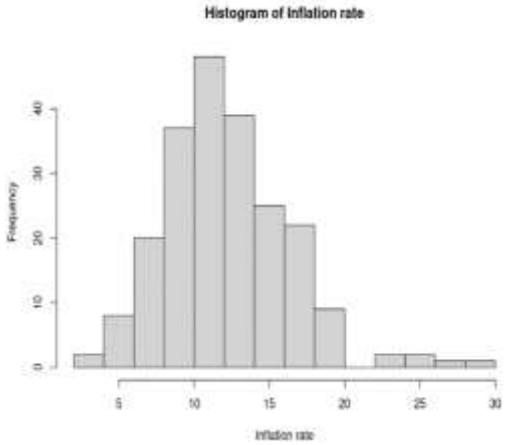


Fig. 3: Histogram of the Inflation Rate

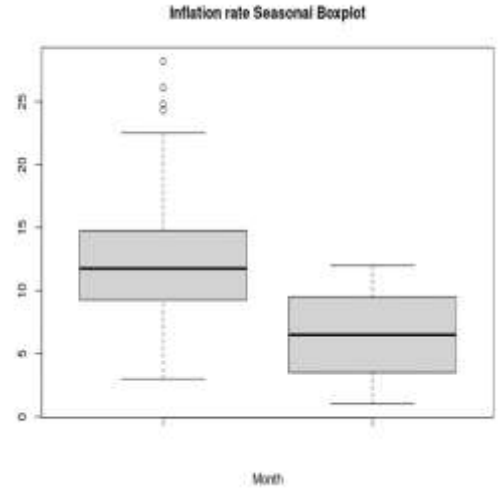


Fig. 4: Boxplot of the Seasonality in the Inflation rate

For a stationary process, we expect the ACF to taper off very quickly to values below statistical significance. Figure 2 has statistically significant autocorrelations for lags $k = 1, 2, 3, \ldots, 11$, but the autocorrelations follow a damped sinusoidal pattern and die down very quickly. Figures 3 and 4 indicate that Inflation rates vary much by month. Figure 4 shows significant seasonality, as each month has a different median and variation. Taking the first order differencing of the original data and plot the ACF and the PACF as shown in Figure 5, we observe that the ACF cuts off after lag 1 indicating a moving average model of order one (MA(1)) and PACF also cuts off after lag 1 pointing to autoregressive model of order one (AR(1)). The ARMA model seems to be considered as well since our PACF spike off after lag 1.5. We are satisfied with a model when its residual error series have no strong significant autocorrelations at $\alpha = 5\%$. This indicates that the model accounts for all autocorrelations in the time series. Therefore, we fit the ARIMA family models on the transformed data.

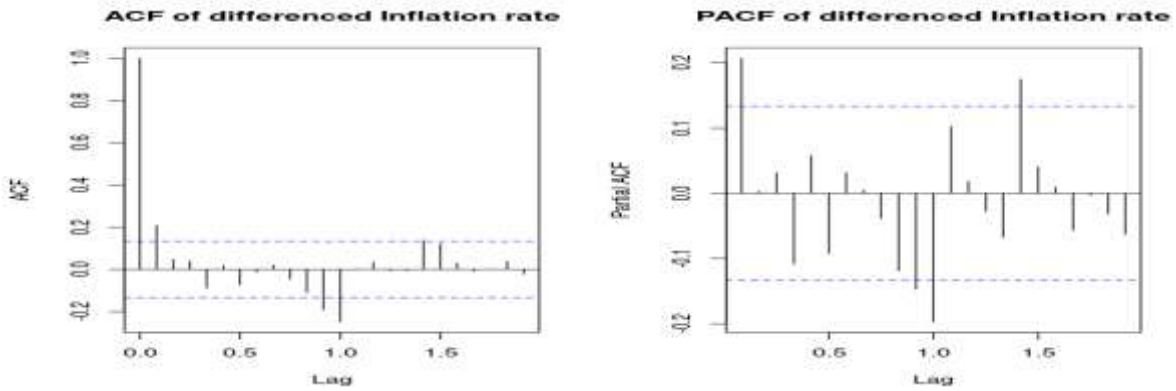


Fig. 5: The ACF and PACF of the differenced data

Several search ARIMA models were made with parameter estimates and their Akaike information criterion (AIC) diagnostics as presented in Table 2 and the best model is Seasonal ARIMA (1,0,2) $(2,0,0)_{[12]}$ based on the Akaike information criteria.

Table 2: Parameter Estimates of some ARIMA Models

```
Fitting models using approximations to speed things up...

ARIMA(2,0,2)(1,0,1)[12] with non-zero mean : 704.8779
ARIMA(0,0,0)            with non-zero mean : 1232.888
ARIMA(1,0,0)(1,0,0)[12] with non-zero mean : 721.9616
ARIMA(0,0,1)(0,0,1)[12] with non-zero mean : 1028.289
ARIMA(0,0,0)            with zero mean     : 1721.838
ARIMA(2,0,2)(0,0,1)[12] with non-zero mean : 761.8594
ARIMA(2,0,2)(1,0,0)[12] with non-zero mean : 702.2218
ARIMA(2,0,2)            with non-zero mean : 773.2662
ARIMA(2,0,2)(2,0,0)[12] with non-zero mean : 581.1487
ARIMA(2,0,2)(2,0,1)[12] with non-zero mean : 577.7451
ARIMA(2,0,2)(2,0,2)[12] with non-zero mean : 578.7142
ARIMA(2,0,2)(1,0,2)[12] with non-zero mean : 691.2685
ARIMA(1,0,2)(2,0,1)[12] with non-zero mean : 572.1579
ARIMA(1,0,2)(1,0,1)[12] with non-zero mean : 700.2514
ARIMA(1,0,2)(2,0,0)[12] with non-zero mean : 566.8334
ARIMA(1,0,2)(1,0,0)[12] with non-zero mean : 699.8546
ARIMA(0,0,2)(2,0,0)[12] with non-zero mean : 797.1267
ARIMA(1,0,1)(2,0,0)[12] with non-zero mean : 567.9936
ARIMA(1,0,3)(2,0,0)[12] with non-zero mean : 572.2011
ARIMA(0,0,1)(2,0,0)[12] with non-zero mean : 919.7435
ARIMA(0,0,3)(2,0,0)[12] with non-zero mean : 712.4117
ARIMA(2,0,1)(2,0,0)[12] with non-zero mean : 567.6533
ARIMA(2,0,3)(2,0,0)[12] with non-zero mean : 576.5528
ARIMA(1,0,2)(2,0,0)[12] with zero mean     : Inf

Now re-fitting the best model(s) without approximations...

ARIMA(1,0,2)(2,0,0)[12] with non-zero mean : 760.2332

Best model: ARIMA(1,0,2)(2,0,0)[12] with non-zero mean
```

The parameters of our selected model [ARIMA(1,0,2)(2,0,0){12}] was estimated and we found that all the parameters significantly predict the volatility in inflation rate in Nigeria. The mathematical expression is:

$$Y_t = 12.3869 + 0.9356Y_{t-1} + e_t + 0.2617e_{t-1} + 0.0296e_{t-2} - 0.3877Y_{t-12} - 0.4071Y_{t-24}$$

The ARIMAX parameter estimates are given as:

```
Series: inflation
Regression with ARIMA(1,0,2)(2,0,0)[12] errors

Coefficients:
         ar1     ma1     ma2     sar1     sar2  intercept    xreg
      0.9375  0.2672  0.0316  -0.3954  -0.4075    10.1920  0.0127
s.e.  0.0321  0.0704  0.0878   0.0745   0.0862     1.7202  0.0056

sigma^2 = 1.551:  log likelihood = -383.6
AIC=783.2   AICc=783.84   BIC=810.85
```

We run the Ljung-Box test on the residuals from our ARIMA(1, 0, 2)$(2,0,0)_{[12]}$ model as presented in Figure 6, which returns a p-value = 0.135. Accordingly, we do not have evidence to reject the null hypothesis, and we conclude that the autocorrelations of the ARIMA(1, 0, 2)$(2,0,0)_{[12]}$'s models' residuals are jointly zero. If the residual series does not have statistically significant autocorrelations, then we cannot reject the null hypothesis. Using the unit root test p-value > 0.05 we fail to reject the null hypothesis that the residuals are normally distributed, we can also see from the shape of the histogram it is bell shaped which indicate that the residuals are normally distributed pointing to autocorrelation in the residual of the model. Since we have good reasons to believe that Inflation rate has volatility, we checked the ACF of squared values to see if they have significant autocorrelations.

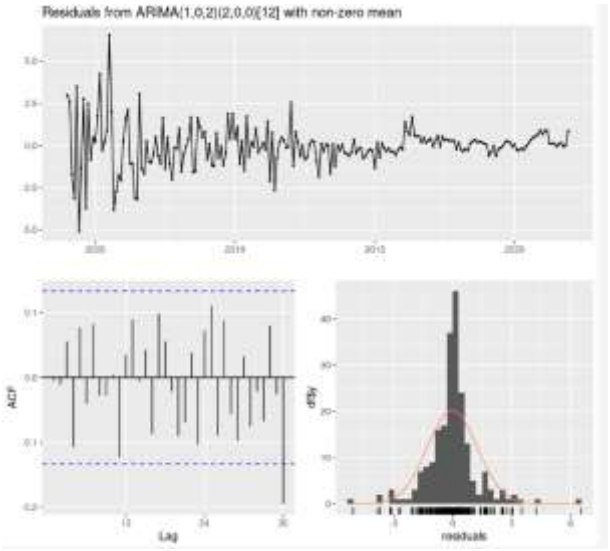


Fig. 6: Residual check for ARIMA model

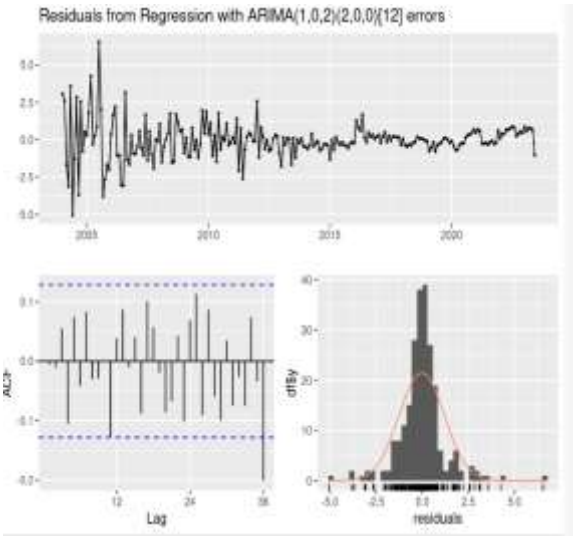


Fig. 7: Residual Check of ARIMAX model

We compare the different models' accuracy with our selected model diagnostics for short-term and long-term forecasts. The results were presented on Tables 3 and 4. This model performs very well in comparison to the other models as seen in model accuracy. The ARIMAX model is slightly more conservative in modeling the trend than other models, which accounts for sub-optimal test set RMSE performance.

Table 3: Long term model accuracy

	ARIMA	*GARCH*	*ARIMAX*	*Exp. Smoothing*	*Regression*
R-square	0.9054	0.8606	0.6278	0.9054	0.3471

RMSE	5.2649	5.2992	3.4838	6.0578	4.3310
MAE	4.5786	4.8862	3.2420	5.8259	3.7964
MAPE	0.2160	0.2354	0.1576	0.2861	0.1800

Table 4: Short Term Model Accuracy

	ARIMA	*GARCH*	*ARIMAX*	*Holt-Winter*	*Regression*
R-square	0.921	0.937	0.018	0.921	0.125
RMSE	2.814	6.771	5.197	28.533	10.634
MAE	1.205	2.184	1.614	3.618	1.092
MAPE	0.068	0.127	0.094	0.214	0.061

The monthly forecast values of selected models from January 2022 - May 2023 are presented in Table 5.

Table 5: Actual value and the forecasted value

Year	*Month*	*Observe value*	*ARIMA*	*GARCH*	*ARIMAX*	*Holt-Winter*	*Regression*
2022	Jan	15.60	15.60	14.38	16.96	12.70	15.63
2022	Feb	15.70	15.57	14.44	14.93	12.86	15.66
2022	Mar	15.92	15.54	14.50	15.12	13.03	15.65
2022	Apr	16.82	15.50	14.57	15.31	13.19	15.64
2022	May	17.71	15.47	14.64	15.43	13.35	15.64
2022	Jun	18.60	15.44	14.71	15.64	13.51	15.64
2022	Jul	19.64	15.41	14.78	15.76	13.68	15.64
2022	Aug	20.52	15.38	14.86	15.65	13.84	15.72
2022	Sep	20.77	15.35	14.94	16.23	14.00	15.86
2022	Oct	21.09	15.32	15.03	16.81	14.16	15.99
2022	Nov	21.47	15.28	15.12	17.09	14.33	16.08
2022	Dec	21.34	15.25	15.21	17.34	14.49	16.15
2023	Jan	21.82	15.22	15.31	17.73	14.65	16.30
2023	Feb	21.91	15.19	15.41	18.05	14.82	16.39
2023	Mar	22.04	15.16	15.52	18.26	14.98	16.39
2023	Apr	22.22	15.13	15.64	18.43	15.14	16.39
2023	May	22.41	15.09	15.76	18.70	15.30	16.39

2023	Jun	22.79	15.06	15.59	19.28	15.47	18.95

Discussion

The aim of this paper is to compare different forecasting models to forecast the future trend of inflation rate in Nigeria. We fit an ARIMA and GARCH family models on Nigeria Inflation rate and made forecast with the selected ARIMA(1,0,2)$(2,0,0)_{(12)}$ model based on our diagnostic checks. We also fit and forecasted with other selected models. We are interested in the model that performs best for a short-term forecast. We choose 6-month period for a short-term forecast, because this is financial data. Table 4 has the short-term forecasts and RMSE for our models. We see in Table 4 that the model with the lowest RMSE is our ARIMA model. This is interesting because the ARIMA model's predictions are the last observed value. This means that when forecasting Inflation rate, even for the next 6 months, the last observed value is a safe and accurate forecast.

Summary and Recommendations

This paper has identified ARIMA $(1,0,2)(2,0,0)_{[12]}$ and ARIMAX$(1,0,2)(2,0,0)_{[12]}$ models as the most appropriate models to forecast short-term and long-term inflation rate using the modelling approach of: model identification, selection, parameter estimation, diagnostic checking and forecasting. The two models were used to examine the inflation dynamics in Nigeria using monthly time series data from January 2004 to May 2021. The performances of the two models showed that ARIMA $(1,0,2)(2,0,0)_{[12]}$ and ARIMAX$(1,0,2)(2,0,0)_{[12]}$ models perform very well.

Based on the findings of this study, the following recommendations are made:

ARIMA (1,0,2)(2,0,0)[12] and ARIMAX(1,0,2)(2,0,0)[12], models can be applied in explaining short inflation dynamics in Nigeria and ARIMAX and regression analysis can applied in explaining long term inflation dynamics in Nigeria. It is recommended that a "one-model-fits-all" for inflation rate dynamics in Nigeria should be discouraged. Expected inflation, exchange rate, interest rate, and liquidity exert significant influence on inflation. Hence, alternative models of inflation dynamics in Nigeria should consider past inflation, exchange rate, interest rate, and liquidity as potential explanatory variables. Efforts should be made by the regulatory authorities to control money supply and ensure exchange rate and interest stability, in order to stem inflationary tendencies. The findings of this paper suggest the economic implications of rising prices, including its impact on economic growth, health and societal well-being. It is therefore, generally accepted that keeping low and stable rates of inflation is the primary objective of the central banks. Economic agents, private and public alike, closely monitor the evolution of prices in the economy, in order to make decisions that allow them to optimize the use of their resources. In this context, it is very important to model inflation rates.

REFERENCES

Ahmed, S. and Tsvetanov D. (2016). The predictive performance of commodity futures risk factors. *Journal of Banking & Finance*, **71(**2): 1-36.

Aklin M., Arias E., Gray J. (2022). Inflation concerns and mass preferences over exchange-rate policy
Econom. Politics, 34 (1): 5-40.

Amadeh H, Amini A, Effati F. ARIMA and ARFIMA prediction of Persian Gulf Gas-Oil F.O.B. *Investment Knowledge* **2(**7): 212-231

Box, G. E. P, Jenkins, G. (1970). Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA.

Caraiani P. (2010). Forecasting Romanian GDP using a BVAR model. *Romanian Journal of Economic Forecasting*. **4**:76-87.

Cochrane, J. H. (1997). *Time Series for Macroeconomics and Finance*. Graduate School of Business, University of Chicago.

Di Tella R., MacCulloch R.J., Oswald A.J. (2001). Preferences over inflation and unemployment: Evidence from surveys of happiness. *Am Econ Rev*, 91 (1): 335-341.

Doepke M., Schneider M. (2006). Inflation and the redistribution of nominal wealth. *J. Polit. Econ.*, 114 (6): 1069-1097.

Engle R. F, Granger C. W. J. (1987) Cointegration and error correction: representation, estimation and testing. *Econometrica*. **55**(2): 251-276

Gandelman N., Hernandez-Murillo R. (2009). The impact of inflation and unemployment on subjective personal and country evaluations. *Review*, **91**: 107-126.

Hamilton, J. D. (1989). A new approach to the economic analysis of non-stationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, vol. **1(**1): 342-8.

Hamzacebi, C. (2008). Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, **178**: 4550-4559.

Kongcharoen C, Kruangpradit T. (2013). Autoregressive integrated moving average with explanatory variable (ARIMAX) model for Thailand export. A paper presented at the 33rd International Symposium on Forecasting, South Korea P. 201

Ljung, G. M., Box G. E. P. (2002). On a measure of lack of fit in time series models. *Biometrika*. **69**: 297- 303.

Onyeka-Ubaka, J. N., Halid, M. A., Ogundeji, R. K. (2021). Optimal Stochastic Forecast Models of Rainfall in South-West Region of Nigeria. *International Journal of Mathematical Analysis and Optimization: Theory and Applications*, https://doi.org/10.52968/28306097; 7(2): 1-20.

Park, S., Lee, W. and Moon, I. C. (2015). Associative topic models with numerical time series. *Information Processing & Management*, **51**(5): 737-55.

Pivarski, J., Bennett, C. and RL. Grossman, R. L. (2016). Deploying Analytics with the Portable Format for Analytics (PFA). *Proceedings of the International Conference of Knowledge Discovery and Data Mining*,

Raicharoen, T, Lursinsap, C, Sanguanbhoki, P. (2003). Application of critical support vector machine to time series prediction", Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on Volume 5, 25-28 May, 2003, pages: V-741-V-744.

Salvatore D, Reagle D. (2002). Schaum's Outline of Theory and Problems of Statistics and Econometrics. 2nd ed. New York: McGraw-Hill.

Yaya, O. S., Fashae, O. A. (2014). Seasonal fractional integrated time series models for rainfall data in *Nigeria. J. Theoret. Appl. Climatol*. **120**(2): 99–108

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* **50**: 159–175.

A COMPARISON OF BAYESIAN REGRESSION AND CLASSICAL ORDINARY LEAST SQUARES METHODS FOR MULTIPLE LINEAR REGRESSION USING POST-UTME DATA

C. O. Odijie[A*] and Ekhosuehi[B]

[a, b] Department of Statistics, University of Benin, Benin City.

[*] Corresponding Author: cyril.odijie@uniben.edu

Abstract

In this paper, the scores of post-UTME candidates of the University of Benin in the 2023/2024 academic session are modelled with ordinary least squares (OLS) method and its Bayesian multiple linear regression counterpart. The predictors envisaged include candidates' first choice of the course of study ($course1$), age at the time of the screening examination (age), $male$ (an indicator variable for male), and $utme$ (Unified Tertiary Matriculation Examination) score. The population involves 13730 students whose required details were completely available. The results were basically similar in the two methods, which should be expected since the whole population size was analyzed and is large enough for a robust inference leading to a convergence in results of the two methods, with JASP default non-informative beta prior ($\alpha = \beta = 1,\ r = 0.354$) used in the Bayesian analysis. All the regression coefficients were statistically significant, ($\beta_{course1}$ = 0.0002516, t = 2.606, p=0.009; β_{age} = 0.099, t = 3.366, p < 0.001; β_{utme} = 0.061, t = 22.312, p < 0.001; β_{male} = - 0.456, t = -3.144, p = 0.002), although they had poor explanatory power on post-UTME scores (F = 154.952, p < 0.001).

Keywords: Bayesian regression, multiple linear regression, estimation, OLS, post-UTME

1. Introduction

When one stochastic variable, usually referred to as the response, predicted or dependent variable in different parlances, hinges on the values of two or more observed variables, usually called the input, predictor or independent variables, a statistical method known as multiple regression is usually conceived, as first proposed by Galton (1886). More specifically, when the model so described is linear in parameters, the model is known as multiple linear regression (MLR) model. Since the renaissance of the alternative school of thought in Statistics, known as Bayesian Statistics, in the second half of the 20th century (Cowles et al., 2019), many aspects of Classical statistics, including MLR, have been retouched with even more promising inference and estimation methods (O'Neill, 2002; Kruschke, 2012; Zhang & Huang, 2016, among others). Of course, the two methods have their relevance in their own way. In fact, it can be shown that results from the two methods in estimation converge when the sample size is large enough (asymptotic convergence). However, it has been demonstrated that Bayesian methods triumph over Classical methods in the face of small samples, giving more credence to the use of correctly specified priors which are peculiar only to Bayesian methods (Kruschke, 2012). In this study, the scores obtained by candidates who sat for the 2023/2024 sessional screening examination

(post-UTME) of the University of Benin, Benin City are modelled with Bayesian method of handling MLR emphasized in comparison with the well-known Classical method of ordinary least squares (OLS). We expect that the results of the two methods will not differ significantly due to the size of the population of study, since the results of both methods are basically the same when this is the case, no matter the choice of priors in Bayesian method. A similar study that considered the post-UTME scores of candidates in relation to only their performance in the prerequisite exam (UTME) across three academic sessions was earlier considered by Odijie and Ekhosuehi (2025).

The succeeding part of this article will take the following structure: section 2 encapsulates the review of relevant studies involving multiple regression and its related studies, section 3 details the methods applied in this study while sections 4 and 5 present the data analysis with discussion of results and the concluding remarks, respectively.

2. Literature Review

Notable studies in both Classical and Bayesian methods for handling MLR models exist in the literature including, but not limited to, Jobson (1991), Bewik et al. (2003); Crawford (2006); O'Brien & Scott (2012); Bezuidenhout & Domleo (2013), Lloyd-Jones et al. (2019), Mbete (2020) and Bergh et al. (2021). In classical methods, it is common to find least squares as prominent method of estimating the parameters of the model alongside with a few others such method of moments and maximum likelihood (Luo, 2016). The least squares method, or ordinary least squares (OLS) method, to delineate it from weighted least squares, minimizes the sum of squares of the random error component of the model. The likelihood function of the model is maximized with respect to the parameter of interest in maximum likelihood method. Method of moments replaces population moments with sample moments in terms of the parameters. In Bayesian contest, on the other hand, an entirely different route is plied – distributions are constructed for parameters of interest which are treated stochastically, while samples drawn from known probabilistic models are taken as fixed after they have been drawn or observed. Point estimates are then obtained from the distribution constructed, as expected values. Based on the idea of subjective probability as first proposed by de Finetti (1931, 1937), Bayesians are at a procedurally-guided liberty to assign a measure of uncertainty known as prior probability or distribution (depending on the nature of the phenomenon or hypothesis of interest) to parameters, a priori. Then, on evidence or information from data (a given sample) the prior uncertainty, or simply the prior probability/distribution is updated for a new one, given the data (evidence) and its likelihood. The resulting probability or probability distribution is what Bayesians call the posterior distribution, on which all inferences are based, such as point estimation and prediction. This is clearly not the case in classical statistics where parameters are seen as fixed immutable values that can only be estimated as point values while constructing some interval of confidence for them, given some probability of false positive cases (aka level of significance) (Lambert, 2018). On the issue of correct or appropriate specification of priors, several methods and guidance have been elicited, including the use of non-

informative priors such as Jeffreys' prior (see Chen et al., 2008 for a good number of literature on this) and informative priors from expert opinions, among others. Also the use of priors has been well justified by notable "fathers" of modern Bayesian statistics such as de Finetti (1931, 1937), Savage (1954), Anscombe and Aumann (1963), Beach et al. (1970), Kahneman and Tversky (1972), Hampton et al. (1973), Neth (2024). These and others have been able to quell the doubt and the question sceptics have asked, namely, "why priors?" and thus resulted in the prevalence of Bayesian thinking/methods in modern Statistics, especially in the computational and applied aspects of Archaeology (Otarola-Castillo and Torquato, 2018), Ecology (Bon et al., 2023), Engineering (Swiler, 2006; Wu et al., 2018), Geophysics (Fang, 2020), Health (Lambert, 2018), Machine Learning (Bharadiya, 2023), Marketing (Green & Frank, 1966), Psychology (van da Berg et al., 2003; Etz & Vandekerckhove, 2018), and Radiology (Donovan and David, 2007; Ma et al., 2023) among very many others.

3. Methodology

3.1. Data

A dataset containing information of candidates who registered for the University of Benin admission screening examination, commonly referred to as post-UTME, was obtained from the Central Records Processing Unit (CRPU) of the University. It was observed that some candidates who registered for the screening did not finally sit for the exams. Since the main variable of interest in this study is the scores of the candidates who took the screening examination, these incomplete records, which would ordinarily be considered as missing cases where totally expunged on the grounds that the population size is known and quite large enough and the missing cases were relatively few to have any significant impact on the results of the analysis. After the "data cleaning" was done, a total population of 13,730 complete records formed the population of study. Population properties were immediately obtained to form a basis for assessing the perfomance of the Bayesian regression and classical OLS that will be conducted.

3.2. Hypotheses

The interest is to ascertain how well the variables, (i) first choice course of study ($course1$), (ii) age of candidate at screening (age) (iii) UTME score of candidate ($utme$) (iv) sex of candidate, (represented by an indicator variable "$male$"), collectively predict the score of a student in the screening exam. Hence, the null hypothesis is that the variables, collectively, do not predict the score of a candidate versus the alternative hyposthesis of interest that they collectively do, in a linear regression model. In other words, the null hypothesis portends that all coefficients of the model are mutually equal to zero. For the Classical analysis, we choose the level of significance for rejecting the null hypothesis to be $\alpha = 0.05$.

3.3. Model

Obviously, the data and hypotheses point directly to a multiple linear regression model, which we state as follows.

$$score_i = \beta_0 + \beta_{course1} * course1_i + \beta_{age} * age_i + \beta_{utme} * utme_i + \beta_{male} * male_i + \varepsilon_i \quad (1)$$

where,

$score_i$ is the post-UTME screening score of the ith candidate,

$course1_i$ is the first choice course of study of the ith candidate,

age_i is the age at post-UTME screening exam of the ith candidate,

$male_i$ is an indicator variable (male = 1 or not = 0) of the ith candidate, and

$\beta_0, \beta_{course1}, \beta_{age}, \beta_{utme}\ and\ \beta_{male}$ are the coefficients of the regression model.

First, noting that the variables $course1$ (originally recorded as course labels) and $male$ (originally recorded as 'sex', with 'm' or 'f' as entries), we first convert or code them into numerical values for the analysis. The $course1$ variable was converted via count encoding where each entry is replaced by the its total count (frequency) while the variable "sex" was replaced by the dummy or binary or indicator variable $male$, with value 1 when true for the ith candidate and 0 otherwise.

3.4. Analysis

The analysis is carried out mainly in JASP, a user-friendly and free statistical software that provides platforms for both Classical and Bayesian analysis, thereby facilitating easy comparison of results. However, some of the diagnostic plots for the population properties were carried out in Python, leveraging on its computing capability and the beauty of plots made by one of its plotting packages, matplotlib.

3.4.1. Classical OLS

The $score$ variable is moved to the dependent variable entry field while the remaining variables are moved to the entry field of the covariates. The parameters of the model are automatically estimated by JASP thereafter. The model summary and coefficients table are displayed as outputs immediately.

3.4.2. Bayesian Regression

For the Bayesian regression, we rely on the default priors defined in the software, namely, the beta prior with both shape parameters equal to 1. This corresponds to choosing a flat prior (assuming equally likely probability of each possible coefficient value over the real line), on the grounds that the data size is large enough to "speak for itself". Also partly due to the fact that not much is known about the distribution of the parameters a priori. Based on this setting, the results of the two methods are expected to be quite close or the same.

4. Results and discussion

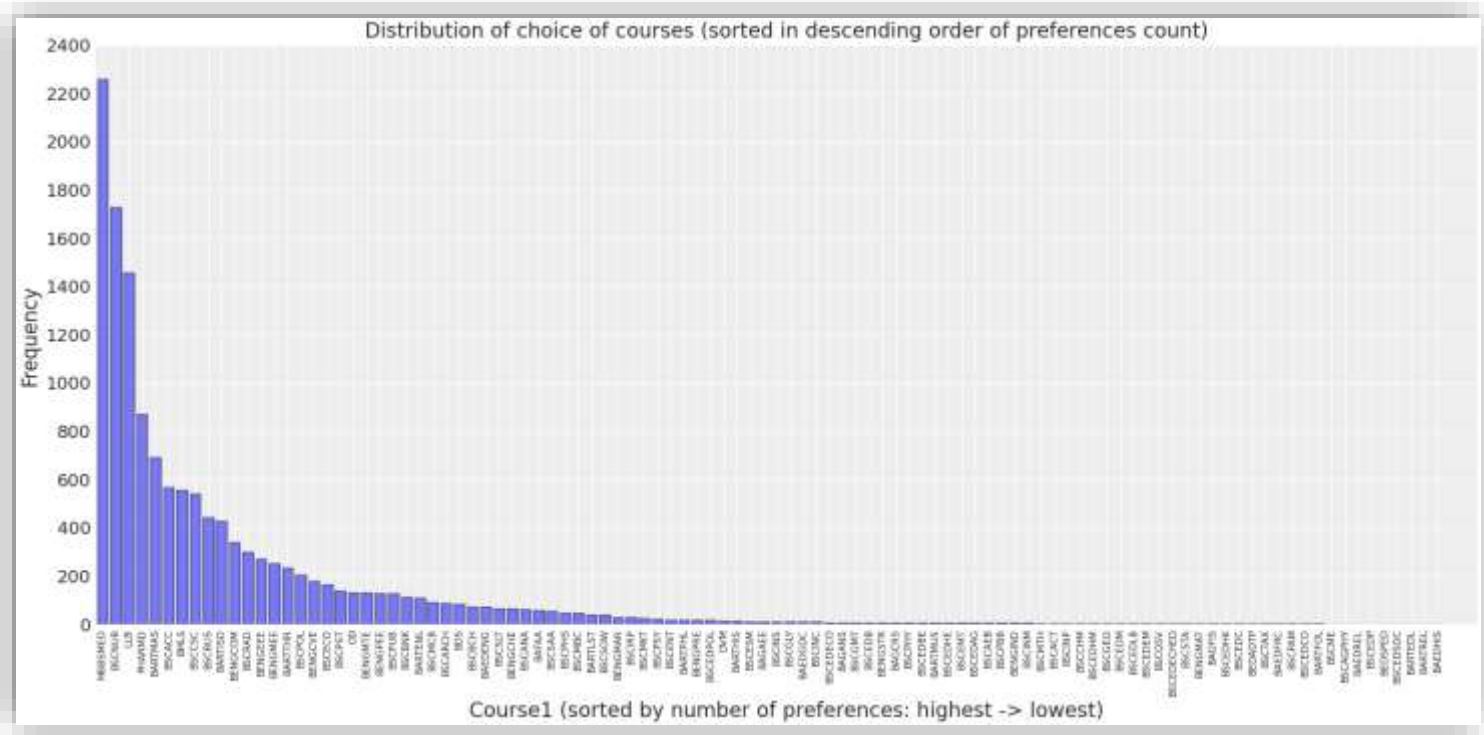
We harnessed software computing power to explore the population properties directly. The results are displayed in Table 1.

Table 1: Population properties of the variables of the study

	course1_coded	age	utme	male	score
Valid	13730	13730	13730	13730	13730
Mode	2259.000 [a]	18.000[a]	203.000[a]	0.000 [a]	50.000 [a]
Median	693.000	18.000	226.000	0.000	53.340
Mean	978.715	18.902	232.746	0.438	53.911
Std. Deviation	784.891	2.419	27.392	0.496	8.432
Skewness	0.447	3.637	1.108	0.252	0.220
Minimum	1.000	15.000	200.000	0.000	3.340
Maximum	2259.000	55.000	352.000	1.000	100.000
25th percentile	271.000	17.000	211.000	0.000	50.000
50th percentile	693.000	18.000	226.000	0.000	53.340
75th percentile	1727.000	20.000	248.000	1.000	59.000

The following are immediately obvious from Table 1.

a. The maximum count of preferences for a particular course of study as first choice is 2259. This is actually the total number of candidates that applied for Medicine (labelled MBBSMED in the dataset) as may be seen in Figure 1 (a bar plot of all the courses applied for, and their preferences count). Whereas, at least one course of study had only 1 preference count.
b. The mean age of all the candidates was 18.9 years, with a minumum age of 15 years and an outlying maximum of 55 years (see boxplot in Figure 3)
c. The minimum UTME score was 200, which is actually the cut-off mark set by the institution before an applicant can apply for post-UTME screening exam. The maximum UTME score was 352. The mean UTME score was 232.746 for that session.
d. The mean (0.438 approx.) of the indicator variable ($male$) indicates that 0.438 x 13730 $\approx$ 6014 were males, which is less than the number of females (7716 approx.). The actual numbers are 6008 males and 7722 females (difference is due to approximation error).
e. The average score for post-UTME (labelled $score$) was 53.911, with an appalling minimum of 3.34 and a perfect maximum of 100.



f. The 25%, 50% and 75% percentiles in the first column, represent the first, second and third quartiles, Q1, Q2=median and Q3, respectively, which are the partitions of the data. The median values represent better central tendency, especially for the distribution of "age" and "utme" variables which are the most positively skewed (age having the

g. largest coefficient of skewness = 3.637) and with few outliers as shown in the boxplots of Figures 2 & 3. Hence the median post-UTME score, age and UTME score were 53.34, 18 years and 226, respectively. For the variables "course1_coded" and "male" which are the converted numerical values in columns 2 and 4 respectively, the median value corresponds to the category with the respective code value, that is BARTMAS (Mass Communication) with code 693 and female (code 0 i.e. not male as an indicator).

Figure 1: Bar plot of first choice course of study of the candidates.

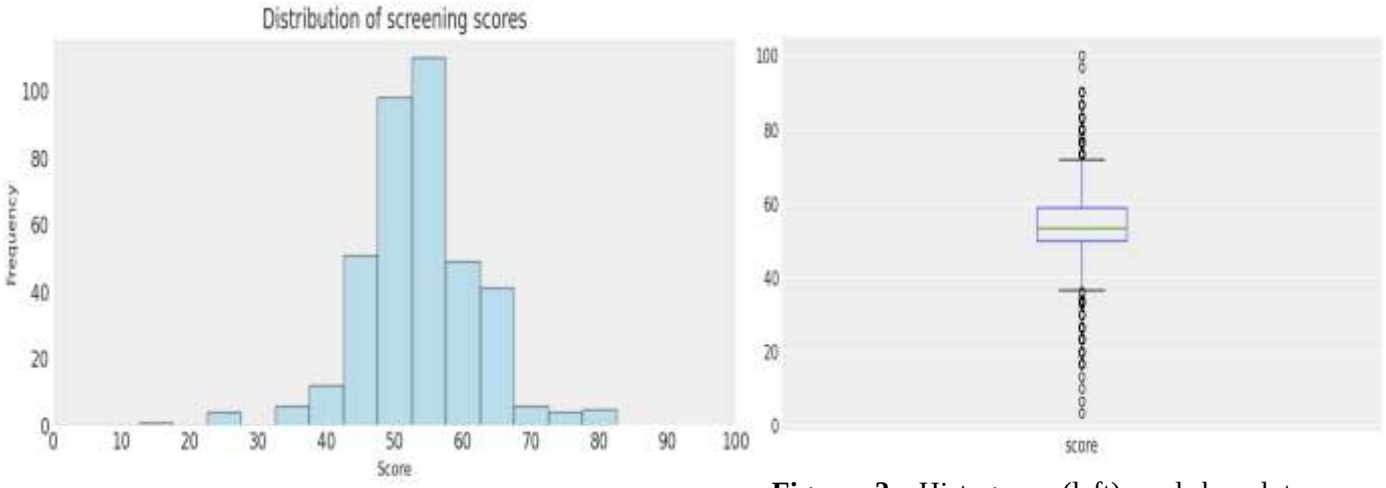


Figure 2: Histogram (left) and boxplot (right) of screening scores

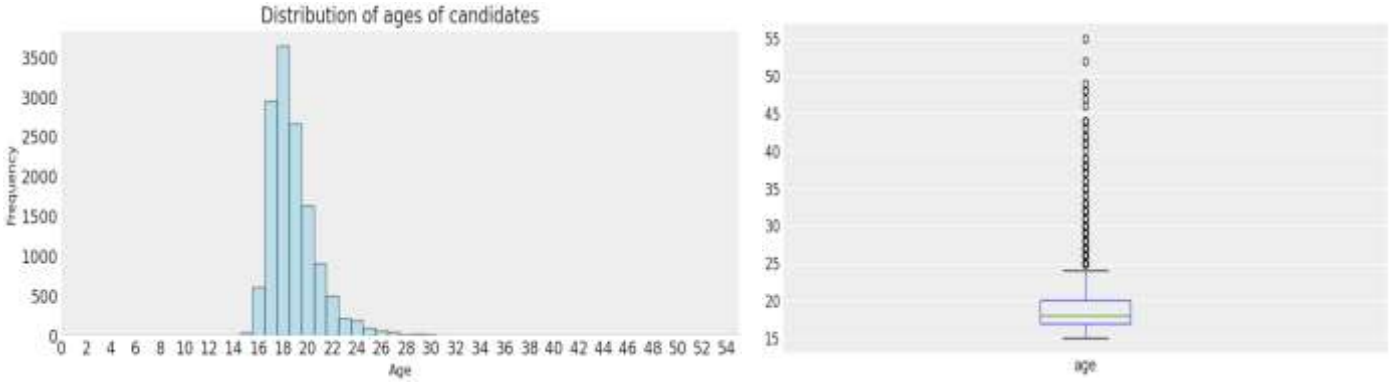


Figure 3: Histogram (left) and boxplot (right) of the ages of the candidate

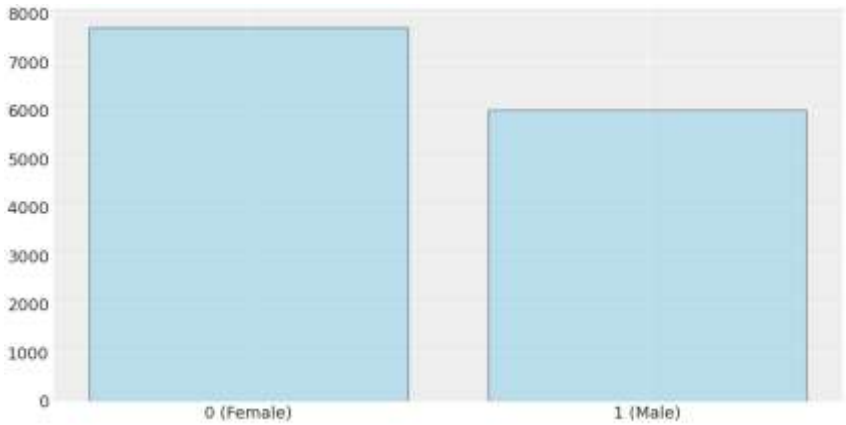


Figure 4: Bar plot of the indicator variable (male)

4.1 Classical OLS results:

The results of the classical MLR with OLS method used for estimating the coefficients of the regression model are given and discussed in this subsection, starting with the model summary.

Table 2: Model Summary - score

Model	R	R^2	Adjusted R^2	RMSE
M_0	0.000	0.000	0.000	8.432
M_1	0.208	0.043	0.043	8.249

Note. M_1 includes course1_coded, age, utme, male

In Table 2, M_0 is the model corresponding to the null hupothesis that all paramter effects are mutually zero while M_1 is that of the alternative (the interest of the study) involving the four predictors listed as a note under the table by JASP. Clearly, the model suggests a very poor explanatory power of the collective variables on post-UTME scores with a meager 4.3% explanation of the variations in the response variable (post-UTME scores) shown by both the R^2 and the adjusted R^2. We can see that the root means square error, a measureof the average deviation of the predicted values from the observed values, when all the variables are hypothesized to have zero effect on post-UTME scores is almost the same as when they are assumed to have some explanation.

In Table 3, the analyis of variance results are displayed. The results show a statistically (but not practically) significant value with $F(4, 13725) = 154.952$, $p < 0.001$. We have said "not practically significant" to butress the fact that although there exists some explanation of the variation in post-UTME scores due to the explanatory variables, it is relatively non-existent when compared to the portion of unexplained variation. Of course, this is one point of limitation of simply rejecting the null hypothesis or failing to do so by the use of p-values, on which Bayesians commonly fault classical hypothesis testing. There is no quantification of the size of significance in probability terms as in Bayesian paradigm. The statement "probability of observing a statistic as extreme as or more extreme than the test statistic", which defines the p-value is vague as there is no clear meaning to the statement in terms of quantifying uncertainly.

Table 3: ANOVA

Model		Sum of Squares	df	Mean Square	F	p
M_1	Regression	42175.098	4	10543.774	154.952	< .001
	Residual	933923.230	13725	68.045		
	Total	976098.327	13729			

The coefficients table (Table 4) further explains what we had discussed in the forgoing (for Table 3). All t-values i.e. test statistics of the coefficients show high statistical significance even in the face of poor overall performance (F-value) and near-zero coefficient estimates (columns 3 and 5) under model

M_1 (alternative hypothesis model). Before we quickly push the blame on the presence of outliers and other possible irregularities in the dataset, we must recall that the entire population was analyzed and these effects are thereby of no strong effect as there are no sampling errors (or sampling variations) in this case.

Table 4: Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M_0	(Intercept)	53.911	0.072		749.175	< .001
M_1	(Intercept)	37.777	0.822		45.940	< .001
	course1_coded	2.516×10^{-4}	9.656×10^{-5}	0.023	2.606	0.009
	age	0.099	0.029	0.028	3.366	< .001
	utme	0.061	0.003	0.198	22.312	< .001
	male	-0.456	0.146	-0.027	-3.114	0.002

In fact, the null model, M_0 alludes to the fact that M_1 explains virtually nothing since the population mean already corresponds to the intercept of the model with the value of 53.911 (refer to Table 1 again for this value in the column for the mean score) and is statistically significant. All values of the estimates are close to zero, indicating the fact that for a unit change in each of the explanatory variable, the response variable only changes by a near-zero amount. Interestingly, the negative value of the variable "male" suggests that post-UTME score is negatively affected (i.e. 0.456 decrease in score) if the ith candidate is a male.

The resulting estimated regression line is given, from equation (1), as

$$score_i = 37.777 + 0.0002516course1_i + 0.099age_i + 0.061utme_i - 0.456male_i \quad (2)$$

4.2. Bayesain MLR method results

Next, we consider the results of the MLR analysis with Bayesian method. The table of results are a little bit different in presentation from their classical counterparts. The model summary show different combinations of the explanatory variables, their posterior probabilities and their Bayes factors.

Table 5: Model Comparison - score

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Null model	0.200	3.357×10^{-126}	1.343×10^{-125}	1.000	0.000
course1_coded + age + utme + male	0.200	0.749	11.906	$2.229\times10^{+125}$	0.043
age + utme + male	0.050	0.163	3.688	$1.937\times10^{+125}$	0.043
course1_coded + age + utme	0.050	0.038	0.754	$4.547\times10^{+124}$	0.043
utme + male	0.033	0.019	0.555	$3.355\times10^{+124}$	0.042
course1_coded + utme + male	0.050	0.017	0.326	$2.011\times10^{+124}$	0.042
course1_coded + utme	0.033	0.007	0.214	$1.306\times10^{+124}$	0.042
utme	0.050	0.004	0.080	$5.002\times10^{+123}$	0.041
age + utme	0.033	0.004	0.104	$6.412\times10^{+123}$	0.042
course1_coded + age	0.033	7.980×10^{-105}	2.314×10^{-103}	$1.426\times10^{+22}$	0.008
course1_coded	0.050	9.042×10^{-106}	1.718×10^{-104}	$1.077\times10^{+21}$	0.008
course1_coded + age + male	0.050	4.734×10^{-106}	8.995×10^{-105}	$5.640\times10^{+20}$	0.009
course1_coded + male	0.033	2.300×10^{-107}	6.669×10^{-106}	$4.110\times10^{+19}$	0.008
age	0.050	5.647×10^{-127}	1.073×10^{-125}	0.673	0.001
age + male	0.033	2.556×10^{-128}	7.412×10^{-127}	0.046	0.001
male	0.050	2.533×10^{-128}	4.813×10^{-127}	0.030	0.000

In Table 5, P(M) is the prior probability assigned to the model before data is observed. They are assumed to be equally likely with respect to the number of explanatory variables combined and they should approximately sum to 1 since they form a discrete probability distribtuion. P(M|Data) is the posterior probability, i.e., the probability of each model after the data have been incorporated in the model. The model with the highest value is considered the best. BF_M is the ratio of the posterior odds to the prior odds for each model. It is not of main interest in decision-making here as are P(M|Data) and BF_{10}. BF_{10}

is the Bayes Factor for each model in relation to the the null model. It is the ratio of the likelihood of the data under each model to the null model. A value greater than 100 marks an extreme evidence for the alternative model given the data. A value between 1 and 3 indicates an anecdoctal evidence for the alternative model given the data. A value of 1 means equal evidence from both the null and alternative model. This last case is always true for the null model (i.e. it compares to itself). See Bergh et al. (2021) for a table showing detailed interpretation of different ranges of Bayes factor. Lastly, the R^2 values for each model are given in the last column of the table. It simply retains its usual meaning of the variation in the response variable attributable to the model in question.

Clearly, with meanings given to each measure, we can see that the best model corresponds to the classical OLS model involving all four explanatory variables, with the posterior probability P(M|Data) = 0.749 and Bayes Factor, $BF_{10} >> 100$. The R^2 is exaclty 0.043 as in OLS. In Bayesian context, we can make a clear probabilistic statement to quantify our uncertainty on this model after observing the data: "we are about 74.9% certain that the post-UTME scores were, or can be, realized under the second model involving all the explanatory variables."

Table 6. Posterior Summaries of Coefficients in the best model

Coefficient	P(incl)	P(excl)	P(incl|data)	P(excl|data)	$BF_{inclusion}$	Mean	SD
Intercept	1.000	0.000	1.000	0.000	1.000	53.911	0.070
course1_coded	0.500	0.500	0.811	0.189	4.288	2.511×10^{-4}	9.646×10^{-5}
age	0.500	0.500	0.953	0.047	20.203	0.099	0.029
utme	0.500	0.500	1.000	0.000	$1.066\times10^{+104}$	0.061	0.003
male	0.500	0.500	0.947	0.053	17.775	-0.455	0.146

Table 6 shows the posterior summary of the coefficeints in the best model. In the table, "P(inc)" and "P(exl)" are the prior probabilities of including the variable and excluding the variable, respectively. "P(inc|data)" and "P(exl|data)" are the posterior probabilities of including the variable and excluding the variable, respectively. The model under Bayesian MLR is obtained from Table 6 ("Mean" column) and the standard deviation (SD) of the mean is given in the last column of the table. Just like the Classical OLS results, the results suggest that all coefficeints should be included in the model. It is clear that the probabilty of including each of the variable increased from the default 0.5 (i.e. 50-50) to nearly. In fact, "utme" coefficient has posterior probability of inclusion exactly equal to 1. Hence, it is sufficient to retain the model as given in equation (2) for the Classical OLS method since they are practically the same.

4.3 Prediction

Here, we predict the scores of 10 randomly selected candidates from the population based on the estimated model in equation (2). The results are given in Table 7.

Table 7. Predicted scores of 10 randomly sampled candidates from the dataset.

course1_coded	age	utme	male	score	pred_score	pred_error
72	23	213	0	53	53.0651152	-0.0651152
84	19	235	1	43.34	53.5581344	-10.218134
541	16	298	0	40	57.6751156	-17.675116
693	17	223	0	46.68	53.2373588	-6.5573588
271	19	206	1	59	51.8361836	7.1638164
566	19	237	0	56.68	54.2574056	2.4225944
2259	17	240	0	53.34	54.6683644	-1.3283644
1727	18	205	0	50	52.4985132	-2.4985132
870	21	256	1	53.34	55.2348920	-1.8948920

The predicted score (pred_score) of the 10 randomly selected candidates showed that scores which are close to the mean score of 53.911 had smaller prediction error (pred_error), which is the difference between the observed and the predicted values. The largest prediction error in absolute terms (17.68 approx.) occurred in the third case with observed score of 40 and predicted score of 58 approximately. We cannot say that the model is a good fit, firstly because R^2 is small as earlier seen and secondly, these predicted values are almost one-sided (no matter the observed score, the predicted scores are mostly higher than the observed score).

Conclusion

The study modelled the scores of post-UTME candidates applying for admission into the University of Benin, Benin City in the 2023/2024 academic session. The entire population of 13730 candidates was analysed in order to unravel the true properties of the population since few abnormalities, including minor skewness of some of the variables and outliers, were noticed. It was observed that both classical OLS and Bayesian regression carried out with the statistical software, JASP, had similar results due to the settings of the data and analysis procedure. The coefficients of the model, though statistically significant, possessed poor explanation of the variations in post-UTME scores of the candidates. The predictors included first choice course of study, age at the screening exam (post-UTME), male (whether candidate is a male or not) and UTME score.

References

Anscombe, F. J., & Aumann, R. J. (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics*, **34**:199-205

Beach, L. R., Wise, J. A., & Barclay, S. (1970). Sample proportions and subjective probability revisions. *Organizational Behaviour and Human Performance.* **5**(2): 183-190.

Bergh, D. V. D., Clyde, M. A., Gupta, A. R. K. N., de Jong, T., Gronau Q. F., Marsman, M., Ly A., & Wagenmakers E. J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behav Res Methods*. **53**(6):2351-2371.

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., & Domleo, R. R. (2013). A demonstration of correlation graphs to human body dimensions. *Scientific Research and Essays*. **8**(27): 1273-1281. https://doi.org/10.5897/SRE2013.5398

Bharadiya, J. P. (2023). A Review of Bayesian Machine Learning Principles, Methods, and Applications. *International Journal of Innovative Science and Research Technology.* **8**(5): 2033-2038.

Bon, J. J., Bretherton, A., Buchhom, K., Cramb, S., Drovandi, C., Hassan, C., Jenner, A. L., Mayfield, H. J., McGree, J. M., Mengersen, K., Price, A., Salomone, R., Santos-Fernandez, E., Vercelloni, J., & Wang, X. (2023). Being Bayesian in the 2020s: opportunities and challenges in the practice of of modern applied Bayesian statistics. *Phil. Trans. R. Soc. A* ***381:*** 20220156. https://doi.org/10.1098/rsta.2022.0156

Chen, M. H., Ibrahim, J. G., & Kim, S. (2008). Properties and Implementation of Jeffreys's Prior in Binomial Models. *J Am Stat Assoc.* **103**(484): 1659-1664.

Cowles, K.; Kass, K., & O'Hagan, T. (2019). What is Bayesian Analysis? https://bayesian.org/what-is-bayesian-analysis/

Crawford S. L. (2006). Correlation and Regression. *Circulation.* **114**(19):2083-8 https://doi.org/10.1161/circulationaha.105.586495

de Finetti, B. (1931). Funzione caratteristica diun fenomeno allatorio. Atti delta R. Accademia Nazionale dei Lincii Ser. 6, Memorie, Classe di Scienze, Fisiche, Matematiche e Naturali, **4**: 251-299.

de Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives, Anales de l'Institut Henri Poincare, **7**: 1-68

Donovan, T., & David, J. (2007). The radiology task: Bayesian theory and perception. *British Journal of Radiology.* **80**(954): 389-391.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Phychonomic Bulletin & Review.* **25**(1): 5-34. https://doi.org/10.3758/s13423-017-1262-3

Fang, Z., Fang H., & Demanet, L. (2020). Chapter Five – Deep generator priors for Bayesian seismic inversion, Moseley B, Krischer L (eds). *Advancement in Geophysics*, Elsevier. **61**: 179-216. https://doi.org/101016

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland*. **15**: 246–263

Green, P. E. & Frank, R. E. (1966). Bayesian Statistics and Marketing Research. *Journal of the Royal Statistical Society. Series C (Applied Statistics).* **15**(3): 173-190

Hampton, J. M., Moore, P. G., & Thomas, H. (1973). Subjective Probability and its Measurement. *Journal of the Royal Statistical Society.* **136**(1): 21-42

Jobson, J. D. (1991). Multiple Linear Regression. In: Applied Multivariate Data Analysis, *Springer*, New York, 219-398. https://doi.org/10.1007/978-1-4612-0955-3

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgement of representativenesss. *Cognitive Psychology.* **3**(3): 430-454

Kruschke, J. K. (2012). Bayesian Estimation Supersedes the t Test. *Journal of Experimental Psychology*: *General*. *Advance online publication*. https://doi.org/10.1037/a0029146

Lambert, B. (2018). A Student's Guide to Bayesian Statistics. SAGE.

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J. & Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* **10**:5086 https://doi.org/10.1038/s41467-019-12653-0

Luo, X. (2016). A Comparison of Three Estimation Methods in Linear Regression Analysis. https://doi.org/10.2991/ICMMITA-16.2016.92

Ma, S. X., Dhanaliwala, A. H., Rudie, J. D., Rauschecker, A. M., Roberts-Wolfe, D., Haddawy, P., & Kahn Jr, C. E. (2023). Bayesian Networks in Radiology. *Radiology: Artificial Intelligence.* **5**(6):e210187

Mbete, D. A. (2020). Parameter Estimation of Bayesian Multiple Regression Model with Informative Inverse Gamma Prior Distribution: Application to Malaria Symptom Dataset. *Asian Journal of Probability and Statistics*. **7**(1):71-86.

Neth, S. (2024). Better Foundations for Subjective Probability. *Austrialian Journal of Philosophy.* **103**(1): 1-22

O'Brien, D. & Scott, P. S. (2012). Correlation and Regression. In Approaches to Quantitative Research – A Guide for Dissertation Students, Ed, Chen, H. Oak Tree Press.

Odijie, C. O. and Ekhosuehi, N. (2025). On the relevance of post-UTME screening examination in Nigeria: A case study of the University of Benin, Benin City. *Proceedings of the 1st annual conference of the Chartered Institute of Statisticians of Nigeria.* Abuja Nigeria. 285-295.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math Biosci.* **180**: 103-14. https://doi.org/10.1016/s0025-5564(02)00109-8

Otarola-Castillo, E. & Torquato, M. G. (2018). Bayesian Statistics in Archeology. *Annual Review of Anthropology*. **47**: 435-453

Savage, L. J. (1954). The foundations of statistics. John Wiley & Sons.

Swiler, L. P. (2006). Bayesian methods in engineering design problems. https://doi.org/10.2172/883142

van den Berg, A. E., Sander, L. K., & van der Wulp, N. Y. (2003). Environmental preference and (How) are they related? *Journal of Environmental Psychology.* **23**(2):135-146

Wu, X., Kozlowski, T., Meidani, H., & Shirvan, K. (2018). Inverse uncertainty quantification using modular Bayesian approach based on Gaussian process, Part 1: Theory. *Nuclear Engineering and Design.* **335**: 339-355. https://doi.org./10.1016/j.nucengdes.2018.0.6.004

A COMPOUND DISTRIBUTION FOR MODELING DISCRETE FAILURE EVENTS

Elebe E. Nwezza[1*], Uchenna U. Uwadi[1], Chukwunenye I. Okonkwo[1], E. J. Ekpenyong[2], Kelechi E. Arua[1], Ikenna E. Chimezie[1], Nnajiofor C. Nwezza[3]

[1]Department of Mathematics and Statistics, Alex Ekwueme Federal University Ndufu-Alike, Ebonyi State.

[2]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Abia State.

[3]Department of Electronic Engineering, University of Nigeria, Nsukka, Enugu state.

[*]**Corresponding author's email**: elebe.nwezza@funai.edu.ng

Abstract

In modeling discrete outcomes, classical discrete distributions are usually associated with a constant parameter defining the probability of occurrence, as in the case of Bernoulli, binomial, geometric, etc., distributions, or the rate of occurrence, as in the case of Poisson distribution. In this study, we generated a new distribution for modeling the occurrence of discrete failure events with varying probabilities. Some properties of the new distribution, such as the moments about the origin and the probability-generating function, were studied. The applicability of the new distribution was considered using two real-life datasets, and comparisons with other existing discrete distributions were performed using goodness-of-fit statistics to determine how well the distribution fitted the datasets. The comparison results show that the new distribution provides a better fit among the competing distributions.

Keywords: Discrete failure event; Geometric distribution; Standard power function distribution; Prior distribution; Posterior distribution;

I. Introduction

Classical distributions, such as Bernoulli, binomial, Poisson, negative-binomial, and geometric distributions, were generated to model specific discrete outcomes with each element in a sample of interest having equal and independent probabilities, P or rate λ, of occurrence. These classical distributions have been applied in failure or survival analysis [1, 2] and in reliability, where interest has been on manufactured units generated by different processes [3].

The equal and independent probability or rate of occurrence of each of the discrete outcomes assumed by these classical discrete distributions may not be appropriate in real-world applications owing to the inherent stochastic nature of the sample units. For instance, in reliability analysis, interest may be on the number of defective items. The assumption that all items have equal and independent probabilities of being defective may be misleading, as each item may have different factors or factors that may be responsible for the defect, and this may cause each item to have different failure times. Bakouch et al.

[4] noted that classical distributions, including geometric and Poisson distributions, have limited applicability as models for reliability, failure times, and counts. Lawless [2, p.33] also noted that, in a heterogeneous population where degenerating components exist, estimation may be difficult because each homogeneous component is assumed to have a different survivor function.

Some discrete distributions, such as Beta-Binomial, Gamma-Poisson, and Beta-Negative binomial, have been generated as a mixture of these classical discrete distributions, in which the parameters of the corresponding classical discrete distribution follow a beta or gamma distribution, respectively [5, p.253]. Ishii and Hayakawa [6] assumed that the binomial parameter is not a constant but a random variable with a beta distribution when comparing sex ratios in human families. Rai [7] considered accident proneness using Poisson distribution and assumed that the Poisson parameter was randomly distributed according to the power function distribution. Withers and Nadarajah [8] considered the number of times it rained in a given period to follow a Poisson distribution, and varied the Poisson parameter according to the gamma distribution. There are many new discrete distributions in the literature that provide more flexibility and a better fit to datasets in various fields of study than classical discrete distributions. Some of these new discrete distributions include the discrete trigeometric distributions generated by discretizing sine–Weibull, cos-Webull, tan-Weibull, and type II tan-Weibull distributions [9]. De Oliveira et al.[10] also obtained a discrete analog of the continuous power Lindley distribution using discretization. Karakaya [11] generated a new discrete distribution by compounding the Epanechnikov-exponential and Poisson distributions. Mushtaq et al. [12] introduced a new discrete generalized class of distributions.

In this study, a new discrete distribution is introduced by considering the geometric distribution parameter as a standard power function distribution random variable. This is to take cognizance of the inherent variability of each event. The remainder of this paper is structured as follows: Section 2 considers the derivation of the new distribution and its properties. Section 3 presents the estimation of the parameters of the new distribution. The posterior distribution is studied in Section 4, while the application of the real datasets to the new distribution is performed in Section 5. The concluding remarks are presented in Section 6.

II. The New model and its properties

Let $X_i\ i = 1,2,3 \ldots n$ represent a sample of n independent individuals and that whenever an event (such as an outbreak of a disease) occurs, each individual may live the event with a probability, p, or fail to live the event with probability, $q = 1 - p$. Then, the probability function of having an individual who lived the event after x observations is given by

$$P(X = x) = (1-p)^{x-1}p\,;\ \ x \in \mathbb{N}, 0 \leq p \leq 1 \qquad (1)$$

Equation (1) has a similar Markovian property of exponential distribution that characterizes it among all other distributions whose support range is restricted to nonnegative integers [5, p.210].

Suppose that the probability, p, at which each individual out lived the event varies and follows a standard power function distribution. According to Johnson et al. [13, p.210], a random variable, P, follows a standard power function distribution if its density function is given as

$$f(p) = \alpha p^{\alpha-1}; \quad 0 \leq p \leq 1, \alpha > 0 \tag{2}$$

The associated expectation of P denoted as $E(P)$ can be obtained as

$$E(P) = \int_0^1 p\, \alpha p^{x-1} dp$$
$$= \frac{\alpha}{\alpha+1} \tag{3}$$

The joint density function of the random variables X and P, denoted by $f(x,p)$ can be obtained using the Baye's theorem. Hence, $f(x,p) = f(x|p)f(p)$ is given by equation (4); where $f(p)$ is as defined by equation (2).

$$f(x,p) = \alpha p(1-p)^{x-1}p^{\alpha-1}; \ \alpha > 0, x \in \mathbb{N}, 0 \leq p \leq 1 \tag{4}$$

The marginal density function of random X, of equation (4) denoted by $f(x)$ is obtained as follows:

$$f(x) = \int_{\forall p} f(x,p)dp$$
$$= \int_0^1 p(1-p)^{x-1}\alpha p^{\alpha-1}\ dp$$
$$= \alpha \frac{\Gamma(\alpha+1)\Gamma(\alpha)}{\Gamma(x+\alpha+1)}$$
$$= \alpha\, \mathrm{B}(\alpha+1, x); \quad x \in \mathbb{N}, \alpha > 0 \tag{5}$$

where $\mathrm{B}(.,.)$ denotes the beta function. Equation (5) is a single-parameter distribution called the geometric-standard power function denoted by G-SPF. The corresponding cumulative density function $F(x) = P(X \leq x)$ in Equation (5) is given by:

$$F(x) = \Gamma(\alpha+1)\left[\frac{\alpha+1}{\Gamma(\alpha+2)} - \frac{(\alpha+x+1)\Gamma(x+1)}{\Gamma(\alpha+x+2)}\right]$$

A plot of the new density function for the selected parameter values is shown in Figure 1. The plot indicates that the frequency curve of the new distribution decreases monotonically. Type equation here.

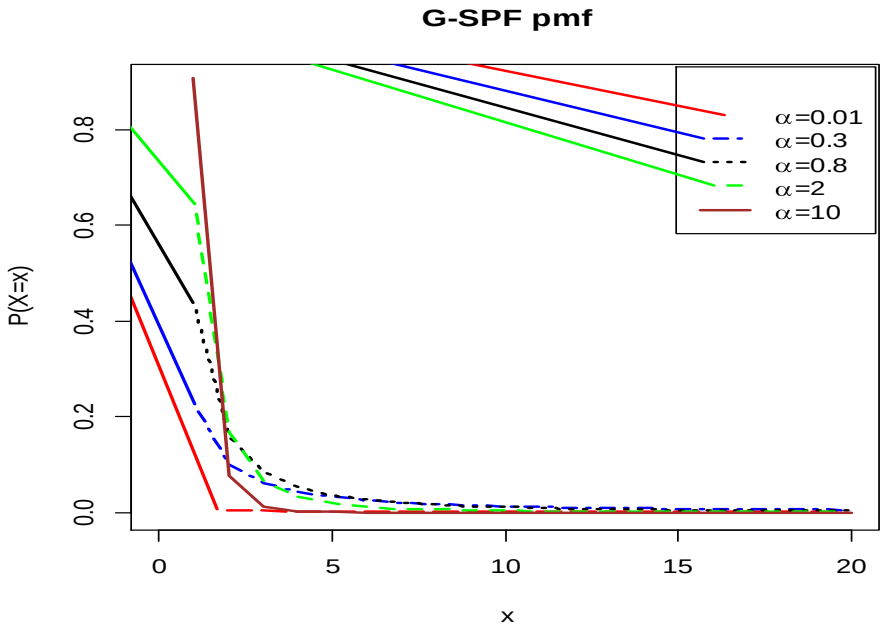


Figure 1: The plot of G-SPF distribution for some selected parameter values

I. Moments about the origin of the new distribution

The r^{th} moment about the origin of a random variable X, with support range $x \in \mathbb{N}$ and density function $f(x)$ is given by

$$E(X^r) = \sum_{\forall x} x^r f(x). \tag{6}$$

For $f(x)$ defined as in equation (5), the first and second moments about the origin are given by

$$E(X) = \frac{\alpha}{\alpha - 1} \tag{7}$$

and

$$E(X^2) = \frac{\alpha^2}{(\alpha - 1)(\alpha - 2)} \tag{8}$$

Proof

$$E(X) = \sum_{x=1}^{\infty} x\, f(x)$$

$$= \sum_{x=1}^{\infty} x \int_0^1 \alpha\, p^{\alpha}(1-p)^{x-1} dp$$

$$= \int_0^1 \alpha\, p^{\alpha} \left[\sum_{x=1}^{\infty} x\,(1-p)^{x-1} \right] dp \qquad (9)$$

Let $q = 1 - p$, then

$$E(X) = \int_0^1 \alpha\, p^{\alpha} \left[\sum_{x=1}^{\infty} x\, q^{x-1} \right] dp \qquad (10)$$

but

$$\sum_{x=1}^{\infty} x\, q^{x-1} = \sum_{x=1}^{\infty} \frac{dq^x}{dq}$$

$$= \frac{d}{dq} \sum_{x=1}^{\infty} q^x$$

$$= \frac{d}{dq} q(1 + q + q^2 + \cdots)$$

$$= \frac{d}{dq} \left(\frac{q}{1-q} \right)$$

$$= \frac{1}{(1-q)^2}$$

$$= \frac{1}{p^2} \qquad (11)$$

Substituting equation (11) in equation(10) results

$$E(X) = \int_0^1 \alpha\, p^{\alpha} p^{-2} dp$$

$$= \frac{\alpha}{\alpha - 1}. \qquad (12)$$

$$E(X^2) = \sum_{x=1}^{\infty} x^2 \int_0^1 \alpha\, p^{\alpha} (1-p)^{x-1} dp$$

$$= \sum_{x=1}^{\infty} [x(x-1)+x] \int_0^1 \alpha\, p^{\alpha}(1-p)^{x-1} dp$$

$$= \sum_{x=1}^{\infty} x(x-1) \int_0^1 \alpha\, p^{\alpha}(1-p)^{x-1} dp + \sum_{x=1}^{\infty} x \int_0^1 \alpha\, p^{\alpha}(1-p)^{x-1} dp$$

$$= \int_0^1 \alpha\, p^{\alpha} \left[\sum_{x=2}^{\infty} x(x-1) q q^{x-2} \right] dp + \frac{\alpha}{\alpha - 1}$$

$$= \int_0^1 \frac{2\alpha p^{\alpha}(1-p)}{p^3} dp + \frac{\alpha}{\alpha - 1}.$$

$$= 2\alpha \frac{\Gamma(\alpha-2)\Gamma(2)}{\Gamma(\alpha)} + \frac{\alpha}{\alpha - 1}$$

$$= \frac{2\alpha}{(\alpha-1)(\alpha-2)} + \frac{\alpha}{\alpha - 1}$$

$$= \frac{\alpha^2}{(\alpha-1)(\alpha-2)}.$$

Consequently, the variance, $V(X) = E(X^2) - (E(X))^2$, of the G-SPF is given by

$$V(X) = \frac{\alpha^2}{(\alpha-1)^2(\alpha-2)} \qquad (13)$$

II. Probability Generating Function

The probability generating function (pgf) of a discrete random variable X provides important information for a discrete random variable including its probability distribution and moment about the origin of the random variable. Suppose a random variable X follows G-SPF then, its pgf, denoted by $G_X(S)$ is given by

$$G_X(S) = E(S^X)$$

$$= \sum_{x=1}^{\infty} S^X f(x)$$

$$= \alpha S\ 2F1\left[1,1+\alpha;2+\alpha;\frac{s}{s-1}\right] \qquad (14)$$

where $2F1[\,.,\ .;\ .;.]$ is a Gaussian hypergeometric function.

III. Estimation of parameter of G-SPF

The maximum likelihood method was used to estimate the parameters of the G-SPF distribution. Suppose that a random sample of size n, denoted by $X_1, X_2, X_3, \ldots, X_n$ is drawn from a population with pdf defined in Equation (5). Then, the log-likelihood function $\ell(\alpha|X)$, is given by:

$$\ell(\alpha|X) = nlog(\alpha) + nlog[\Gamma(\alpha+1)] + \sum_{i=1}^{n}\log[\Gamma(x_i)] - \sum_{i=1}^{n}\log[\Gamma(x_i+\alpha+1)]. \qquad (15)$$

The score function corresponding to equation (15) is obtained by taken the derivative of equation (15) $w.r.t.\ \alpha$ which is given by

$$\frac{d\ell(\alpha|X)}{d\alpha} = \frac{n}{\alpha} + \frac{n\Gamma'(\alpha+1)}{\Gamma(\alpha+1)} + \sum_{i=1}^{n}\frac{\Gamma'(x_i+\alpha+1)}{\Gamma(x_i+\alpha+1)} \qquad (16)$$

The estimator of α, $\hat{\alpha}$ is obtained by equating equation (16) to zero and solving for α. However, equation (16) is a nonlinear equation, and the solution for $\hat{\alpha}$ does not exist in closed form. However, the estimate of α can be obtained using any of the different optimization methods, including Newton Raphson and BFGS, which are available in statistical software such as R and Python.

IV. Posterior distribution of p

Using the Bayes theorem, the posterior distribution of P is given by

$$f^*(p) = \frac{f(x|p)f(p)}{\int_0^1 f(x|p)f(p)\,dp}$$

$$= \frac{\Gamma(x+\alpha+1)}{\Gamma(\alpha+1)\Gamma(x)}p^{\alpha}(1-p)^{x-1}$$

$$= \frac{1}{\mathrm{B}(\alpha+1,x)}p^{\alpha}(1-p)^{x-1};\quad 0\le p\le 1, x\in\mathbb{N}, \alpha>0. \qquad (16)$$

Equation (17) is a beta distribution with $\alpha+1$ and x as its parameters, where $\mathrm{B}(.,.)$ is a beta function. The expectation of the posterior distribution of P denoted by $E(P^*)$ is obtained as

$$E(P^*) = \frac{\alpha+1}{x+\alpha+1} \qquad (18)$$

V. Applications

Two datasets were used to illustrate the potential of the new distribution, and how the new distribution adequately fits the datasets was evaluated by comparing its fits with the fits of some other discrete distributions using goodness-of-fit criteria such as the Aikaike Information Criterion (AIC) and negative

log-likelihood values. The discrete distributions considered were Negative Binomial, Poisson, Poisson-Geometric (GPois) because of [14]. The probability mass functions of these distributions are expressed as

$$P(x;n,p)=\frac{\Gamma(x+n)}{\Gamma(n)x!}p^n(1-p)^x; n>0, 0\leq p\leq 1, x=0,1,2,\ldots \quad (19)$$

$$P(x;\theta)=\frac{\theta^x e^{-\theta}}{x!};\ \theta>0, x=0,1,2,3,\ldots \quad (20)$$

$$P(x;\lambda,p)=\frac{p\lambda e^{-\lambda}}{x![1-(1-p)\Gamma(x+1,\lambda)]^2}; p,\lambda>0, x=0,1,2,3,\ldots \quad (21)$$

The first dataset was on oil pipeline failure caused by natural hazards and was reported by Nwezza *et al*. [14]. Dataset 1: 6, 10, 3, 2, 5, 8, 4, 11, 65, 5, 7, 36, 12, 7, 18, 5, 7, 13, 12, 5.

The second dataset is based on the number of daily deaths in Egypt due to Covid-19 infections as reported by Abdulhakim *et al*. [15]. Dataset 2: 1 ,1, 2, 2, 1, 1, 2, 4, 5, 1, 1, 3, 6, 6, 4, 1, 5, 6, 6, 8, 5, 7, 7, 9, 9, 15, 17, 11, 13, 5, 14, 5, 13, 9, 19, 15, 11, 14, 12, 11, 7, 13, 10, 20, 22, 21, 12.

Tables 1 and 2 present the parameter estimates of the competing distributions, AIC, and negative log-likelihood values of the datasets, respectively.

Table 1: Results of estimation and goodness-of-fit for the first dataset

Distribution	*Estimate*	*Standard error*	*AIC*	*-log-likelihood*
G-SPF($\boldsymbol{\alpha}$)	0.4415	0.1081	163.1054	80.5527
NegBinomial($\boldsymbol{p}$)	0.62402	0.0191	205.5163	101.7581
Poisson($\boldsymbol{\lambda}$)	12.0500	0.7789	292.5974	145.2987
GPois($\boldsymbol{p},\boldsymbol{\lambda}$)	0.0024	0.0010	253.3176	124.6588
	19.1243	0.9809		

Table 2: Results of estimation and goodness-of-fit for the second dataset

Distribution	*Estimate*	*Standard error*	*AIC*	*-log-likelihood*
G-SPF($\boldsymbol{\alpha}$)	0.5118	0.08321	337.0943	167.7572
NegBinomial($\boldsymbol{p}$)	0.8392	0.007	353.2962	175.6481
Poisson($\boldsymbol{\lambda}$)	8.3404	0.4237	374.0914	186.0457
GPois($\boldsymbol{p},\boldsymbol{\lambda}$)	0.7353	0.3816	375.735	185.8697
	8.6264	0.6545		

The AIC goodness-of-fit criterion is estimated using

$$AIC = -2\ell + 2s\,; \qquad (22)$$

where ℓ is the log-likelihood function and s is the number of parameters associated with the distribution.

The values in Tables 1 and 2 indicate that the AIC and negative log-likelihood values of the new distribution, G-SPF, are the smallest among the competing distributions. This implies that the new distribution provides the best fit for the two datasets.

The estimated density plots of the new distribution adequately connected the bins of the histogram of the datasets. The histogram and plots of the competing estimated density functions, as shown in Figures 2 and 3 for datasets 1 and 2, respectively, also indicate that the new distribution provided better fits.

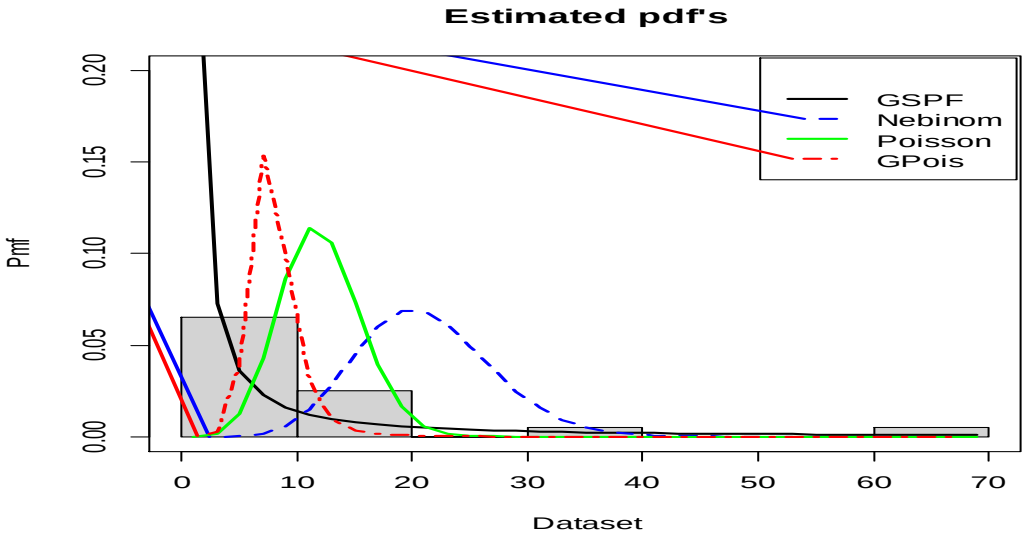


Figure 2: Plots of estimated pdfs for the first dataset

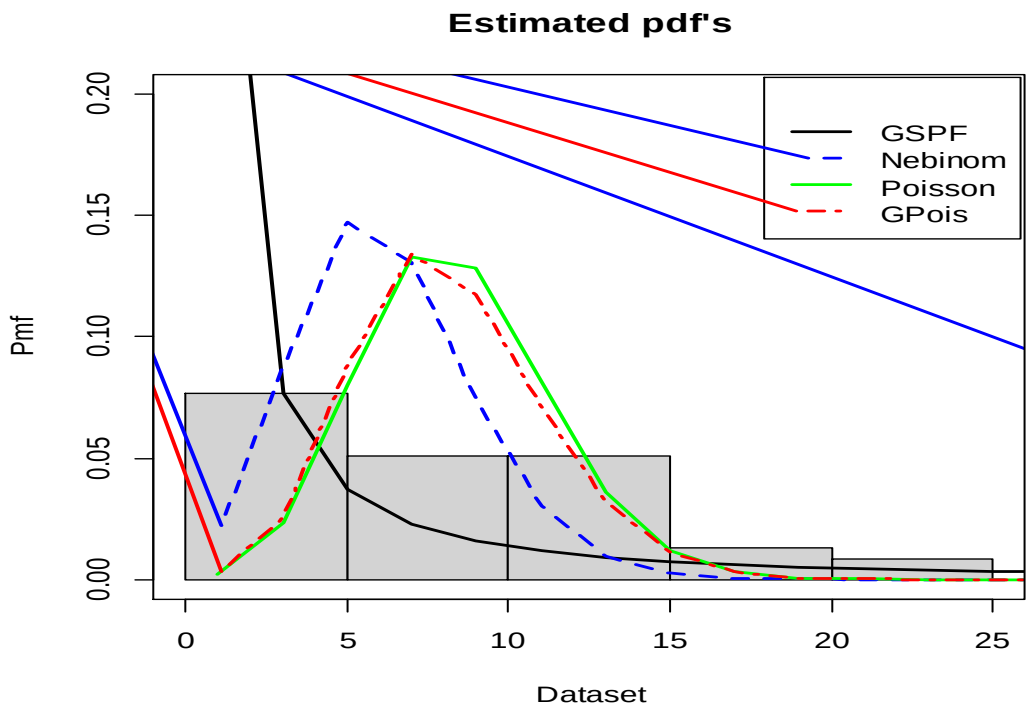


Figure 3: Plots of estimated pdfs for the second dataset

VI. Conclusion

A new discrete distribution for modeling discrete failure events was generated based on the variability of the geometric distribution parameter. The cumulative density function of the new distribution and some of its properties, including the moments and probability-generating function, were derived. The associated posterior distribution of the geometric distribution and its expectations were obtained. The expectation of the posterior distribution appeared to be larger than the expectation of the prior distribution of the geometric distribution. Furthermore, the new distribution was applied to two datasets to demonstrate its applicability to real-life problems, and a comparison to determine which distribution provided the best fit to the datasets was made with other existing discrete distributions that were applied to the same datasets using goodness-of-fit criteria. The results of the comparison show that the new distribution provides the best fit among the competing distributions.

References

[1] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd ed. New Jerssy(NJ). John Wiley; 2002

[2] Lawless JF. Statistical Models and Methods for lifetime data. 2nd ed. New Jersey (NJ).John Wiley; 2003.

[3] Meeker WQ, Escobar LA. Statistical methods for reliability Data. New Jerssy(NJ). John Wiley; 1998.

[4] Bakouch HS, Jazi MA, Nadarajah S. A new discrete distribution. Statistics; A J. theo. and appl. stat., doi/10.1080/02331888.2012.716677.

[5] Johnson Nl, Kemp AW, Kotz S. Univeriate discrete distributions, 3rd ed. New Jersey (NJ). John Wiley; 2005.

[6] Ishii G, Hayakawa R. On the compound binomial distribution. Annals inst. Stat. maths, 1960; 12: 69-80

[7] Rai G. A mathematical model for accident proneness. Trad. Estad. Invest. Oper. 1971; 22: 207-212.

[8]Withers CS, Nadarajah S. On the compound Poisson-Gamma distribution.
Kybernetika. 2011; 47:15-37.

[9] Chesneau C, Gillariose J, Joseph J. et al. (2024). New discrete trigeometric distribution: estimation with application to count data. Int. J. mod. and simul.. doi/10.1080/02286203.2024.2315328.

[10]De Oliveira RP, Mazucheli J, dos Santos MLA, et al. A discrete analogue of the continuous power Lindley distribution and its applications. Revista Brasileira de Biometria. 2018; 36(3): 649-667.

[11]Karakaya K. A new discrete distribution with application to radiation, smoking and health data. J. Radiation Research and Applied Sciences. 2023; 16(4)

[12]Mushtaq A, kayid M, Alomani G. A new discrete generalized class of distribution with application to radiation and covid-19 data. J. Radiation Research and Applied Sciences. 2025; 18(2).

[13]Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate distributions, 2nd
ed. New Jersey (NJ). John Wiley; 1995.

[14]Nwezza EE, Ugwuowo FI (2022). Modeling the number of component failure: A Poisson-geometric distribution. Scientific African. 2022;16: 1-8.

[15]Abdulhakim AA, Abdul Hadi NA., Ahmed ZA. A new discrete analog of the
continuous Lindley distribution with reliability applications. Entropy. 2020; 22(6): 603. Doi:
10.3390/e22060603

A MODIFIED GAUSSIAN KERNEL WEIGHTS FOR IMPROVING GOODNESS-OF-FIT OF LOCAL LINEAR REGRESSION

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's Email: edionwe.efosa@fupre.edu.ng

Abstract

The local linear regression model (LLR) is regarded as the nonparametric version of the weighted least squares in which the weight matrix (one for each data point) is derived from some decreasing functions such as the Gaussian kernel function (GKF) using appropriate size of bandwidth (smoothing parameter). It is observed that as the bandwidth tends to 1 and above, the kernel weights all tend approximately to the same (constant) value for which the LLR produces essentially the same goodness-of-fit results as those of the ordinary least squares (OLS) model. This observation implies that a measure of the flexibility of LLR over the OLS is a function of the degree of dispersion of the kernel weights. Therefore, a modified version of the GKF that guarantees kernel weights of higher measure of dispersion (flexibility) is proposed. Comparisons of the goodness-of-fit of four responses from a widely-analyzed response surface problem from the literature as well as those from simulated data show substantial improvements in the performance of the LLR that utilizes the kernel weights generated from the modified GKF.

Keywords: Simplified Gaussian kernel function, goodness-of-fit, kernel weights, local linear regression, ordinary least squares, response surface methodology, smoothing parameter

1. Introduction

The first phase of response surface methodology (RSM) is the experimental design phase where a researcher takes n measurements, y_i, $i = 1,2, \dots, n$, on the response (output) variable y resulting from n combinations (treatments) of values, x_{i1}, $x_{i2}, \dots, x_{iq}$, of the q explanatory (independent, input) variables $x_1, x_2, \dots, x_q$, and such measurements are used in the modelling phase of RSM for the purpose of establishing an empirical relationship between the two categories of variables in a process referred to as model fitting (Del Castillo, 2007; Ashish et al., 2024).

In the modelling phase, it is assumed that the relationship between the explanatory variables and the response variable can be represented by the formula:

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{iq}) + \varepsilon_i, \ \ i = 1,2, \dots, n, \quad (1)$$

where f is the true but unknown function of the relationship between y and the explanatory variables $x_1, x_2, \dots, x_q$, and ε_i are the random errors such that $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$(Khuri, 2017; Ugur et al., 2024).

For data emanating from response surface studies, there are basically two categories of model for estimating f in Eq. (1), namely the parametric regression model (e.g. Ordinary Least Squares, OLS) and nonparametric regression model (e.g. Local Linear Regression, LLR).

The challenges encountered in the application of the OLS in RSM include high bias and inaccurate predictions if the assumed model does not accurately represent the data under study as well as when the data set contains complex or hidden patterns (Pickle et al. 2008; He et al. 2012).

According to Myers (1999), Anderson-Cook and Prewitt (2005), Pickle et al. (2008) and Wan and Birch (2011), LLR is applied if the objectives of the research include the following:

1. The functional form of the relationship between the explanatory variables and the response variable is unknown or known but not well behaved.
2. The researcher's interest is a curve that can be used for predicting the values of responses within the design space;
3. The researcher is less interested in the interpretive function of the estimated model coefficients (parameters) and more interested in studying the shape of the response surface.

LLR estimate of y_i, denoted by $\hat{y}_i^{(LLR)}$, $i = 1,2, \dots, n$, is given by:

$$\hat{y}_i^{(LLR)} = \boldsymbol{x_i}(\boldsymbol{X^T W_i X})^{-1}\boldsymbol{X^T W_i y}, \quad (2)$$

In matrix notation, the n by 1 vector of estimates of the response from Eq. (2) is expressed as:

$$\hat{\boldsymbol{y}}^{(LLR)} = \begin{bmatrix} \boldsymbol{h_1^{(LLR)}} \\ \boldsymbol{h_2^{(LLR)}} \\ \vdots \\ \boldsymbol{h_n^{(LLR)}} \end{bmatrix} \boldsymbol{y}, \quad (3)$$

$$= \boldsymbol{H^{(LLR)}y}, \quad (4)$$

where $\boldsymbol{H}^{(LLR)}$ is the $n \times n$ LLR Hat matrix, and $\boldsymbol{h_i^{(LLR)}} = \boldsymbol{x_i}(\boldsymbol{X^T W_i X})^{-1}\boldsymbol{X^T W_i}$ is the i^{th} row vector of $\boldsymbol{H}^{(LLR)}$ for estimating the response in the i^{th} data point, $\boldsymbol{x_i}$ is the i^{th} row of the LLR $n \times (q+1)$ model matrix $\boldsymbol{X}$ (with transpose $\boldsymbol{X^T}$) given by:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix},$$

$\boldsymbol{W_i}$ is an $n \times n$ diagonal matrix of the kernel weights for estimating y_i given by:

$$\boldsymbol{W_i} = \begin{bmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & w_{nn} \end{bmatrix}, \quad (5)$$

where w_{ii} of the kernel weight matrix, $\boldsymbol{W_i}$ for estimating y_i in Eq. (2) is obtained from the product kernel of the simplified GKF given by:

$$w_{ii} = \prod_{j=1}^{q} K\left(\frac{x_{oj}-x_{ij}}{b}\right) \Big/ \sum_{i=1}^{n}\prod_{j=1}^{q} K\left(\frac{x_{oj}-x_{ij}}{b}\right), \quad (6)$$

where $K\left(\frac{x_{oj}-x_{ij}}{b}\right) = e^{-\left(\frac{x_{oj}-x_{ij}}{b}\right)^2}$ is the simplified Gaussian kernel function (GKF), $x_{oj}, j = 1,2, \dots, q$, is the value of the j^{th} explanatory variable at the i^{th} data point, $i = 1,2, \dots, n$, where the estimation of y_i, $i = 1,2, \dots, n$, is sought and b, such that $0 < b \leq 1$, is the bandwidth (smoothing parameter) (Zheng and Kulasekera, 2013; Edionwe and Mbegbu, 2014; Eguasa and Edionwe[a], 2025).

The optimal bandwidth b^{opt} is the bandwidth that minimizes the Penalized Prediction Error Sum of Squares ($PRESS^{**}(b)$) given by:

$$PRESS^{**}(b) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(LLR)}\right)^2}{\left(n - trace\left(H^{(LLR)}(b)\right)\right) + (n-k-1)\left(\frac{SSE_{max} - SSE_b}{SSE_{max}}\right)}, \quad (7)$$

$$= \frac{PRESS}{\left(n - trace\left(H^{(LLR)}(b)\right)\right) + (n-k-1)\left(\frac{SSE_{max} - SSE_b}{SSE_{max}}\right)}, \quad (8)$$

where SSE_{max} is the maximum Sum of Squared Errors obtained as b tends to infinity, SSE_b is the Sum of Squared Errors associated with a particular value of b, $tr\left(\boldsymbol{H}^{(\boldsymbol{LLR})}b\right)$ is the trace of the LLR Hat matrix, and $\hat{y}_{i,-i}^{(LLR)}$ is the leave-one-out cross-validation estimate of y_i with the i^{th} observation left out (Mays and Birch, 2002; Edionwe and Eguasa, 2023; Eguasa and Edionwe[b], 2025).

Nonparametric regression models such as LLR are developed for modelling data in studies that involve a single explanatory variable and a very large sample size n (Cleveland, 1979; Fan and Gijbels, 1995; Hardle et al., 2005). Therefore, the application of LLR is still very unpopular in RSM where sample size could be as small as twenty with as many as three explanatory variables (Anderson-cook and Prewitt, 2005). In fact, a comparison of the goodness-of-fit clearly shows that LLR does not perform even as well as the OLS as far as data from RSM is concerned especially when the data consists of more than one explanatory variables (Pickle et al., 2008; Wan and Birch, 2011; Eguasa et al., 2022; Edionwe et al., 2016).

Hence, the current structure of LLR requires an upgrade for a better adaptation to data emanating from RSM. In particular, this paper seeks to address the low performance of LLR from the perspective of the kernel weights computed from the simplified GKF.

The remainder of the paper is organized as follows: Section 2 presents a review of kernel weights derived from the current form of the simplified GKF. A modified version of the simplified GKF is proposed in Section 3. Comparisons of the goodness-of-fit of LLR utilizing the proposed GKFs and those from the OLS and LLR utilizing the existing form of the GKF are presented in Section 4. The paper concludes in Section 5.

2. A Review of the Weights Matrix derived from the GKF

The application of the Gaussian kernel function (GKF) as a weighting function with respect to nonparametric regression such as LLR is based on three assumptions. One, that data points closer to a given point (say x_o), defined as data point of interest in LLR, contain more information about the value of the response variable at the data point x_o and therefore should be assigned larger kernel weights than data points that are farther away from x_o. Two, that the unknown function, $f\left(x_{i1}, x_{i2}, \dots, x_{iq}\right)$ in Eq. (1), is smooth, meaning that small changes in the explanatory variables should lead to small changes in the response. Three, that the unknown function, $f\left(x_{i1}, x_{i2}, \dots, x_{iq}\right)$ in Eq. (1), is continuous, meaning that y_i and $x_{i1}, x_{i2}, \dots, x_{in}$ are measurable quantity, taking on infinitely any value within a given range and not limited to specific, distinct values (Hardle, 2005; Hofmann et al., 2008).

Assuming that we require the LLR estimate of $y_o, i = o, x_{oj}$ being the value of the j^{th} explanatory variable at the data point $i = o$. Consider two other data points $i = s$ and $i = t$, where the value of the j^{th} explanatory variable is assumed to be x_{sj} and x_{tj}, respectively. Assuming $\left|(x_{oj} - x_{sj})\right| < \left|(x_{oj} - x_{tj})\right|$, that is x_{sj} is closer or nearer to x_{oj} than x_{tj}. So, the GKF in Eq. (6), being a decreasing function, assigns a relatively larger weight (say w_{ss}) to the data point $i = s$ that is closer to x_{oj} than the weight (say w_{tt}) that it assigns to data point $i = t$ that is farther away from x_{oj} (Fan and Gijbels, 1995; Hardle et al., 2005). That is $w_{ss} > w_{tt}$ for $(x_o - x_s) < (x_o - x_t)$ where, respect to y_o, the data point $i = o$ is called the data point of interest in the LLR literature (Mays and Birch, 2002; Pickle et al., 2008).

Since, in a given data, we have data points that are relatively nearer to as well as data points that are relatively farther from a given data point of interest, we get a weight matrix, say $\boldsymbol{W_o}$, consisting of both relatively large and relatively small kernel weights.

Essentially, the diagonal weights matrix derived from Eq. (6) can be expressed as:

$$\boldsymbol{W_i} = \begin{bmatrix} \left(\frac{\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{1j}}{b}\right)^2}}{\sum_{i=1}^{n}\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{1j}}{b}\right)^2}}\right) & 0 & \cdots & 0 \\ 0 & \left(\frac{\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{2j}}{b}\right)^2}}{\sum_{i=1}^{n}\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{2j}}{b}\right)^2}}\right) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{nj}}{b}\right)^2}}{\sum_{i=1}^{n}\prod_{j=1}^{q} e^{-\left(\frac{x_{ij}-x_{nj}}{b}\right)^2}}\right) \end{bmatrix}$$

A study of the behavior of the weight matrix reveals that as the kernel weights take on the same constant value, say α, the value of LLR estimate of y_i, $i = 1,2,\ldots,n$, tend to the value of OLS estimate of y_i, $i = 1,2,\ldots,n$. In other words, as $w_{11} = w_{22} = \cdots = w_{nn} = \alpha$ for all W_i, we have:

$$\hat{y}_i^{(LLR)} = \boldsymbol{x_i}(\boldsymbol{X^T W_i X})^{-1}\boldsymbol{X^T W_i y} = \boldsymbol{x_i}(\boldsymbol{X^T X})_{-1}\boldsymbol{X^T y} = \hat{y}_i^{(OLS)}, \quad (9)$$

which gives rise to a situation where $\hat{y}_1^{(LLR)} = \hat{y}_1^{(OLS)}, \hat{y}_2^{(LLR)} = \hat{y}_2^{(OLS)}, \ldots, \hat{y}_n^{(LLR)} = \hat{y}_n^{(OLS)}$.

Although, this situation in Eq. (9) occurs at a data-based value of b usually close to and greater than 1, but what it implies is that the flexibility of LLR derives from the variability of the kernel weights in $\boldsymbol{W_i}$, $i = 1,2,\ldots,n$. This is so for the simple reason that if we compute a measure of dispersion (such as variance) of the weight matrix arising from the scenario in Eq. (9), we expect to get a zero variance of the weights, and by extension, a zero or minimum flexibility of LLR.

Therefore, the aim of this paper is to modify the GKF such that the generated weight matrices are of relatively larger variance than the variance of the current form of the GKF in Eq. (6) so as to address the poor performance of the LLR.

3. Methodology

In the modification of the GKF, we take advantage of the idea that the flexibility of LLR is a measure of the variability or dispersion of the kernel weights.

Recall that the kernel weight is given by:

$$w_{ii} = \prod_{j=1}^{q} K\left(\frac{x_{oj}-x_{ij}}{b}\right) \Big/ \sum_{i=1}^{n}\prod_{j=1}^{q}\left(\frac{x_{oj}-x_{ij}}{b}\right), \quad i = 1,\ldots,n$$

where $i = o$ is the data point of interest where the estimate of y_o is sought, and the value of the explanatory variable at $i = o$ is $x_{oj}, j = 1,2,\ldots,q$.

We know that the sum of the product kernel weights $\sum_{i=1}^{n}\prod_{j=1}^{q}\left(\frac{x_{oj}-x_{ij}}{b}\right)$ is a function of the n product kernel weights $\prod_{j=1}^{q} K\left(\frac{x_{oj}-x_{ij}}{b}\right)$. Therefore, for ease of computation, we express the kernel weight as:

$$w_{ii} \propto K\left(\frac{x_{oj}-x_{ij}}{b}\right) = \prod_{j=1}^{q} e^{-\left(\frac{x_{oj}-x_{ij}}{b}\right)^2}, \quad i = 1,2,\ldots,n, \quad (10)$$

Next, we express w_{ii} in Eq. (10) in a more general form as:

$$w_{ii}^{*} = \prod_{j=1}^{q} e^{-\left(\frac{a_1 x_{oj} - a_2 x_{ij}}{b}\right)^2},\ a_1 \geq 0, a_2 \geq 0, \qquad i = 1,2,\dots,n,$$

(11)

The weights w_{ii}^{*} in Eq. (11) reduces to w_{ii} in Eq. (10) if $a_1 = a_2 = 1$.

From Eq. (11), we have

$$w_{ii}^{*} = e^{-\left(\frac{a_1 x_{o1} - a_2 x_{i1}}{b}\right)^2} \times e^{-\left(\frac{a_1 x_{o2} - a_2 x_{i2}}{b}\right)^2} \times \dots \times e^{-\left(\frac{a_1 x_{oq} - a_2 x_{iq}}{b}\right)^2},$$

(12)

$$w_{ii}^{*} = e^{-\left[\left(\frac{a x_{o1} - a_2 x_{i1}}{b}\right)^2 + \left(\frac{a_1 x_{o2} - a_2 x_{i2}}{b}\right)^2 + \cdots + \left(\frac{a_1 x_{oq} - a_2 x_{iq}}{b}\right)^2\right]},$$

(13)

$$\ln w_{ii}^{*} = \ln\left\{e^{-\frac{1}{b^2}\left[(a_1 x_{o1} - a_2 x_{i1})^2 + (a_1 x_{o2} - a_2 x_{i2})^2 + \cdots + \left(a_1 x_{oq} - a_2 x_{iq}\right)^2\right]}\right\},$$

$$\ln w_{ii}^{*} = -\frac{1}{b^2}\left[(a_1 x_{o1} - a_2 x_{i1})^2 + (a_1 x_{o2} - a_1 x_{i2})^2 + \cdots + \left(a_1 x_{oq} - a_2 x_{iq}\right)^2\right],$$

(14)

(15)

On expansion, Eq. (15) gives:

$$lnw_{ii}^{*} = -\frac{1}{b^2}\left[\left(a_1^2 x_{o1}^2 - 2a_1 a_2 x_{o1} x_{i1} + a_2^2 x_{i1}^2\right) + \left(a_1^2 x_{o2}^2 - 2a_1 a_2 x_{o2} x_{i2} + a_2^2 x_{i2}^2\right) + \cdots + \left(a_1^2 x_{oq}^2 - 2a_1 a_2 x_{oq} x_{iq} + a_2^2 x_{iq}^2\right)\right],$$

(16)

$$lnw_{ii}^{*} = \frac{1}{b^2}\left[(2a_1 a_2 x_{o1} x_{i1} - a_2^2 x_{i1}^2 - a_1^2 x_{o1}^2) + (2a_1 a_2 x_{o2} x_{i2} - a_2^2 x_{i2}^2 - a_1^2 x_{o2}^2) + \cdots + \left(2a_1 a_2 x_{oq} x_{iq} - a_2^2 x_{iq}^2 - a_1^2 x_{oq}^2\right)\right],$$

(17)

Collecting like terms in Eq. (17) gives:

$$lnw_{ii}^{*} = \frac{1}{b^2}\left[2a_1 a_2\left(x_{o1} x_{i1} + x_{o2} x_{i2} + \cdots + x_{oq} x_{iq}\right) - a_2^2\left(x_{i1}^2 + x_{i2}^2 + \cdots + x_{iq}^2\right) - a_1^2\left(x_{o1}^2 + x_{o2}^2 + \cdots + x_{oq}^2\right)\right],$$

(18)

$$\ln w_{ii}^{*} = \frac{1}{b^2}\left[2a_1 a_2 \sum_{j=1}^{q} x_{oj} x_{ij} - a_2^2 \sum_{j=1}^{q} x_{ij}^2 - a_1^2 \sum_{j=1}^{q} x_{oj}^2\right],$$

(19)

From Eq. (19), we get a vector of weights, $[\ln w_{11}^{*}, \ln w_{22}^{*}, \dots, \ln w_{nn}^{*}]$ which forms the diagonal elements of the $n \times n$ weight matrix $\boldsymbol{W}_i$, $i = 1,2,\dots,n$.

Since $\sum_{j=1}^{q} x_{oj}^2 = x_{o1}^2 + x_{o2}^2 + \cdots + x_{oq}^2 = C\ (say)$, we can express Eq. (19) as:

$$\ln w_{ii}^{*} = \frac{1}{b^2}\left[2a_1 a_2 \sum_{j=1}^{q} x_{oj} x_{ij} - a_2^2 \sum_{j=1}^{q} x_{ij}^2 - a_1^2 C\right],$$

(20)

The variance of both sides of Eq. (20) gives:

$$Var([\ln w_{11}^{*}, \ln w_{22}^{*}, \dots, \ln w_{nn}^{*}]) = Var\left[\frac{1}{b^2}\left(2a_1 a_2 \sum_{j=1}^{q} x_{oj} x_{ij} - a_2^2 \sum_{j=1}^{q} x_{ij}^2 - a_1^2 C\right)\right],$$

(21)

From $Var(a\boldsymbol{x}) = a^2 Var(\boldsymbol{x})$, where a is a constant, we have:

$$Var([\ln w_{11}^{*}, \ln w_{22}^{*}, \dots, \ln w_{nn}^{*}]) = \frac{1}{b^4}\left[4a_1^2 a_2^2 \sum_{j=1}^{q} Var(x_{oj} x_{ij}) - a_2^4 \sum_{j=1}^{q} Var\left(x_{ij}^2\right) - a_1^4 Var(C)\right],$$

(22)

For data with a single explanatory variable, $q = 1$, Eq. (22) reduces to:

$$Var([\ln w_{11}^{*}, \ln w_{22}^{*}, \dots, \ln w_{nn}^{*}]) = \frac{a_2^2}{b^4}\left[4a_1^2 Var(x_o x_i) - a_2^2 Var(x_i^2)\right],$$

(23)

$$Var([\ln w_{11}^*, \ln w_{22}^*, \dots, \ln w_{nn}^*]) = \frac{a_2^2}{b^4}\left[4a_1^2x_o^2Var(x_i) - a_2^2Var(x_i^2)\right],\ i = 1,2,\dots,n,$$

(24)

For the existing version of the GKF in Eq. (10), if we substitute $a_1 = a_2 = 1$ in Eq. (24), we obtain the equivalent of the estimates of the variance of the natural logarithm of kernel weights:

$$Var([\ln w_{11}^*, \ln w_{22}^*, \dots, \ln w_{nn}^*]) = \left|\frac{1}{b^4}\left[4x_o^2Var(x_i) - Var(x_i^2)\right]\right|, \qquad i = 1,2,\dots,n,$$

(25)

Notice that the size of the variance is just the difference between the two terms in the brackets in Eq. (25). Therefore, based on the aim of the current paper, the following steps ensure that the variance in Eq. (24) is as large as the data permits:

Step 1 - set $a_2 = 1$ in Eq. (24) to get:

$$Var([\ln w_{11}^*, \dots, \ln w_{nn}^*]) = \frac{1}{b^4}\left[4a_1^2x_o^2Var(x_i) - Var(x_i^2)\right],$$

(26)

Step 2 –using suitable optimization tool, determine the optimal values of $0 < b \leq 1$ and $a_1 > 0$ in order to increase the variance in Eq. (26) (equivalently in Eq. (5) or Eq. (11)).

Therefore, the modified kernel weights derived from the proposed modified GKF are given by:

$$w_{ii}^* = \prod_{j=1}^{q} K\left(\frac{a_1x_{oj} - x_{ij}}{b}\right) \Big/ \sum_{i=1}^{n}\prod_{j=1}^{q}\left(\frac{a_1x_{oj} - x_{ij}}{b}\right),$$

(27)

Although, Eq. (26) was derived for $q = 1$ (that is for a single response) but is also true for $q > 1$ since the kernel weights are independently generated for each of the q explanatory variables to get the product kernel weights in Eq. (27).

With reference to PRESS**(b) in Eq. (7), using suitable optimization tool, the optimal values of the parameters a_1 and b in Eq. (27) are determined by the minimization of $PRESS^{**}(a_1, b)$ given by:

$$PRESS^{**}(a_1, b) = \frac{PRESS}{\left(n - trace\left(H^{(LLR)}(a_1, b)\right)\right) + (n-k-1)\left(\frac{SSE_{max} - SSE_{a_1,b}}{SSE_{max}}\right)},$$

(28)

where every term retains the previous definition.

The application of modified GKF in Eq. (27) is based on the same assumptions stated for the existing GKF in Eq. (6) in the opening paragraph of Section 2. The variance of the kernel weights, denoted by $\sigma^2(w_{ii})$, $i = 1,2,\dots,n$, generated from both the existing GKF and the modified GKF lie in the interval $0 \leq w_{ii} \leq 1$ since the kernel weights lie in the interval $0 \leq w_{ii} \leq 1$.

In this paper, the minimization of PRESS** for the selection of the optimal parameters b in Eq. (8), $a_1$and b in Eq. (28) were carried out using the Genetic Algorithm (GA) optimization toolbox embedded in Matlab. GA is specifically suitable when dealing with optimization of regression models (such as LLR) that cannot be expressed in closed form (Alvarez et al., 2009; Yeniay 2014; Sestelo et al., 2017).

A computer program written in codes executable in Matlab for implementing LLR is presented in the Appendix. The computer program is coupled to the GA tool in Matlab 2014 version using the @ link.

4. Application

In the first part of this Section, a real multiple response data from the literature is used in order to compare the goodness-of-fit of the OLS, the LLR that utilizes the kernel weights derived from the proposed version of GKF (denoted by LLR_1 for ease of reference) with those of LLR that utilizes kernel weights from the current form of GKF. The goodness-of-fit used for comparison include Sum of Squares of Errors (SSE), Coefficient of Determination (R^2), and the $PRESS^{**}$ criteria given in Eq. (8) and Eq. (28), respectively for LLR and LLR_1. For the OLS, PRESS** is obtained using the formula

$PRESS^{**} = \frac{PRESS}{\left(n-trace\left(H^{(OLS)}\right)\right)}$, where $n - trace\left(H^{(OLS)}\right)$ is the degree of freedom (Wan and Birch, 2011).

The SSE and R^2 both give an indication of how close the estimates of the response are to the observed values while PRESS** criterion is a measure of the predictive capability of the fitted models (Montgomery. 2009; Zahra et al., 2020). The best result for each of the goodness-of-fit is shown in bold font.

A table showing the variances of the weight matrix at each data point, $(Var(\boldsymbol{W_i}), i = 1,2,\dots,n,)$ from LLR and LLR_1 as well as the plots of the residuals of the estimated response from each of the regression models are also presented. For instance, the residual plots of LLR_1 for each of the four response variables are obtained by plotting LLR residuals, $e_i^{(LLR_1)} = y_i - y_i^{(LLR_1)}, i = 1,2,\dots,n,$ against the data points i, respectively.

This Section concludes with further comparisons of results based on a simulation study involving data that comprises a single and three explanatory variables.

4.1 The Minced Fish Quality Data

The Minced Fish Quality Data originated from the work of Shah et al. (2004). The LLR results are presented in several papers including Wan and Birch (2011). The study involves three explanatory variables x_1 (washing temperature), x_2(washing time) and x_3 (washing ratio of water volume to sample kernel weights) and four responses each measuring four aspects of quality of minced fish, namely, springiness (y_1), thiobarbituric acid number (y_2), cooking loss (y_3), and whiteness index (y_4).

The polynomials specified for OLS for both y_1 and y_4 include the intercept, x_1 and x_1^2. The one specified for y_2 includes the intercept, x_1, x_2, x_1^2, and x_1x_2, and for y_3, we have the intercept, $x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3, x_3^2$. The data generated from a CCD is presented in Table 1.

The optimal smoothing parameters of LLR and LLR_1 for each response are presented in Table 2. Table 3 shows the variances of the weight matrices at each data point while the comparison of the results for goodness-of-fit are presented in Table 4.

Table 1: The Minced Fish 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	1.52	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

Table 2: Optimal values of the bandwidths for the minced fish data

Response	*Model*	a_{opt}	b_{opt}
y_1	LLR	-	0.1463
	LLR_1	1.0091	0.0815
y_2	LLR	-	0.4363
	LLR_1	1.6875	0.1701
y_3	LLR	-	0.5371
	LLR_1	1.6062	0.1502
y_4	LLR	-	0.1197
	LLR_1	1.0498	0.0879

Table 3: Variances of the Weight Matrices for the minced fish quality data

I	y_1		y_2		y_3		y_4	
	LLR	LLR_1	LLR	LLR_1	LLR	LLR_1	LLR	LLR_1
1	0.0102	0.0119	0.0028	0.0053	0.0025	0.0292	0.0114	0.0119
2	0.0102	0.0119	0.0028	0.0583	0.0025	0.0583	0.0114	0.0117
3	0.0102	0.0119	0.0028	0.0583	0.0025	0.0583	0.0114	0.0119
4	0.0102	0.0119	0.0028	0.0276	0.0025	0.0588	0.0114	0.0117
5	0.0102	0.0119	0.0028	0.0053	0.0025	0.0583	0.0114	0.0119
6	0.0102	0.0119	0.0028	0.0583	0.0025	0.0588	0.0114	0.0117
7	0.0102	0.0119	0.0028	0.0583	0.0025	0.0588	0.0114	0.0119
8	0.0102	0.0119	0.0028	0.0276	0.0025	0.0588	0.0114	0.0117
9	0.0234	0.0578	0.0034	0.0254	0.0024	0.0586	0.0385	0.0565
10	0.0234	0.0580	0.0034	0.0586	0.0024	0.0576	0.0385	0.0586
11	0.0049	0.0053	0.0034	0.0254	0.0024	0.0586	0.0052	0.0053
12	0.0049	0.0053	0.0034	0.0586	0.0024	0.0576	0.0052	0.0053
13	0.0049	0.0053	0.0011	0.0270	0.0024	0.0586	0.0052	0.0053
14	0.0049	0.0053	0.0011	0.0270	0.0024	0.0576	0.0052	0.0053
15	0.0049	0.0053	0.0011	0.0270	0.0007	0.0588	0.0052	0.0053
16	0.0049	0.0053	0.0011	0.0270	0.0007	0.0588	0.0052	0.0053
17	0.0049	0.0053	0.0011	0.0270	0.0007	0.0588	0.0052	0.0053
sum	0.1627	0.2481	0.0415	0.6020	0.0365	0.9643	0.2048	0.2466

Table 4: Goodness-of-fit for the minced fish data

Response	*Model*	*PRESS***	R^2	*SSE*
y_1	OLS	0.0042	92.1256	0.0231
	LLR	**0.0026**	95.6990	0.0126
	LLR_1	**0.0026**	**95.7916**	**0.0123**
y_2	OLS	19.6113	93.3851	90.9033
	LLR	36.4407	82.1456	245.3568
	LLR_1	**13.8133**	**96.5101**	**47.9581**
y_3	OLS	20.3074	84.0607	41.1338
	LLR	17.0573	68.1622	82.1622
	LLR_1	**13.2095**	**92.2568**	**19.9824**
y_4	OLS	48.9401	54.1259	198.8048
	LLR	17.1477	97.1704	12.2627
	LLR_1	**17.0615**	**97.1990**	**12.1387**

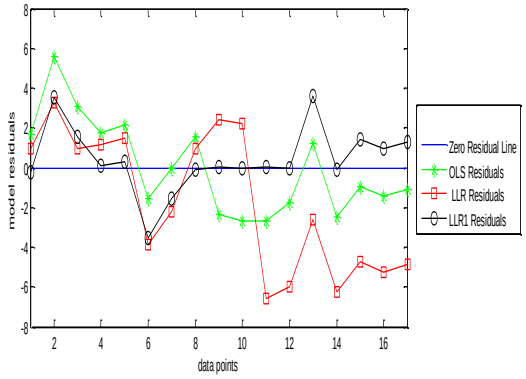


Figure 1: Graph of model residuals for y_2

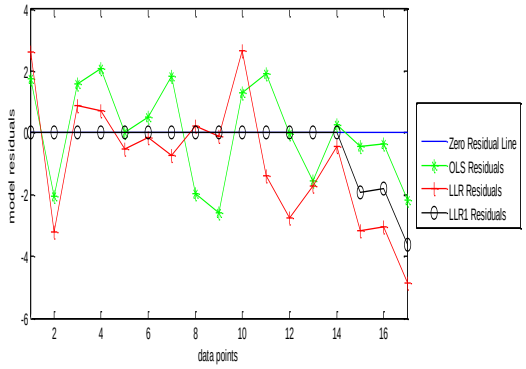


Figure 2: Graph of model residuals for y_3

4.2 Discussion of Results

Table 3 presents the sum of the variances of the kernel weights across the four responses. As expected, LLR_1 gives the kernel weights with larger sum of variances than those of LLR. In this paper, the aim of the modification of the existing GKF is to obtain a GKF that generates kernel weights of larger variances. Therefore, the results in Table 3 serve as a numerical validation to that effect.

Table 4 shows that LLR_1 gives the best results in all the goodness-of-fit and across the four responses except for PRESS** of y_1 where it jointly gives the best value of 0.0026 with LLR. Of the three regression models considered, best values of PRESS** show that LLR_1 possesses the highest predictive capability which is the essence of data modelling in RSM. Furthermore, based on the results obtained for R^2, we observe that LLR_1 is able to explain approximately 96%, 97%, 92% and 97% of the inherent variations in y_1, y_2, y_3 and y_4, respectively. Again, comparison of the SSE indicates that the LLR_1 estimates of the response are the closest to the observed values of the response in the study.

Notice that if LLR_1 is removed from the picture, OLS will emerge the better regression model for y_2 and y_3, both of which consist of more than one explanatory variables. This is the reason the LLR is not a popular regression model for RSM whose data usually consists of two or more explanatory variables.

The model residuals of estimates of y_2 and y_3 from the three regression models are graphed against the seventeen data points and the plots are shown in Figures 1 and 2, respectively. Overall, the LLR_1 residuals from the two Figures are seen to lie closest to the zero residual line with its residuals lying virtually on the zero residual line in nine and fourteen data points in Figure 1 and Figure 2, respectively. This reflects in the smaller SSE of the LLR_1 particularly in the y_1 and $y_2$responses. The

residual plots of LLR for both y_1 and y_4 are quite close to those of LLR_1 for the simple reason that both y_1 and y_4 involve a single explanatory variable. This closeness in residuals also reflects in their SSE for y_1 and y_4.

4.3 Simulation Study

The aim of the simulation study is to show how each of the regression models performs when random errors of varying variances are added to data simulated from polynomials given by:

$$y_i = 22 + 20x_{i1} - 18x_{i1}^2 + \varepsilon_i, \quad (29)$$

$$y_i = 48 - 67x_{i1} - 15x_{i2} - 38x_{i3} + 44x_{i1}^2 + 13x_{i2}^2 + 21x_{i3}^2 + 15x_{i1}x_{i2} + 35x_{i1}x_{i3} - 8x_{i2}x_{i3} + \boldsymbol{\varepsilon_i} \quad (30)$$

where x_{i1}, x_{i2}, and x_{i3}, $i = 1, \ldots, n$, are the respective values of the of the explanatory variables x_1, x_2, and x_3 at the i^{th} data point in Table 1, ε_i, $i = 1, \ldots, n$, is the i^{th} value of the normally distributed random error, where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$.

In the simulation study, random errors of four different variances are used and 500 data sets were simulated for each random error variance. The respective average SSE (AVESSE) and the average PRESS** (AVEPRESS**) of each regression model and for each size of the random error variance are presented in Table 5 (for polynomial in Eq. (29)) and Table 6 (for polynomial in (30)). Also included in the tables is the average sum of the variances of the kernel weights (denoted by AVESUMKW) for LLR and LLR_1 to provide further proof that the kernel weights generated by the modified GKF are of larger variances than their counterparts from the existing GKF.

Table 5: Results from Simulated Data

$\sigma^2(\boldsymbol{\varepsilon})$	*Model*	*AVEPRESS***	*AVESSE*	*AVESUMKW*
1	OLS	15.2037	69.6546	-
	LLR	10.3397	62.2154	0.0007
	LLR_1	**8.0511**	**33.8001**	0.0104
5	OLS	23.3487	141.0549	-
	LLR	15.8968	118.4308	0.0006
	LLR_1	**14.1418**	**78.4247**	0.0094
10	OLS	32.6949	228.4214	-
	LLR	22.5295	**188.7178**	0.0005
	LLR_1	**21.4985**	196.9886	0.0023
15	OLS	41.7502	315.1319	-
	LLR	29.0449	**258.8174**	0.0005

	LLR_1	**28.1290**	278.7111	0.0013

Table 6: Results of simulated data from Eq. (30)

$\sigma^2(\boldsymbol{\varepsilon})$	*Model*	*AVEPRESS***	*AVESSE*	*AVESUMKW*
1	OLS	50.0898	47.0046	-
	LLR	22.3111	104.8015	0.0007
	LLR_1	**16.0485**	**11.1099**	0.0077
5	OLS	82.2054	85.2200	-
	LLR	32.6564	154.4627	0.0007
	LLR_1	**27.2541**	**33.4807**	0.0066
10	OLS	115.4911	127.4185	-
	LLR	43.2442	208.0761	0.0007
	LLR_1	**39.1434**	**64.6071**	0.0057
15	OLS	146.3884	167.6773	-
	LLR	53.0219	259.0865	0.0006
	LLR_1	**50.1520**	**99.5198**	0.0050

In Table 5, it is seen that the LLR_1 gives the best AVEPRESS** across the four different data sets and the best AVESSE across the data sets with $\sigma^2(\boldsymbol{\varepsilon}) = 1$, and $\sigma^2(\boldsymbol{\varepsilon}) = 5$, thus, emerging the overall best model in the study. The LLR gives the second best AVEPRESS** and the best AVESSE across the data sets with $\sigma^2(\boldsymbol{\varepsilon}) = 10$, and $\sigma^2(\boldsymbol{\varepsilon}) = 15$. The LLR gives the second best results because the data sets involve a single explanatory variable x_1.

Results in Table 6 display the exclusive superiority of the LLR_1 where it is seen to produce the best AVESSE and AVEPRESS** across the four data sets. As expected when we have more than one explanatory variable. LLR performance is seen to be inferior to that of LLR_1 for results in Eq. (30) where it is observed to give the poorest AVESSE across the four data sets under study.

In both Tables 5 and 6, we observe that the average sum of variances of the kernel weights from the modified GKF is larger than their counterparts from the existing GKF

1. Conclusion and Future Perspective

This paper proposes a modified version of the GKF for the purpose of generating kernel weights of enhanced flexibility for the LLR model. The proposed version of the GKFs was applied in LLR which is designated LLR_1 in this paper for each of reference.

LLR_1 consistently gives estimates with better goodness-of-fit than the OLS and LLR irrespective of the number of explanatory variables. These empirical results are proofs that the relative increase in variance of the weights from the proposed version of the GKF translates to a substantial increase in the performance of the LLR model.

Lastly, we emphasize that the estimate of the variance of the natural log of the kernel weights in Eq. (26) is just a means to an end. It was derived to give a lead on the direction or intervals to search for the optimal values of a_1, a_2 and b so as to get kernel weights of relatively higher variance than the variance of the kernel weights from the existing GKF, and the results in Table 3 validate the claim of higher variance. However, we accept that the assumption made in order to arrive at Eq. (10) oversimplified the variance obtained in Eq. (27). Fortunately, the variance of the kernel weights of the weight matrix W_i is neither required in the application of LLR or LLR_1nor in the computations of the goodness-of-fit in Table 4.

Future work will focus on the use of the proposed GKF in the application of hybrids regression model of OLS and LLR for handling data in areas including but not restricted to RSM settings.

References

Alvarez, M.J., Izarbe L., Viles, E. and Tanco, M. (2009). The use of genetic algorithm in response surface methodology. *Journal of Quality Technology and Quantitative Management*, 6(3), 295-309.

Anderson-Cook, C.M. and 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.

Ashish D., Harsh, B., Rajnik, H., Krina, V., Manoj, K.J.C. and Ramesh, V. (2024). Optimizing process conditions for ghewar development using response surface methodology. *European Journal of Nutrition and Food Safety*, 16(8), 98 – 109.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829 – 836.

Del Castillo E. (2007). *Process optimization: A statistical method*. New York: Springer International Series in Operations Research and Management Science.

Edionwe, E. and Mbegbu, J.I. (2014). Local bandwidths for improving the performance statistics of model robust regression 2. *Journal of Modern Applied Statistical Methods*, 13(2), 506-527.

Edionwe, E., Mbegbu, J.I. and 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. and Eguasa, O. (2023). Local quadratic Regression: Maximizing performance via a modified PRESS** for bandwidths selection. *The Philippine Statistician*, 72(1), 51 – 69.

Eguasa, O. and Edionwe, E. (2025). The adaptive nonparametric regression model and its residuals with a mixing parameter for response surface methodology: A novel blend. *FUPRE Journal of Scientific and Industrial Research*, 9(1), 305 – 320.

Eguasa, O. and Edionwe, E. (2025). Adaptive nonparametric regression model via a global mixing parameter for the multi-response problem. *Journal of the Nigerian Association of Mathematical Physics*, 69(1), 19 – 34.

Eguasa, O., Edionwe E. and Mbegbu J.I. (2022). Local linear regression and the problem of dimensionality: A remedial strategy via a new locally adaptive bandwidths selector. *Journal of Applied Statistics*, 50(6), 1283 – 1309.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: A variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, 57(2), 371 – 394.

Hardle, W., Muller, M., Sperlich, S. and Werwatz, A. (2005). *Nonparametric and semiparametric Models: An Introduction*. Berlin: Springer-Verlag.

He, Z., Zhu, P.F. and Park, S.H. (2012). A robust desirability function for multi-response surface optimization. *European Journal of Operational Research*, 221, 241-247.

Hofmann, T., Scholkofp, B. and Smola, A. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3), 1171 – 1220.

Khuri, A.I. (2017). Response surface methodology and its applications in agricultural and food sciences. *Biometrics & Biostatistics International Journal*, 5(5), 155 – 163.

Mays, J.E. and Birch, J.B. (2002). Smoothing for small samples with model misspecification: Nonparametric and semiparametric concerns. *Journal of Applied Statistics*, 29(7), 1023 – 1045.

Montgomery, D.C. (2009). *Introduction to statistical quality control*. 7th edition. New York: John Wiley & Sons.

Myers, R.H. (1999). Response surface methodology - Current status and future directions. *Journal of Quality Technology*, 31, 30-44.

Pickle, S.M., Robinson, T.J., Birch, J.B. and Anderson-Cook, C.M. (2008). A semi-parametric model to robust parameter design. *Journal of Statistical Planning and Inference*, 138, 114-131.

Sestelo, M., Villanueva, N.M., Meira-Machado, L. and Roca-Pardinas, J. (2017). An R package for nonparametric estimation and inference in life sciences. *Journal of Statistical Software*, 82(12), 1-27.

Shah, K.H., Montgomery, D.C. and Carlyle, W.M. (2004). Response surface modelling and optimization in multi-response experiments using seemingly unrelated regressions. *Quality Engineering*, 16, 387-397.

Ugur, O., Okan, B., Fethiye, G., Sesen, C. and Sahin, H.T. (2023). Application of response surface methodology for optimizing turbidity of daper recycling wastewater using microwave technology. *Asian Journal of Applied Chemistry Research*, 13(1), 13 – 22.

Wan, W. and Birch, J.B. (2011). A semi-parametric technique for multi-response optimization. *Journal of Quality and Reliability Engineering International*, 27, 47-59.

Yeniay, O. (2014). Comparative study of algorithm for response surface optimization. *Journal of Mathematical and Computational Applications*, 19, 93-104.

Zahra, S.A., Hossein, P. and Farrokh, A. (2020). Application of response surface methodology for optimization of zinc elimination from a polluted soil using tartaric acid. *Adsorption Science and Technology*, 38(3-4), 79 – 93.

Zheng, Q., Gallagher, C. and Kulasekera, K.B. (2013). Adaptively kernel weighted kernel regression. *Journal of Nonparametric Statistics*, 25(4), 855 – 872.

Appendix

Matlab Codes for LLR Regression

```
function pressdd=my_llr_modified_kernel_weights (D)
x1; vector of values of explanatory variable
x2; vector of values of explanatory variable
⋮
Xj; vector of values of explanatory variable
y;  vector of values of response variable
e=2.7183;
k=j; % No of explanatory variables
n=length(x1); % No of data points
const=ones(n,1); % vector of ones
X=[const x1 x2 … xj]; % LLR model matrix
% use #13 - #18 to preallocate dimensions for vectors/arrays
yLLR_max=zeros(n,1); % a n by 1 empty vector to store y estimates for large bandwidth
yLLRcv=zeros(n,1);   % a n by 1 empty vector to store leave-one-out estimates of y
d=zeros(n,1);        % a n by 1 empty vector to store diagonal elements of LLR Hat matrix
yLLR=zeros(n,1);     % a n by 1 empty vector to store y estimates for bandwidths
varkerweights=zeros(n,1); % a n by 1 empty vector to store variance of kernel weights
% use #19 - #29 to get the maximum SSE (SSEmax)
B=10;
for i=1:n;
w1max=((1/e).^(((x1(i)-x1)/B).^2)).*((1/e).^(((x2(i)-x2)/B).^2))*.
… .*((1/e).^(((xj(i)-xj)/B).^2));
WWmax=sum(w1max);
kerweight_max=w1max./WWmax;
Wmax=diag(kerweight_max); % n by n diagonal weight matrix
HatLLR_max=X(i,:)*((X'*Wmax*X)\(X'*Wmax)); % ith row of the LLR Hat matrix
yLLR_max(i)=HatLLR_max*y; % LLR estimate of y for large bandwidth;
end
```

28. SSEmax=sum((y-yLLR_max).^2);
29. % use #32 - #35 to search for optimal a(1) (that is a_1) and a(2) (that is b) for the kernel weights
30. %%%%%%%%%%
31. for i=1:n;
32. w1=((1/e).^(((a(1)*x1(i)-x1)./a(2)).^2)).*((1/e).^(((a(1)*x2(i)-x2)./a(2)).^2))*. ....*((1/e).^(((a(1)*xj(i)-xj)./a(2)).^2))
33. WW=sum(w1);
34. kerweight=w1./WW;
35. %%%% use #37 - #38 to get the variance of kernel weights of each diagonal weight matrix (W_i)
36. varkerweights(i)=var(kerweight); % variance of the vector of kernel weights for W_i
37. sum_var=sum(varkerweights); % sum of variances of all the weight matrices,
38. % use #40 - #43 to obtain leave-one-out estimates $y_{i,-i}^{LLR}$ of y_i.
39. kerweight(i,:)=[];y(i,:)=[];X(i,:)=[];
40. W=diag(kerweight); % diagonal weight matrix for leave-one-out estimate of y
41. par_est=(X'*W*X)\X'*W*y;
42. % use #44 - #46 to restore original dimensions of arrays
43. X=[const x1 x2 … xj];
44. y;
45. kerweight=w1./WW;
46. yLLRcv(i)=X(i,:)*per_est;
47. % use #48 -#53 get LLR estimates of responses
48. W=diag(kerweight); % n by n diagonal weight matrix for estimate of y
49. HLLR=X(i,:)*((X'*W*X)\X'*W);
50. d(i)=HLLR(1,i);
51. yLLR(i)=HLLR*y;
52. end
53. % use #55 - #62 to get minimum PRESS**
54. PRESS=sum((y-yLLRcv).^2);
55. SSE=sum((y-yLLR).^2);
56. % #58 - #61 ensures PRESS** is within acceptable range
57. PRESS_SV=PRESS/(((n-sum(a)))+((n-k-1)*((SSEmax-SSE)/SSEmax)));
58. if PRESS**<0; pdstar=99999999999999;
59. else pdstar=PRESS/(((n-sum(a)))+((n-k-1)*((SSEmax-SSE)/SSEmax)));
60. end
61. pressdd=pdstar,
62. SSE,
63. ymean=mean(y);
64. ySSM=sum((y-ymean).^2);
65. Rsqr=100*(1-(SSE/ySSM));
66. a(1)=a_1,
67. a(2)=b,

* a_1 and b for LLR_1 in Eq. (27) are coded as a(1) and a(2), respectively, in Line No. 32 in the syntax that Matlab accepts.

A PARAMETERIZED EXTENSION OF EXPONENTIAL DISCRIMINANT ANALYSIS FOR ENHANCED CLASSIFICATION

F. Meka[1] , J. E. Osemwenkhae[2], A. Iduseri[3]

[1] Department of Mathematics and Statistics, University of Delta, Agbor.
[2] Department of Statistics, University of Benin, Benin City.
[1]Corresponding author Email: fortune.meka@unidel.edu.ng
[2]Corresponding author Email: Joseph.osemwenkhae@uniben.edu.ng
[3]Corresponding author Email: Augustine.iduseri@uniben.edu.ng

Abstract

Exponential Discriminant Analysis (EDA) has emerged as a powerful technique for dimensionality reduction and classification, particularly in scenarios involving high-dimensional data. Traditional EDA employs the matrix exponential of scatter matrices to avoid singularity issues. However, this fixed exponential form lacks flexibility and may not fully capture the intrinsic structure of diverse datasets. In this paper, we propose a Modified Exponential Discriminant Analysis (MEDA) framework that generalizes EDA by incorporating tunable parameters as coefficients to the exponential base, and scatter matrices thereby controlling the scale and growth rate of the transformation, respectively. This formulation allows adaptive scaling of the scatter information, enabling better separation of classes. Experiments on benchmark datasets demonstrate that MEDA outperforms its counterparts in terms of classification accuracy.

Keywords: Linear Discriminant Analysis, Exponential Discriminant Analysis, Singularity, Classification.

1. Introduction

Discriminant Analysis is a multivariate statistical technique used to describe group separation or to predict group membership (Huberty and Olejnik, 2006). It is broadly categorized into **Linear Discriminant Analysis (LDA),** applicable when the covariance matrices of the groups are equal, and **Quadratic Discriminant Analysis (QDA),** used when the covariance matrices unequal. LDA itself has two major applications: **Predictive Discriminant Analysis (PDA)** and **Descriptive Discriminant Analysis (DDA)**. In PDA, the focus is on classifying individuals or objects into predefined groups based on allocation rules (Lobna et al., 2016; Ekhosuehi and Iduseri, 2022), while DDA emphasizes identifying the variables responsible for group differences (Stevens, 1996; Huberty and Olejnik, 2006; Iduseri and Osemwenkhae, 2016; Osemwenkhae et al., 2019).

Despite its wide applicability, a major drawback of the classical LDA is the **singularity of the within-class scatter matrix**, particularly when dealing with **high-dimensional data,** where the number of variables exceeds the number of samples. Such data often violate core LDA assumptions, including multivariate normality and independence of observations. This "curse of dimensionality" not only leads to singularity but also reduces the discriminative ability of the method (Park et al., 2007). Consequently, classical LDA becomes unsuitable for modern data environments such as image recognition, genomics, and other fields involving large feature spaces.

To overcome these limitations, several enhancements to LDA have been developed. One of the earliest solutions involved **two-stage methods**, such as Principal Component Analysis followed by LDA (PCA+LDA), which first reduces data dimensionality before classification (Belhumeur et al., 1997). Other researchers introduced **regularization-based techniques**, where a small constant or penalty term is added to the within-class scatter matrix to prevent singularity and improve robustness (Zhang et al., 2010; Mahadi et al., 2024). Another class of approaches includes **transformational methods**. These applies mathematical transformations to the scatter matrices, such as the matrix exponential, to ensure positive definiteness and enhance group separability (Zhang et al., 2010; Ran, 2018).

Among these variants, **Exponential Discriminant Analysis (EDA)** stands out for addressing the singularity problem by employing the matrix exponential transformation, which ensures a positive definite matrix, and achieves enhanced hit rates. However, direct matrix exponentiation can lead to instability due to the rapid growth of matrix powers within the exponential function, especially for large matrices. These numerical challenges can compromise precision and limit the practical utility of EDA in some applications.

To provide a more robust and adaptable framework, this study introduces the **Modified Exponential Discriminant Analysis (MEDA)**. The proposed method extends the concept of EDA by incorporating a flexible scaling approach that adjusts both the magnitude and rate of transformation. This adaptive feature enables better control over the transformation process, enhances stability, and promotes improved class separability in high-dimensional classification tasks.

2. Methodology

In discriminant analysis, the central objective is to identify a transformation that enhances the separation among multiple groups while minimizing variation within each group. Consider a classification problem involving n distinct classes, denoted as $C_1, C_2, \dots, C_n$ containing $N_1, N_2, \dots N_n$ samples respectively. Each sample represents a feature vector in a multidimensional space. The task is to determine an optimal projection that maps these high-dimensional samples onto a lower-dimensional subspace in which the projected class means are as distinct as possible, and the overlap between classes is minimized.

This is done by maximizing Fisher's linear discriminant function:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{1}$$

This calculation requires first obtaining the scatter matrix of the data

$$S_i = \sum_j^N \left(x_j^i - \mu_i\right)\left(x_j^i - \mu_i\right)^T \tag{2}$$

where μ_i represents the average of all samples from class C_i

$$\mu_i = \frac{1}{N_i}\sum_j^N x_j \tag{3}$$

These scatter matrices sum to the within-class scatter matrix (4). The difference in these scatter matrices is similarly the between-class variance (5).

The between class scatter matrix S_B and the within class scatter matrix S_W is given by:

$$S_W = \sum_{i=1}^n \sum_{j=1}^{N_i} \left(x_j^i - \mu_i\right)\left(x_j^i - \mu_i\right)^T \tag{4}$$

$$S_B = \sum_{i=1}^n N_i\,(\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

The optimal LDA projection can then be obtained by solving the characteristics equation:

$$|S_W^{-1} S_B - \lambda I| = 0 \tag{6}$$

Whenever the inverse of S_W cannot be taken due to singularity, the failure of equation (6) will lead to the failure of LDA. As such, leveraging on the property of matrix exponential becomes needful. As the exponential of a matrix is given by:

$$\exp(A) = \sum_{j=0}^{\infty} \frac{A^j}{j!} = I + A + \frac{A^2}{2!} + \cdots + \frac{A^s}{s!} + \cdots \tag{7}$$

where I is the identity matrix. Clearly, from exponent rule

$$\exp(-A) \cdot \exp(A) = \exp(A - A) = I$$

This implies that

$$exp^{-1}(A) = \exp(-A)$$

As such, every matrix exponential is invertible. Therefore, the Exponential Discriminant Analysis (Zhang *et al.*, 2010) redefined the LDA criterion as:

$$\varrho = \max_{v \in R^{d \times t}} \frac{tr(V^T \exp(S_B) V)}{tr(V^T \exp(S_W) V)}, \quad (8)$$

and in this case, v can be obtained from solving the exponential eigenproblem

$$\exp(S_B)x = \lambda \exp(S_w)x \quad (9)$$

This scatter matrix transformation amplifies the differences in class scatter structures by exponentially weighting the eigenvalue contributions, thereby improving class separability. In the present study, we generalize this concept through the adoption of a parameterized exponential, $a[\exp(k(S_B))]$ and $a[\exp(k(S_W))]$, where a and k are adjustable scalar parameters. This generalization, herein referred to as the **Modified Exponential Discriminant Analysis (MEDA),** introduces additional flexibility in modeling the exponential behavior of scatter matrices, with the objective of maximizing interclass distances and achieving superior discriminative performance. The parameter $k > 0$ serves as a **scaling factor** applied to the scatter matrices before the exponential transformation. It controls the rate of growth of the matrix exponential and thereby stabilizes the numerical behaviour of the exponentiation operation, particularly under small sample conditions or high dimensional settings. A small positive value of k prevents the exponential from diverging excessively.

2.1 Modified Exponential Discriminant Analysis (MEDA)

Theorem 1. Let $f(x) = ae^{kx}$ be a general exponential function. If $a > 0$ and $x > 0$, the inequality $ae^{kx} > e^x$ holds.

The Figure 1 below shows this property of the exponential function. Figure 1 reveals that varying the values of k brings about deviations in the exponential curve.

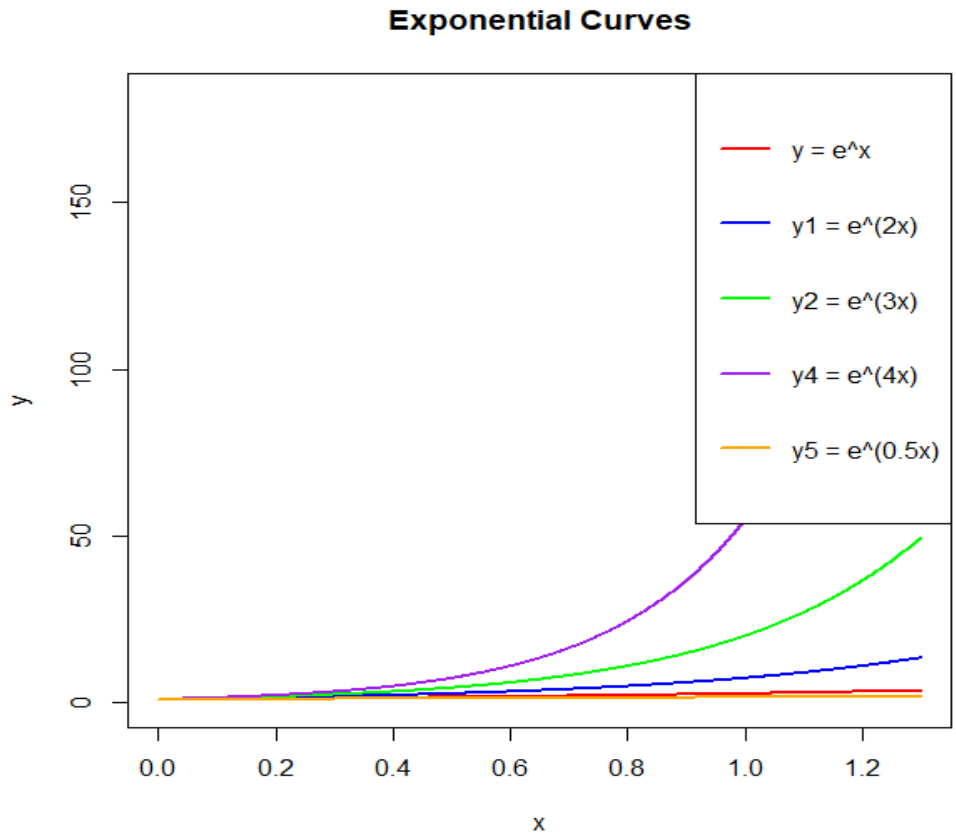


Figure 1: Exponential curves for varied values of k

If $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d\times n}$ is the data matrix for LDA, where x_j represents the jth training image, the first step in the LDA rule is to compute the between and within class scatter matrices given by equations (4) and (5)

Following the exponential function

$$y = ae^{kx} \tag{10}$$

where $a > 0$, $k > 0$ and x will be replaced by the scatter matrices.

We obtain the matrices kS_W and kS_B their exponentials as exp (kS_w) and exp (kS_B) and finally obtain the matrices a[exp (S_w)] and a[exp (kS_B)]. If we denote

$$\mathrm{W} = \mathrm{a}[\exp\,(kS_w)] \tag{11}$$

and

$$\mathrm{B} = \mathrm{a}[\exp\,(kS_B)] \tag{12}$$

The MEDA criterion may now be written as:

$$C = \max_{v\in R^{d\times t}} \frac{tr(V^T BV)}{tr(V^T WV)} \tag{13}$$

Here, V is the optimal projection matrix, which is obtained by solving the eigenvalue problem.

This is subject to the constraint:

$$V^T WV = I_t \tag{14}$$

Where, I_t is a $t \times t$ identity matrix.

This constraint ensures that the solution is scale-invariant and that the discriminant projection vectors are mutually orthogonal in the metric induced by the within-class scatter matrix W.

$$Bx = \lambda Wx \tag{15}$$

The MEDA rule is stated as follows:

Input: The data matrix $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d\times n}$ where x_j represents the j-th training image.

Step1. Compute the matrices S_B, S_W, kS_B, kS_W, exp(kS_B), exp(kS_W); a[exp(kS_B)] and a[exp(kS_W)]

Step 2. Compute the eigenvectors $\{x_i\}$ and eigenvalues $\{\lambda_i\}$ of $[\mathrm{a}\{\exp(kS_W)\}]^{-1}\mathrm{a}[\exp(kS_B)]$

Step 3. Sort the eigenvectors $V = \{x_i\}$ according to $\{\lambda_i\}$ in decreasing order;

Step 4. Orthogonalize the columns of the projection matrix V.

Output: The projection matrix V

If a = k = 1, the MEDA translates the EDA.

The MEDA method avoids the singularity of the within-class scatter matrix due to the exponential property and enhanced group separation from the EDA through the effect of k.

2.3 Discriminant Strength of the MEDA

LDA finds an optimal projection by simultaneously maximizing the between-class distance and minimizing the within-class distance. The two distances can be measured by the traces of two scatter matrices as:

$$trace(s_b) = \lambda_{b1} + \lambda_{b2} + \cdots + \lambda_{bn} \tag{16}$$

$$trace(s_w) = \lambda_{w1} + \lambda_{w2} + \cdots + \lambda_{wn} \tag{17}$$

Following the general exponential function $f(x) = ae^{kx}$, the distances can be given as

$$trace\left(ae^{ks_b}\right) = ae^{k\lambda_{b1}} + ae^{k\lambda_{b2}} + \cdots + ae^{k\lambda_{bn}} \tag{18}$$

$$trace\left(ae^{ks_w}\right) = ae^{k\lambda_{w1}} + ae^{k\lambda_{w2}} + \cdots + ae^{k\lambda_{wn}} \tag{19}$$

In general, Since the eigenvalues of $trace\left(ae^{ks_b}\right)$ are often used to describe the separation between classes, while the eigenvalues of $trace\left(ae^{ks_w}\right)$ are often used to describe the closeness of the samples within classes. Hence, the discriminant vector that corresponds to the bigger ratio of $\frac{\lambda_{bi}}{\lambda_{bi}}$ owns stronger

discriminant power. Since $\lambda_{bi} > \lambda_{wi}$, then the inequality $e^{\lambda_{bi}} > e^{\lambda_{wi}}$ holds. And if this inequality holds, then $ae^{k\lambda_{bi}} > ae^{k\lambda_{wi}}\ \forall\ k > 0$, having in mind that the $a's$ will cancel out. It is easy to conclude that since

$$ae^{k\lambda_{bi}} > e^{\lambda_{bi}} > \lambda_{bi}$$

Then

$$\frac{ae^{k\lambda_{bi}}}{ae^{k\lambda_{wi}}} > \frac{e^{\lambda_{bi}}}{e^{\lambda_{wi}}} > \frac{\lambda_{bi}}{\lambda_{wi}} \tag{20}$$

2.4 Performance Metrics

The following performance metrics were used to evaluate our proposed method (Sokolova and Lapalme 2009)

Accuracy: This is the proportion of correct prediction. It is determined using the formula

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \times 100$$

$$= \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{21}$$

Precision: The number of the selected items that are relevant is determined using the formula:

$$Precision = \frac{TP}{TP+FP} \times 100 \tag{22}$$

Recall: The number of relevant items were selected are given by the relation

$$Recall = \frac{TP}{TP+FN} \times 100 \tag{23}$$

Where TP and TN stands for true positive and true negative respectively. Also FP and FN stands for false positive and false negative respectively.

F1 Score: This is the harmonic mean of precision and recall. It balances the trade-off between Precision and Recall.

$$F1\ Score = 2 \times \frac{Precition \times Recall}{Precition + Recall} \times 100 \tag{24}$$

Runtime: This reveals the **total time** taken by the examined procedures to run from start to finish (Gougeon, 2013). The system.time() function in R was used to determine the runtime for each procedure.

4. Results and Discussions

In this section, we carried out some numerical experiments to evaluate the performance of our proposed method. All numerical experiments were carried out on R software 4.3.1 implementation.

The projection matrix in all the methods comprises of the $C - 1$ dominant discriminant vectors, where C is the number of classes. To assess the effectiveness of our proposed rule, we employed three benchmark datasets: the Iris dataset, the Wine dataset, and the ORL face dataset.

The Iris dataset is one of the most well-known benchmark datasets in pattern recognition and machine learning. It consists of 150 samples of iris flowers, equally divided among three species: Iris setosa, Iris versicolor, and Iris virginica. Each sample is described by four numerical features measured in centimeters. They are: sepal length, sepal width, petal length, and petal width. The dataset is linearly

separable for some classes, making it suitable for testing classification algorithms, dimensionality reduction methods, and discriminant analysis techniques.
The Wine dataset is a classic multivariate dataset often used for evaluating classification and feature selection methods. It contains 178 samples of wine derived from three different cultivars grown in the same region of Italy. Each sample is characterized by 13 continuous features that represent various chemical and physical properties of the wine, such as alcohol content, malic acid, ash, flavanoids, and color intensity. The dataset poses a moderate challenge due to overlapping feature distributions, making it a valuable resource for testing algorithms designed to handle real-world variability.

The ORL (Olivetti Research Laboratory) Face dataset is widely used in face recognition and image classification research. It contains a collection of 400 grayscale images of 40 individuals, with 10 different images per person. The images were captured under varying conditions, including changes in facial expressions, lighting, and pose, while maintaining a consistent background. Each image has a resolution of 92×112 pixels, though they are often resized for computational efficiency. The dataset is particularly useful for evaluating dimensionality reduction, feature extraction, and discriminant analysis methods in the context of human face recognition.
The proposed MEDA method is compared with the classical LDA and the EDA. In order to show the performance of MEDA, its values on performance metrics (Accuracy, Precision, Recall, F1 score and Runtime) was compared with that of its counterparts illustrated using the three datasets.

CASE 1: The Iris Dataset

The R software outputs of the performance metrics for the three methods used are presented in Table 1 and further illustrated in Figure 2 and 3 respectively.

Table 1. Performance Metrics for three Methods used on the Iris Dataset

Method	Accuracy	Precision	Recall	F1 Score	Runtime
LDA	89.333	89.443	89.333	89.239	0.065
EDA	87.333	87.665	87.333	87.124	0.052
MEDA	**98.000**	**98.182**	**98.000**	**97.995**	**0.048**

Table 1 shows the different performance metrics of the three classification methods on the Iris Dataset. The best performance are in bold prints. The MEDA competed favorably and outperformed other methods in each of the metrics used. MEDA achieved a classification accuracy of 98.00%, compared to 89.33% and 87.33% for LDA and EDA respectively. This illustrates MEDA's superior inter-class separation via the parameterized exponential approach, which is inspired by the exponential transformation used to improve class separability. Figure 2 gives visual representation of this result.

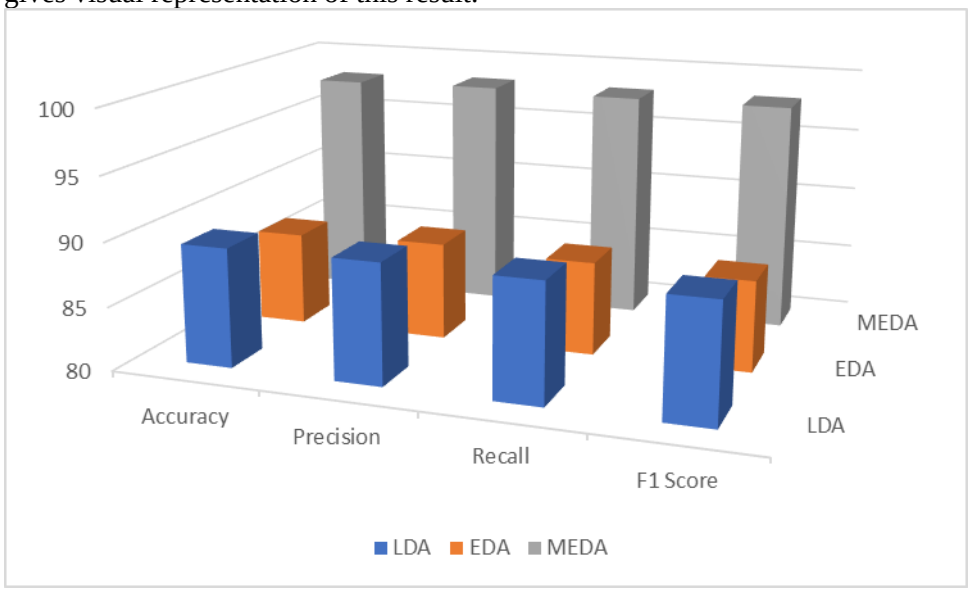


Fig 2. Graphical Comparison of the four Different Performance Metrics on

the Iris Dataset

Figure 3 below also gives a visual representation of the runtime comparison forLDA, EDA and MEDA. In the context of their metrics, Figure 3 shows that MEDA had a shorter training time which is important for frequent retraining and time sensitive system/models. In addition, it reflects MEDA's efficiency, scalability, and suitability on the Iris dataset compared to LDA and EDA.

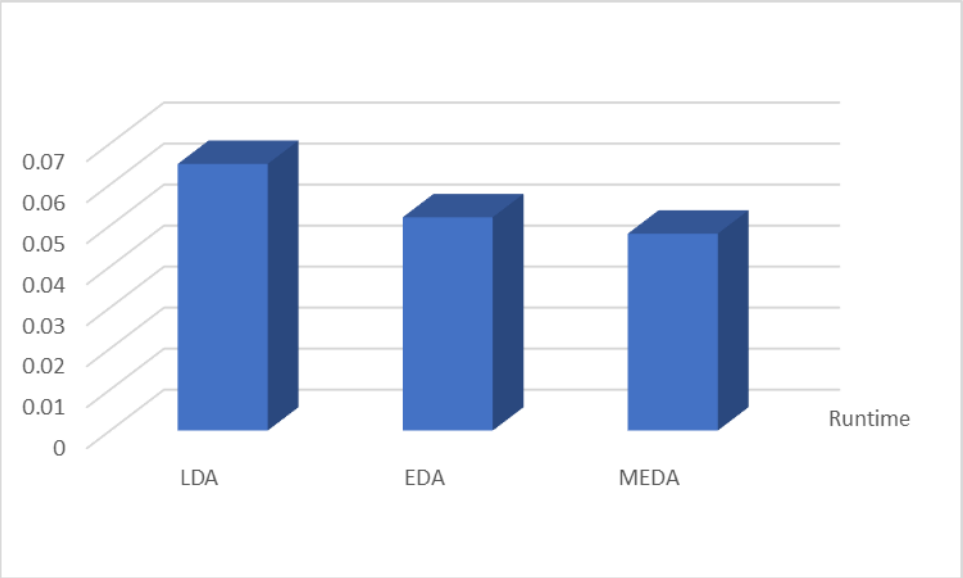


Fig 3: Runtime Comparisons for the three Methods using the Iris Dataset

To further assess the performance of the MEDA, results from the wine dataset and the ORL dataset are presented in Table 2 and 3 below.

CASE 2: The Wine Dataset

The performance metric results generated in R for the three applied methods are shown in Table 2 and further depicted in Figures 4 and 5.

Table 2. Performance Metrics for three Methods used on the Wine Dataset

Method	Accuracy	Precision	Recall	F1 Score	Runtime
LDA	66.831	68.872	70.031	68.402	0.061
EDA	85.361	86.342	85.290	85.167	0.050
MEDA	**90.967**	**91.998**	**91.404**	**91.115**	**0.049**

The results in **Table 2** show a clear performance improvement from LDA to EDA and further to MEDA on the Wine dataset, across all evaluation metrics:

Accuracy increases significantly from **66.83% (LDA)** to **85.36% (EDA),** and then to **90.97% (MEDA),** indicating that both EDA and MEDA are more effective at capturing the class structure in the data.

Precision, Recall, and **F1 Score** follow a similar upward trend, confirming that the improvement is consistent and not skewed toward a specific class performance aspect.

Runtime is comparable for all methods, with MEDA even slightly faster (0.049s) than EDA (0.050s) and LDA (0.061s), showing that the added complexity of MEDA does not introduce a computational burden.

Overall, **MEDA outperforms both LDA and EDA** in classification performance while maintaining high efficiency, suggesting that incorporating parameter tuning and matrix exponentiation (as in MEDA) enhances class separation. Figure 4 gives a visual representation of this result.

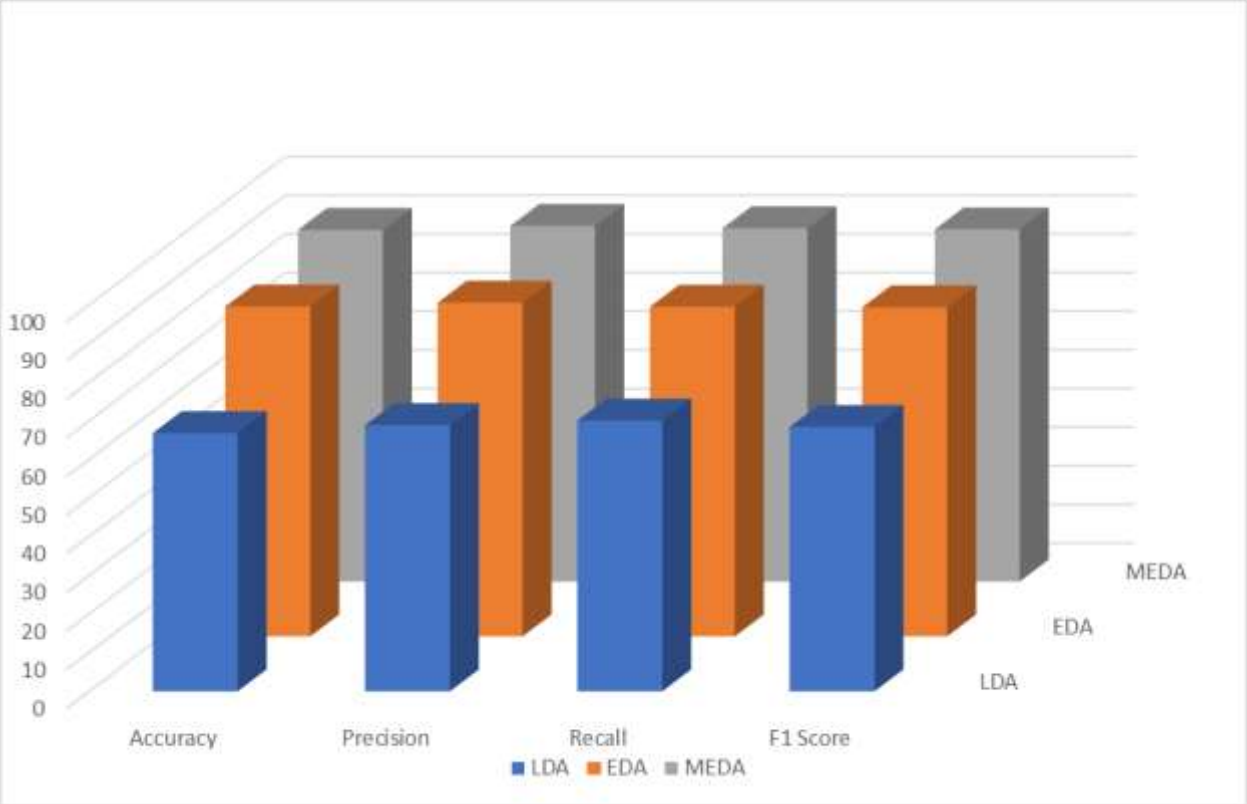


Fig 4A. Graphical Comparison of the four Different Performance Metrics on the Wine Dataset

Figure 5 provides a visual comparison of the runtimes for LDA, EDA, and MEDA. The results indicate that MEDA required less training time, a key advantage for applications involving frequent retraining or time-sensitive models.

This also highlights MEDA's efficiency, scalability, and suitability for the Wine dataset relative to LDA and EDA.

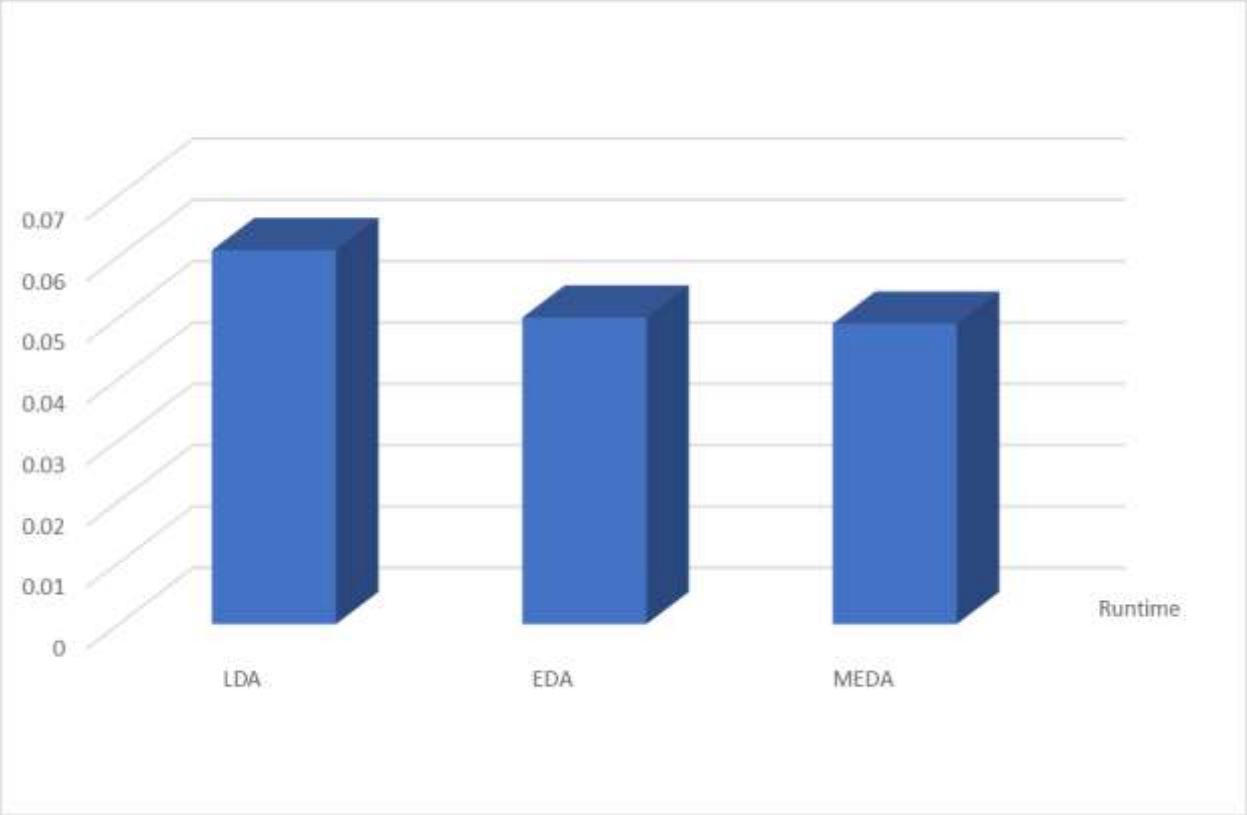


Fig 5: Runtime Comparisons for the three Methods using the Wine Dataset

CASE 3: The ORL Dataset

The performance metrics obtained using the ORL dataset for the three methods are summarized in Table 3 and further visualized in Figures 6 and 7.

Table 3. Performance Metrics for three Methods used on the ORL Dataset

Method	Accuracy	Precision	Recall	F1 Score	Runtime
LDA	6.341	5.491	7.733	44.028	0.644
EDA	17.073	17.617	17.941	51.036	0.656
MEDA	**34.390**	**36.163**	**35.350**	**63.914**	**0.631**

The results in **Table 3** using the ORL face dataset demonstrate the limitations of traditional LDA and EDA when applied to complex, high-dimensional image data:

LDA performs poorly, with only **6.34% accuracy**, and abnormally inflated **F1 Score (44.03)** compared to its low precision and recall. This likely reflects misaligned predictions across classes or extreme imbalance in confusion matrices.

EDA improves accuracy to **17.07%,** indicating that using exponential scatter matrices helps capture more structure than standard LDA, but still struggles to model the high variability in facial images.

MEDA shows a **notable leap to 34.39% accuracy**, along with balanced gains in precision, recall, and F1 Score. This suggests that MEDA is better at extracting meaningful features from high-dimensional image data, likely due to its ability to amplify class-separating components.

Despite these improvements, **overall accuracy remains modest**, emphasizing the ORL dataset's complexity and the need for either more powerful nonlinear models or better preprocessing (e.g., dimensionality reduction or deep features). **Runtime across methods is similar,** so gains in accuracy do not come at a computational cost. Figure 6 shows the graphical representation of this result.

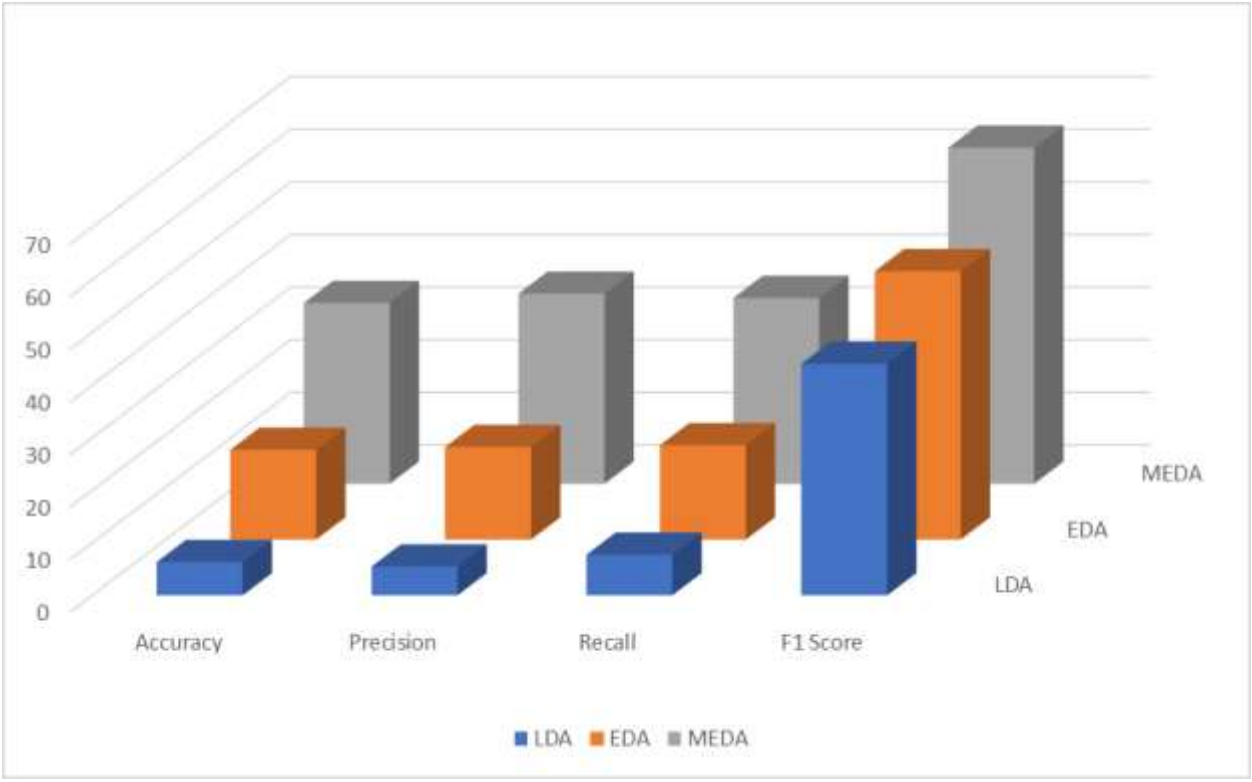


Fig 6. Graphical Comparison of the four Different Performance Metrics on the ORL Dataset

Figure 7 below also gives a visual representation of the runtime comparison for LDA, EDA and MED A. In the context of their metrics. It shows that MEDA had a shorter training time which is important for frequent retraining and time sensitive system/models. In addition, it reflects MEDA's efficiency, scalability, and suitability on the ORL data compared to LDA and EDA.

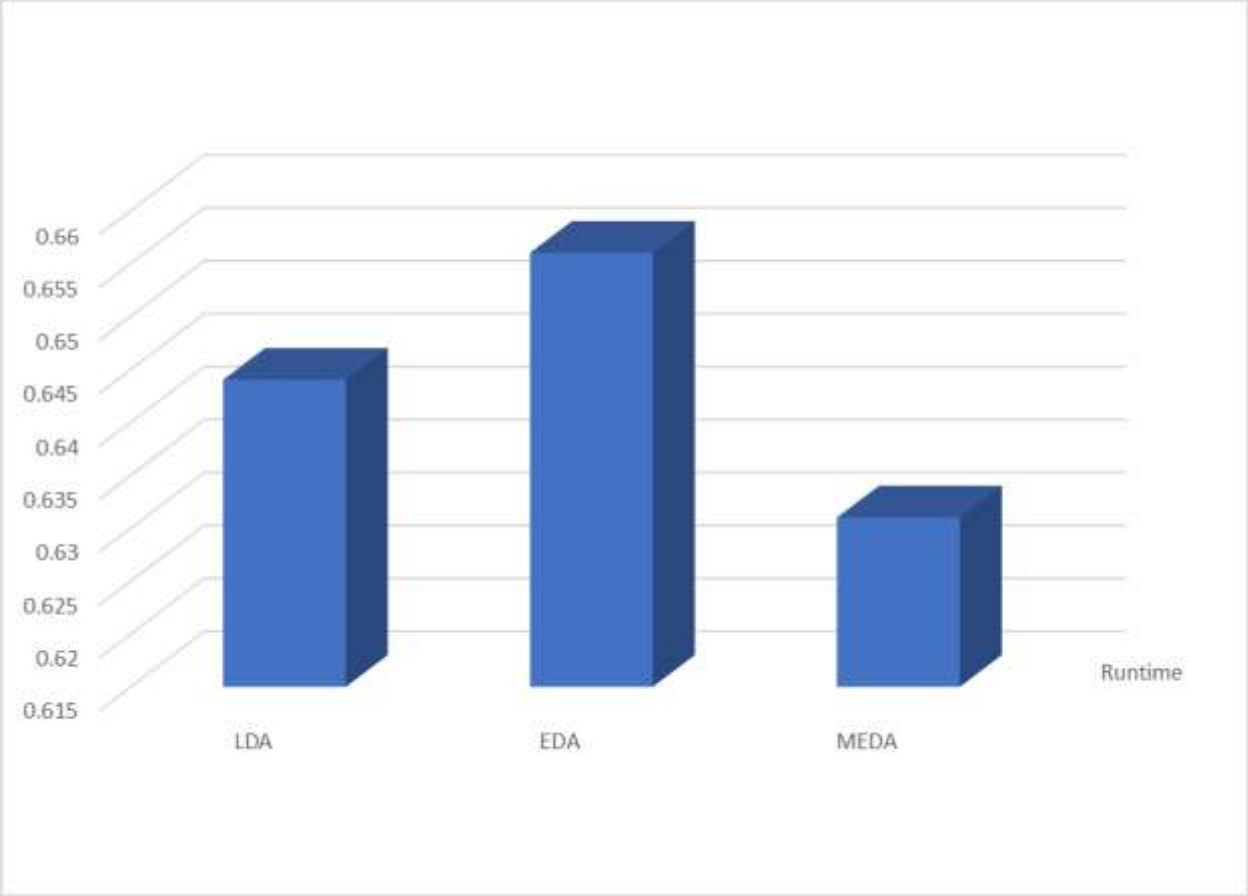


Fig 7: Runtime Comparisons for the three Methods using the ORL Dataset

CASE 4: Simulation Studies

To further evaluate our model, a simulation study was carried out on the methods, using varying sample sizes. The classical EDA method was initially included as a benchmark. However, under small sample size scenarios the direct matrix exponential of the within-class scatter matrix became numerically unstable, often resulting in non-invertible exponential matrices and non-finite eigenvalues. This instability has also been reported in the literature for high-dimensional or small-n cases. Therefore EDA was excluded from the simulation comparison because it could not produce valid results under the simulated conditions. The comparison therefore focuses on LDA (classical baseline) and the proposed MEDA method. The result is presented in Figure 8 below:

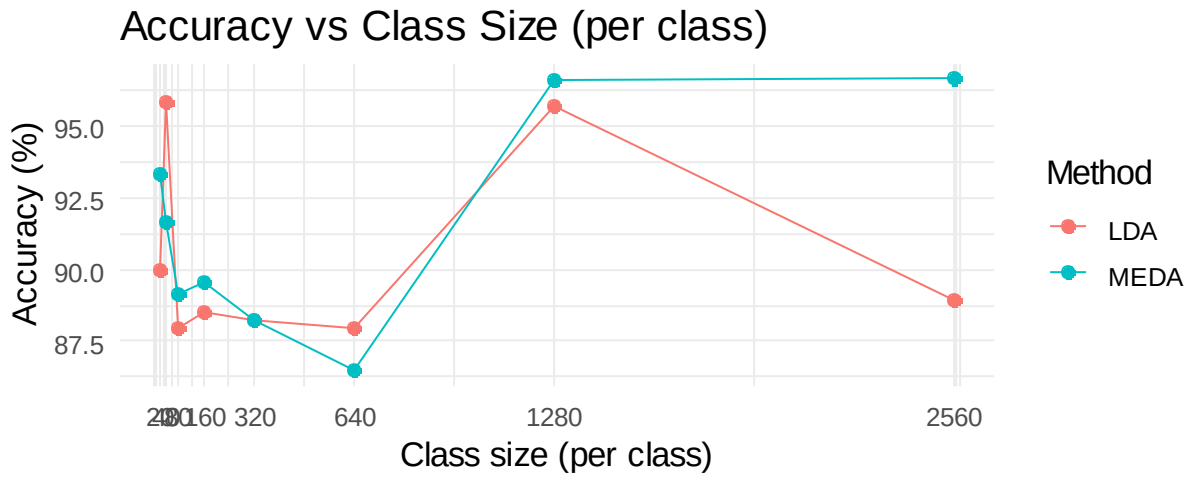


Figure 8: Comparing Accuracy of LDA and MEDA on Simulated Data

Figure 8 demonstrates the superiority of the MEDA over the LDA especially when sample size was above 1000. EDA could not be included because it failed numerically in small sample settings due to matrix exponential instability, which is a known theoretical limitation. MEDA fixes that limitation by adding a scaling approach.

5. Summary and Recommendations

In this paper, a modified exponential discriminant analysis (MEDA) is proposed to enhance the performance of the EDA method. The proposed MEDA method introduces scalar coefficients a and k to the exponential function and the scatter matrices respectively. Consequently, the exponential growth rate is now controlled and more stable, giving room for better performance. *Using a very small scalar* k *in the exponential mapping tends to linearize the matrix exponential and regularizes the scatter matrix, improving classification performance to 98%. In contrast, varying the scalar* a *demonstrated minimal impact, as it tends to cancel out in the discriminant ratio. The MEDA is therefore recommended for classification when dealing with high dimensional data.*

References

Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs_Fisherface: recognition using Class-Specifc Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 711-720.

Ekhosuehi, V. U. and Iduseri, A. (2022). Logistic Regression and Discriminant Analysis of academic staff mix by rank via research performance in a university setting. *Journal of the Nigerian Statistical Association*, 34, 40-60.

Gougeon, P. (2013). *Measuring and optimizing computational performance in R. R Journal*, 5(1), 27–36. https://journal.r-project.org/archive/2013/RJ-2013-001/index.html

Huberty, C.J. and Olejnik, O. (2006). *Applied MANOVA and Discriminant Analysis*, John Wiley and Sons, Inc., Hoboken, New Jersey.

Iduseri, A. and Osemwenkhae, J. E. (2016). Using discriminant analysis to identify major prerequisites for success in specific courses of study in a university system. *Journal of Nigerian Association of Mathematical Physics*, 33, 99-106.

Liu, J., Cai, X., and Niranjan, M. (2023). *GO-LDA: Generalised Optimal Linear Discriminant Analysis*. arXiv preprint arXiv:2305.14568

Lobna, A, Afif, M, and Sonia, G. (2016). The consumer loans payment default predictive model. An application in Tunisian commercial bank. *Asian Economic and Financial Review*, 6(1), 27- 42.

Mahadi, M., Ballal, T., Moinuddin, M., Al-Naffouri, T. Y., and Al-Saggaf, U. M. (2024). Regularized linear discriminant analysis using a nonlinear covariance matrix estimator. *IEEE Transactions on Signal Processing*, 72, 1049–1062. https://doi.org/10.1109/TSP.2024.3361715

Martínez, A. M., and Kak, A. C. (2001). *PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 228–233. https://doi.org/10.1109/34.908974

Osemwenkhae, J. E., Iduseri, A. and Meka, F. (2019). Determinants of the level of stress experienced by teachers at different educational levels: a descriptive discriminant approach. *Journal of the Nigerian Statistical Association*, 31, 64-80.

Park, H. Drake, B. L., Lee, S. and Park, C. H. (2007). Fast Linear Discriminant Analysis using QR Decomposition and RegularizationCollege of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332, U.S.A.

Ran, R., Fang, B., Wu, X. and Zhang, S. (2018). A simple and effective generalization of exponential matrix discriminant analysis and its application to face recognition. *IEICE TRANS. INF. & SYST.*, 101(1).

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences.* (3rd ed.), Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sokolova, M., and Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks. Information Processing and Management*, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Ye J. and Li Q. (2005). A two-stage linear discriminant analysis via QR decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 929-941.

Zhang, D., and Ye, J. (2014). Learning regularized LDA by clustering. *Pattern Recognition*, 47(9), 3034–3042. https://doi.org/10.1016/j.patcog.2014.03.006

Zhang, T.P., Fang, B., Tang, Y.Y., Shang, Z., and Xu, B. (2010). Generalized discriminant analysis: A matrix exponential approach. *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, 40(1),186–197.

AI-POWERED CLIMATE AND WEATHER FORECASTING TOOLS, FOR IMPROVED AGRICULTURAL PLANNING IN SOUTH-SOUTH OF NIGERIA

Kerry Christopher Chinedu[1] , Nkemjika Chukwukammadu Onyedikachukwu[2] and Amagoh Maureen Nkechi[3]

Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-uku, Delta State.

kerryedu1@gmail.com[1] onyedikachukwukammadu.nkemjika@gmail.com[2]
nkechimaureen0001@gmail.com[3]

Abstract

The increasing unpredictability of weather patterns caused by climate change poses serious challenges to sustainable agricultural planning in Nigeria. As farming activities in the South-South region remain highly climate-sensitive, the integration of artificial intelligence (AI)-powered climate and weather forecasting tools has emerged as a viable means for improving decision-making and building climate resilience. This study examined the availability, utilization, benefits, and challenges of AI-powered climate and weather forecasting tools for improved agricultural planning in South-South Nigeria. Five specific objectives guided the study: to identify the existing AI-powered forecasting tools, examine their extent of utilization, determine their perceived benefits, assess the challenges hindering their adoption, and propose strategies for improved integration into agricultural systems. The study adopted a descriptive survey design. The population comprised 132 registered crop farmers and 100 agricultural extension agents across the six South-South states, giving a total of 232 respondents. Data were collected using a structured questionnaire and analyzed with descriptive statistics (mean and standard deviation) to answer the research questions, while the Z-test was used to test five null hypotheses at a 0.05 level of significance. Findings revealed that while AI-powered climate and weather forecasting tools are moderately available and utilized in the region, their adoption is hindered by lack of infrastructure, literacy, and policy-related challenges. Nevertheless, both farmers and extension agents acknowledged the significant benefits of these tools in improving farm-level planning and productivity. The study concludes that a supportive ecosystem comprising digital infrastructure, capacity building, and enabling policy frameworks is vital for strengthening the integration of AI-powered tools in agricultural planning and enhancing climate resilience in South-South Nigeria.

Keywords: Artificial Intelligence, Climate Forecasting Tools, Agricultural Planning, Weather Forecasting, South-South Nigeria

Introduction

Agriculture plays a central role in the economic development and sustenance of rural livelihoods in Nigeria, particularly within regions such as South-South Nigeria, where farming supports a significant proportion of the population. Despite the abundance of natural resources and fertile soil in the region, the increasing unpredictability of climatic conditions has continued to pose serious constraints to agricultural productivity (Alleh et al., 2019). Erratic rainfall patterns, flooding, and prolonged dry seasons have disrupted traditional farming calendars and decision-making processes (Ogundeji et al., 2020). Many farmers in the region rely on indigenous or historical weather knowledge, which has become increasingly unreliable in the face of changing climatic conditions (Olaniyi et al., 2018). As extreme weather events become more frequent and intense, there is a growing need for innovative tools that enable farmers to make informed and proactive decisions in agricultural planning (Adenle, 2020; Ejiofor et al., 2020).

Globally, artificial intelligence (AI)-powered climate and weather forecasting tools have emerged as transformative technologies in agriculture. These tools employ machine learning algorithms, satellite imagery, big data analytics, and cloud-based systems to generate hyper-localized and accurate weather predictions that support precision farming. Some of the widely known and applied AI-powered forecasting tools include IBM's Watson Decision Platform for Agriculture, Google's DeepMind Weather AI, Microsoft's AI for Earth, aWhere Climate Intelligence, and Climate FieldView used by Bayer and other agribusinesses for predictive analytics. Others such as Farmerline, CropIn SmartFarm, PlantVillage Nuru, TAMSAT, and CHIRPS integrate weather data and satellite monitoring to provide real-time forecasts and advisory services to smallholder farmers across Africa and Asia (World Bank, 2021; Kshetri, 2021; Gholami et al., 2017). In countries like Kenya, AI-powered platforms such as Farmerline have successfully delivered mobile-based weather and market information to over 500,000 farmers, resulting in improved productivity and reduced losses (World Bank, 2021). Similarly, in India, machine learning models have been deployed to predict monsoon patterns, assisting millions of farmers in timing planting and harvest activities (Kshetri, 2021). These international experiences demonstrate that AI-powered climate and weather forecasting tools can significantly enhance resilience and

productivity when supported by robust digital infrastructure and institutional frameworks (Ayanlade et al., 2017).

In Nigeria, however, the integration of AI into agriculture is still evolving, and the adoption of such forecasting tools remains limited (Ekekwe, 2019; Nwakile et al., 2020). Some of the AI-powered tools that have been introduced or piloted within the country include Crop2Weather, Farmerline, AgroCloud, PlantVillage Nuru, aWhere, and satellite-based services using CHIRPS and TAMSAT rainfall datasets. These tools are used by agricultural development projects, research institutions, and private agritech firms to provide climate and weather information. However, within South-South Nigeria, access to and use of these technologies are still moderate and fragmented. Mobile-based alert systems, AI-driven rainfall predictors, and satellite-imaging applications are among the few tools recognized and used by farmers and agricultural extension agents in the region. Their use is largely concentrated in government and NGO-supported programmes, while rural farmers remain constrained by poor digital literacy, low internet connectivity, and inadequate awareness (Aina et al., 2023; Fadairo & Oluwatayo, 2022).

This limited adoption contrasts sharply with the global trend and underscores the need for empirical evidence on how these tools are currently utilized and perceived within the local agricultural context. As observed by Ayinde et al. (2022), digital innovations can only transform agricultural productivity if they are effectively domesticated and made accessible to end users. Similarly, Obadimu et al. (2020) emphasized that the success of climate-smart agriculture in Nigeria depends on integrating AI tools into extension systems and local institutions.

For any AI-powered solution to thrive within Nigeria's agricultural ecosystem, it must be context-specific, accessible, and supported by enabling policies and partnerships. The South-South region, with its dual exposure to inland and coastal climatic challenges, presents a strategic environment for studying the integration of AI-powered climate and weather forecasting tools in agricultural planning. Yet, there is a paucity of empirical research that documents the types of tools available, their level of utilization, perceived benefits, and challenges to adoption among farmers and extension officers (Aina et al., 2023). Without such understanding, interventions aimed at promoting digital agriculture may fail to yield tangible impacts on food production and climate resilience (Obadimu et al., 2020).

This study therefore focuses on AI-powered climate and weather forecasting tools, for improved agricultural planning in South-South of Nigeria. Specifically, it investigates the tools currently available, the extent of their utilization, the perceived benefits and challenges associated with their use, and strategies for their effective integration into agricultural systems within the region.

Purpose of the Study

The general purpose of this study is to analyze the use of AI-powered climate and weather forecasting tools for improving agricultural planning in South-South Nigeria. Specifically, the study seeks to;

1. Identify the AI-powered climate and weather forecasting tools currently available to farmers and extension agents in South-South Nigeria.
2. Examine the extent to which these AI-powered tools are utilized in agricultural planning within the region.
3. Determine the perceived benefits of using AI-powered climate and weather forecasting tools on agricultural productivity among farmers and extension agents.
4. Assess the challenges hindering the adoption and effective utilization of AI-powered climate and weather forecasting tools in South-South Nigeria.
5. Propose strategies for enhancing the integration and sustainability of AI-powered climate and weather forecasting tools in agricultural planning across the region

Research Questions

The following research questions will guide the study:

1. What are the AI-powered climate and weather forecasting tools currently available to farmers and extension agents in South-South Nigeria?
2. To what extent are these AI-powered tools utilized in agricultural planning within the region?
3. What are the perceived benefits of using AI-powered climate and weather forecasting tools on agricultural productivity among farmers and extension agents?
4. What are the challenges hindering the adoption and effective utilization of AI-powered climate and weather forecasting tools in South-South Nigeria?
5. What are the strategies for enhancing the integration and sustainability of AI-powered climate and weather forecasting tools in agricultural planning across the region

Hypotheses

The following null hypotheses were tested at the 0.05 level of significance;

H_{o1}: There's no significant difference between the mean responses of farmers and extension agents on the AI-powered climate and weather forecasting tools available in South-South Nigeria.

H_{o2}: There's no significant difference between the mean responses of farmers and extension agents on the extent of utilization of AI-powered climate and weather forecasting tools for agricultural planning.

H_{o3}: There's no significant difference between the mean responses of farmers and extension agents on the perceived benefits of AI-powered climate and weather forecasting tools on agricultural productivity.

H_{o4}: There's no significant difference between the mean responses of farmers and extension agents on the challenges affecting the adoption and use of AI-powered climate and weather forecasting tools in South-South Nigeria.

H_{o5}: There's no significant difference between the mean responses of farmers and extension agents on the strategies for enhancing the integration of AI-powered climate and weather forecasting tools into agricultural planning in the region.

Methodology

Given the systematic collection and description of data, this study adopted a descriptive survey research design, which involves determining the opinions, characteristics, or status of a population without manipulating variables (Nworgu 2015). The aforementioned method is considered appropriate since this study seeks to obtain the views of farmers and extension agents on the availability, utilization, benefits, and challenges of using AI-powered climate and weather forecasting tools for agricultural planning in South-South Nigeria.

This study was conducted in South-South region of Nigeria, which comprises of six states: Akwa Ibom, Bayelsa, Cross River, Delta, Edo, and Rivers. The region was selected due to its strong agricultural potential and vulnerability to climate variability, flooding, and erratic rainfall—conditions that make it suitable for assessing the role of AI-powered forecasting tools in agricultural planning and decision-making. From the population of interest, 256 respondents were sampled, although 232 questionnaires retrieved indicated 132 registered crop farmers actively engaged in agricultural production and 100

agricultural extension agents working with government agencies and development programs across the six South-South states.

The collection of data was carried out by trained research assistants, with the application of simple random selection method. A structured questionnaire developed based on insights from relevant literature on AI applications in agriculture was used for data collection. The questionnaire was organized into five sections corresponding to the five research objectives. Each item was rated using a four-point Likert scale of; Strongly Agree (4), Agree (3), Disagree (2), and Strongly Disagree (1).

The questionnaire was validated by professionals in relevant fields of interest, including data analyst. Their suggestions led to revisions that improved the clarity, structure, and content validity of the instrument. The reliability of the instrument was established using the Cronbach Alpha method, which yielded a coefficient of 0.84, indicating a high level of internal consistency.

Data were analyzed using descriptive statistics (mean and standard deviation) to answer the research questions. For decision-making, item(s) with mean score of 2.50 and above was interpreted as "Agreed," otherwise (< 2.50) considered "Disagreed." While the null hypotheses was carried out using Z-test at 0.05 level of significance. A null hypothesis was upheld if the calculated p-value is greater than 0.05, otherwise fail to accept.

Results

Table 1: Mean Ratings and Z-Test Analysis of Farmers and Extension Agents on AI-Powered Climate and Weather Forecasting Tools Available in South-South Nigeria

N = 232 (132 Farmers & 100 Extension Agents)

s/n	AI-Powered Forecasting Tools Available to Farmers	$\bar{X}$ (G)	SD	$\bar{X}_1$	SD_1	$\bar{X}_2$	SD_2	Sig	Rem	Dec
1	Mobile-based weather alert systems (e.g., Crop2Weather, Farmerline)	3.32	0.78	3.28	0.80	3.38	0.74	0.19	**A**	**NS**
2	AI-driven platforms for short-term rainfall prediction	3.21	0.82	3.15	0.85	3.29	0.79	0.11	**A**	**NS**
3	Satellite imaging apps for monitoring weather conditions	3.05	0.91	2.98	0.94	3.13	0.87	0.13	**A**	**NS**
4	Predictive analytics integrated into agricultural extension services	3.10	0.86	3.03	0.88	3.18	0.83	0.09	**A**	**NS**

5	AI-enabled weather dashboards used by cooperatives or farm clusters	2.94	0.93	2.90	0.96	3.00	0.89	0.32	**A**	**NS**
6	AI-assisted radio forecast summaries tailored to local zones	3.16	0.77	3.11	0.81	3.22	0.72	0.21	**A**	**NS**
7	Voice-assisted AI weather information on mobile devices	2.87	0.96	2.80	0.99	2.97	0.91	0.17	**A**	**NS**
8	Interactive voice response (IVR) platforms providing weather data	3.08	0.84	3.05	0.88	3.13	0.80	0.28	**A**	**NS**
9	Chatbots integrated into agricultural apps for weather guidance	2.85	0.92	2.79	0.94	2.94	0.88	0.25	**A**	**NS**
10	AI-supported community weather stations linked to cloud servers	3.00	0.88	2.95	0.90	3.06	0.86	0.26	**A**	**NS**
11	Machine learning models used by NGOs for regional forecasting	3.12	0.81	3.07	0.85	3.18	0.77	0.23	**A**	**NS**

Key: $\bar{X}$ (G) = Grand Mean, SD = Standard Deviation, Sig = Significance Level, NS = Not Significant, $\bar{X}_1$ = Farmers' mean, X_2 = Extension Agent' mean, SD_1 = Farmers' SD, SD_2 = Extension Agents' SD, Rem = Remarks, A = Agreed

The data in Table 1 reveal that all eleven items had mean values ranging from 2.85 to 3.32, indicating that respondents agreed that AI-powered climate and weather forecasting tools are moderately available in South-South Nigeria. The standard deviation values (0.72–0.99) indicate moderate agreement among respondents. The significance levels (p-values) for the eleven items ranged from 0.09 to 0.32, all greater than 0.05, indicating no significant difference between the mean responses of farmers and extension agents. Therefore, the null hypothesis (H_{01}) is upheld.

Table 2: Mean Ratings and Z-Test Analysis of Farmers and Extension Agents on the Extent of Utilization of AI-Powered Weather Forecasting Tools for Agricultural Planning

N = 232 (132 Farmers & 100 Extension Agents)

s/n	Utilization of AI-Powered Forecasting Tools in Agricultural Planning	$\bar{X}$ (G)	SD	$\bar{X}_1$	SD_1	$\bar{X}_2$	SD_2	Sig	Rem	Dec
1	Used to determine optimal planting dates	3.26	0.80	3.22	0.82	3.31	0.76	0.20	**A**	**NS**
2	Guides decisions on irrigation scheduling	3.12	0.83	3.07	0.84	3.19	0.81	0.16	**A**	**NS**
3	Supports planning of fertilizer application	3.01	0.86	2.98	0.88	3.05	0.83	0.39	**A**	**NS**
4	Aids in early harvesting to avoid rainfall damage	3.18	0.82	3.11	0.86	3.27	0.77	0.14	**A**	**NS**

5	Helps identify weather risks for pest/disease control	3.24	0.81	3.20	0.84	3.30	0.78	0.23	**A**	**NS**
6	Used for market planning tied to weather trends	2.92	0.88	2.87	0.90	2.99	0.86	0.30	**A**	**NS**
7	Integrated into farm management apps used by farmers	2.84	0.91	2.79	0.94	2.91	0.88	0.26	**A**	**NS**
8	Referenced by cooperatives when planning group planting activities	3.09	0.83	3.03	0.87	3.17	0.78	0.18	**A**	**NS**
9	Used in community training programmes by extension agents	3.20	0.78	3.11	0.82	3.33	0.72	0.06	**A**	**NS**
10	Consulted during advisory services and field visits	3.28	0.76	3.24	0.78	3.34	0.73	0.13	**A**	**NS**
11	Incorporated in risk assessment for crop insurance	2.93	0.89	2.90	0.91	2.96	0.87	0.47	**A**	**NS**
12	Used for seasonal planning across multiple farming cycles	3.15	0.79	3.08	0.83	3.23	0.74	0.17	**A**	**NS**

Key: $\bar{X}$ (G) = Grand Mean, SD = Standard Deviation, Sig = Significance Level, NS = Not Significant, $\bar{X}_1$ = Farmers' mean, X_2 = Extension Agent' mean, SD_1 = Farmers' SD, SD_2 = Extension Agents' SD, Rem = Remarks, A = Agreed

The results in Table 2 show that all twelve items had mean scores between 2.84 and 3.28, reflecting moderate utilization of AI-powered climate and weather forecasting tools across planning activities such as planting, irrigation, pest control, and harvesting. Standard deviation values (0.72–0.94) show acceptable consistency in responses. All p-values (0.06–0.47) exceed 0.05, indicating no statistically significant difference between farmers and extension agents. Hence, the null hypothesis (H_{02}) is upheld.

Table 3: Mean Ratings and Z-Test Analysis of Farmers and Extension Agents on Perceived Benefits of AI-Powered Climate Forecasting on Agricultural Productivity

N = 232 (132 Farmers & 100 Extension Agents)

s/n	Perceived Benefits of AI-Powered Climate Forecasting	$\bar{X}$ (G)	SD	$\bar{X}_1$	SD_1	$\bar{X}_2$	SD_2	Sig	Rem	Dec
1	Improved accuracy of seasonal forecasts	3.42	0.74	3.40	0.76	3.45	0.71	0.51	**A**	**NS**
2	Reduced crop failure due to early warning alerts	3.37	0.76	3.33	0.78	3.41	0.74	0.35	**A**	**NS**
3	Enhanced farm-level planning and preparedness	3.33	0.78	3.29	0.81	3.38	0.74	0.31	**A**	**NS**
4	Better timing of input use (e.g., seeds, fertilizers)	3.30	0.80	3.25	0.83	3.36	0.76	0.28	**A**	**NS**
5	Increased productivity through timely farming decisions	3.39	0.73	3.36	0.75	3.44	0.70	0.29	**A**	**NS**

6	Reduced post-harvest losses due to early weather alerts	3.25	0.82	3.18	0.85	3.33	0.78	0.16	**A**	**NS**
7	Boost in farmers' confidence to take calculated risks	3.15	0.84	3.10	0.88	3.22	0.79	0.22	**A**	**NS**
8	Increased market access through informed harvest timing	3.08	0.87	3.05	0.90	3.12	0.83	0.42	**A**	**NS**
9	Strengthened advisory services using forecast data	3.29	0.78	3.23	0.81	3.37	0.74	0.15	**A**	**NS**
10	Reduction in unnecessary farm labour and input waste	3.12	0.85	3.06	0.87	3.19	0.81	0.19	**A**	**NS**

Key: $\bar{X}$ (G) = Grand Mean, SD = Standard Deviation, Sig = Significance Level, NS = Not Significant, $\bar{X}_1$ = Farmers' mean, X_2 = Extension Agent' mean, SD_1 = Farmers' SD, SD_2 = Extension Agents' SD, Rem = Remarks, A = Agreed

All ten items recorded mean scores between 3.08 and 3.42, indicating that respondents agreed on the benefits of AI-powered climate and weather forecasting tools for agricultural productivity. These benefits include improved accuracy of forecasts, reduced crop losses, better input timing, and increased confidence in decision-making. All p-values (0.15–0.51) are greater than 0.05, showing no significant difference between the mean responses of both groups. The null hypothesis (H_{03}) is therefore upheld.

Table 4: Mean Ratings and Z-Test Analysis of Farmers and Extension Agents on Challenges Affecting Adoption of AI-Powered Weather Forecasting Tools

N = 232 (132 Farmers & 100 Extension Agents)

s/n	Challenges to AI Forecasting Tool Adoption	$\bar{X}$ (G)	SD	$\bar{X}_1$	SD_1	$\bar{X}_2$	SD_2	Sig	Rem	Dec
1	Limited access to smartphones and digital devices	3.35	0.79	3.39	0.78	3.30	0.80	0.33	**A**	**NS**
2	Poor network and internet connectivity in rural areas	3.42	0.72	3.45	0.70	3.38	0.74	0.40	**A**	**NS**
3	High cost of data subscription and device maintenance	3.31	0.77	3.36	0.76	3.25	0.78	0.28	**A**	**NS**
4	Inadequate awareness of AI-based weather tools	3.28	0.80	3.33	0.78	3.22	0.82	0.25	**A**	**NS**
5	Low digital literacy among farmers	3.37	0.76	3.41	0.74	3.32	0.78	0.36	**A**	**NS**
6	Lack of training on use of forecasting platforms	3.39	0.73	3.42	0.72	3.35	0.74	0.41	**A**	**NS**
7	Absence of government support or subsidy for digital tools	3.20	0.84	3.16	0.87	3.25	0.80	0.32	**A**	**NS**
8	Language barriers and user interface issues in tools	3.05	0.89	3.01	0.91	3.10	0.87	0.43	**A**	**NS**
9	Lack of trust in AI-generated weather information	2.91	0.90	2.87	0.91	2.96	0.89	0.39	**A**	**NS**

10	Limited integration with traditional extension systems	3.18	0.83	3.12	0.86	3.26	0.80	0.22	**A**	**NS**

Key: $\bar{X}$ (G) = Grand Mean, SD = Standard Deviation, Sig = Significance Level, NS = Not Significant, $\bar{X}_1$ = Farmers' mean, X_2 = Extension Agent' mean, SD_1 = Farmers' SD, SD_2 = Extension Agents' SD, Rem = Remarks, A = Agreed

Results in Table 4 show mean values ranging from 2.91 to 3.42, indicating that respondents agreed these challenges hinder adoption. Key constraints identified include poor network access, high cost of data, limited training, and low digital literacy. All p-values (0.22–0.43) exceed 0.05, indicating no significant difference between farmers and extension agents. Therefore, the null hypothesis (H_{04}) is upheld.

Table 5: Mean Ratings and t-test Analysis of Farmers and Extension Agents on Strategies for Enhancing Integration of AI-Driven Climate Forecasting Tools

N = 232 (132 Farmers & 100 Extension Agents)

s/n	Strategies to Enhance AI Forecasting Integration	$\bar{X}$ (G)	SD	$\bar{X}_1$	SD_1	$\bar{X}_2$	SD_2	Sig	Rem	Dec
1	Organize regular training for farmers on use of AI tools	3.47	0.71	3.44	0.73	3.51	0.69	0.34	**A**	**NS**
2	Incorporate AI tools into existing agricultural extension services	3.41	0.74	3.36	0.76	3.47	0.71	0.26	**A**	**NS**
3	Provide digital literacy programmes at community level	3.45	0.72	3.42	0.74	3.49	0.70	0.38	**A**	**NS**
4	Increase access to smartphones through cooperative support	3.29	0.80	3.25	0.83	3.34	0.76	0.30	**A**	**NS**
5	Improve rural internet infrastructure for connectivity	3.39	0.77	3.36	0.79	3.43	0.75	0.42	**A**	**NS**
6	Translate AI interfaces into local languages	3.20	0.85	3.18	0.88	3.22	0.82	0.46	**A**	**NS**
7	Develop farmer-friendly AI platforms with voice prompts	3.26	0.82	3.23	0.84	3.30	0.79	0.40	**A**	**NS**
8	Subsidize cost of AI weather apps for farmers	3.33	0.78	3.30	0.80	3.36	0.76	0.41	**A**	**NS**
9	Collaborate with telecom providers for weather SMS alerts	3.37	0.76	3.32	0.78	3.44	0.73	0.29	**A**	**NS**
10	Establish AI-supported community weather centres	3.30	0.81	3.26	0.84	3.36	0.78	0.27	**A**	**NS**
11	Integrate AI forecast data into national agricultural policy	3.42	0.75	3.39	0.77	3.46	0.72	0.35	**A**	**NS**

Key: $\bar{X}$ (G) = Grand Mean, SD = Standard Deviation, Sig = Significance Level, NS = Not Significant, $\bar{X}_1$ = Farmers' mean, X_2 = Extension Agent' mean, SD_1 = Farmers' SD, SD_2 = Extension Agents' SD, Rem = Remarks, A = Agreed

All eleven items recorded mean scores above 3.20, showing strong agreement on strategies to enhance integration of AI-powered climate and weather forecasting tools. Key strategies include farmer training, improved rural connectivity, local language interfaces, and policy support. All p-values (0.26–0.46) exceed 0.05, implying no significant difference between both groups. Thus, the null hypothesis (H_{05}) is upheld.

Discussion of Findings

The findings from the study revealed that AI-powered weather forecasting tools are moderately available to farmers in South-South Nigeria. Tools such as mobile-based alert systems, AI-driven rainfall predictors, satellite imaging apps, and AI-enabled community dashboards were identified by both farmers and extension agents. The general agreement among respondents implies a growing presence of digital climate services in the region, though not yet widespread. This finding aligns with the observations of Obadimu et al. (2020), who noted that although AI-enabled tools are emerging in Nigeria, their presence remains concentrated in select pilot areas and donor-supported programmes. In a related view, Ayanlade et al. (2017) observed that the growing availability of digital climate services offers a unique opportunity to supplement traditional knowledge, particularly in regions vulnerable to erratic weather patterns. These tools, though still underutilized, represent a critical step toward climate-resilient agriculture in the region.

Results from research question two showed that the extent of utilization of AI-powered weather forecasting tools among farmers is moderate. Respondents indicated that these tools were mainly used for determining planting periods, guiding irrigation schedules, input planning, and accessing advisory services. Although extension agents reported slightly higher use levels than farmers, the overall pattern shows that AI tools are yet to be deeply integrated into day-to-day farm planning. This supports the findings of Fadairo and Oluwatayo (2022), who reported that while digital tools exist, farmers often lack the capacity to apply them consistently for informed planning. According to Ayinde et al. (2022), the limited uptake of AI in Nigerian agriculture can be traced to gaps in training, infrastructure, and

institutional coordination, even when awareness exists. Therefore, without structured efforts to scale access and build technical know-how, the tools' potential may not be fully harnessed.

The study revealed that both farmers and extension agents perceive AI-powered weather forecasting tools as beneficial to agricultural productivity. Among the benefits identified were improved forecast accuracy, reduced crop losses, enhanced preparedness, better input use, and increased confidence in decision-making. These perceived benefits validate the assertions made in international experiences. For instance, Kshetri (2021) noted that AI models deployed in India and Kenya have helped farmers anticipate weather shocks and increase their productivity. Similarly, the World Bank (2021) observed that when AI is used to generate timely forecasts, farmers in vulnerable regions improve yield outcomes through smarter planning. In the Nigerian context, Ekekwe (2019) stressed that bridging traditional practices with predictive technology can empower rural communities to navigate climate unpredictability more effectively. These findings therefore suggest that farmers are not only aware of the benefits but are willing to embrace tools that offer tangible improvements.

The findings indicated that several challenges affect the adoption of AI-powered weather forecasting tools in South-South Nigeria. These include poor internet access, low digital literacy, lack of training, high data costs, and limited trust in AI-generated outputs. These challenges are consistent with those identified by Obadimu et al. (2020), who linked low adoption rates in rural Nigeria to infrastructural and capacity-related gaps. According to Aina et al. (2023), without addressing the digital divide, introducing AI tools will only benefit a small segment of the farming population. In addition, Olaniyi et al. (2018) highlighted that the absence of community-level support structures and low awareness further hinder the translation of climate data into actionable farm-level practices. These findings underscore the need for targeted interventions that address both technological and human dimensions of AI adoption.

The study found that multiple strategies could enhance the integration of AI-driven climate forecasting tools into agricultural planning. Respondents agreed on the importance of digital literacy training, community-based weather centres, AI-integrated extension services, and policy support. These strategies are supported by Raedmaekers, Svatikova and Yearwood (2015), who emphasized the importance of inclusive frameworks and institutional coordination in facilitating AI adoption in

developing countries. Ekekwe (2019) similarly noted that localized training and access to low-cost tools are essential for meaningful digital integration. Furthermore, Adenle (2020) advocated for strategic public-private partnerships to provide infrastructure and capacity-building in AI-enabled agricultural systems. These findings highlight the urgency for stakeholders to collaboratively design context-specific frameworks that support the operationalization of AI tools at the grassroots level.

Conclusion

Findings from the study revealed that AI-powered weather forecasting tools such as mobile-based alerts, rainfall prediction platforms, and AI-enabled dashboards are moderately available in the study area. However, the extent of utilization of these tools by farmers remains moderate, as most rely only on basic features and lack deep integration into full-scale planning. Nevertheless, both farmers and extension agents recognized the positive impact of AI forecasting on productivity, particularly in supporting timely planting, reducing crop losses, and improving overall farm-level decision-making.

Despite these benefits, challenges such as poor internet infrastructures, high costs, low awareness, and inadequate training were identified as significant barriers. The study also established several strategies that can enhance the integration of AI tools in agriculture, including training/educating farmers, rural internet investments, digital literacy programmes, and policies that support local innovation.

In conclusion, while AI-powered climate forecasting tools hold great promise for improving agricultural planning in South-South Nigeria, their impact will remain limited unless structural, educational, and institutional challenges are addressed. With the right support, these tools can serve as transformative instruments for building climate resilience and increasing productivity in Nigeria's agricultural sector.

Recommendations

Based on the findings of the study, the following recommendations are made;

1. Government agencies should invest in rural internet and digital infrastructure to improve access to AI-powered forecasting platforms for farmers in underserved communities. And profound policies that will affluence the implementation/sustainability of AI-powered climate and weather forecasting tools for farmers.
2. Tertiary institutions and research centres should collaborate with technological firms to develop simple, farmer-friendly AI interfaces that are adaptable to local needs and languages.

3. Extension services should be strengthened to include training on AI weather forecasting tools as part of regular outreach and advisory activities.
4. Development partners and NGOs should support capacity-building programmes focused on digital literacy, especially for smallholder farmers with limited exposure to technology.
5. Telecommunication providers should partner with agricultural agencies to deliver forecast information via SMS, voice calls, and radio formats accessible to farmers with limited smartphone access.

References

Adenle, A. A. (2020). Unlocking the potential of agricultural innovation for climate change adaptation in Nigeria. *Climate and Development, 12*(3), 241–252.

Aina, O. S., Lawal, A. M., & Bolarinwa, K. K. (2023). Digital technologies and agricultural productivity in sub-Saharan Africa: A systematic review. *African Journal of Science, Technology, Innovation and Development, 15*(1), 32–44.

Alleh, A.O., Ejiofor, T.E., Nwakile, T.C. Ogbonna, E.K. (2019). Information needs of agricultural science senior secondary school students on agro-processing and pest/diseases control for self-reliance in a green economy in the federal capital territory, Abuja. *Journal of Association of Vocational and Technical Educators of Nigeria, 24*(2), 113 - 120.

Ayanlade, A., Radeny, M., & Morton, J. F. (2017). Comparing smallholder farmers' perception of climate change with meteorological data: A case study from southwestern Nigeria. *Weather and Climate Extremes, 15*, 24–33. https://doi.org/10.1016/j.wace.2016.12.001

Ayinde, O. E., Omotesho, O. A., & Falola, A. (2022). Barriers to climate-smart agriculture adoption among smallholder farmers in Nigeria: Policy implications. *Environmental Challenges, 8*, 100563. https://doi.org/10.1016/j.envc.2022.100563

Ejiofor, T.E., Alleh, A. O. & Nwakile, T.C. (2020). Information needs on green economy for self-reliance by agricultural science students in senior secondary schools in the Federal Capital Territory, Abuja. *Journal of Agricultural Education Teachers Association of Nigeria, 4*(1), 190 – 198. https://www.researchgate.net/publication/358736556

Ekekwe, N. (2019). *How digital technology is changing farming in Africa*. Harvard Business Review. https://hbr.org/2019/05/how-digital-technology-is-changing-farming-in-africa

Fadairo, O. S., & Oluwatayo, I. B. (2022). Role of climate information services in enhancing farmers' adaptive capacity to climate variability in Nigeria. *Climate Risk Management, 35*, 100404. https://doi.org/10.1016/j.crm.2022.100404

Gholami, R., Watson, R. T., Hasan, H., Molla, A., & Bjørn-Andersen, N. (2017). Information systems solutions for environmental sustainability: How can we do more? *Journal of the Association for Information Systems, 17*(8), 521–536. https://doi.org/10.17705/1jais.00436

Kshetri, N. (2021). 1.5 billion farmers around the world need better weather forecasts — AI can help. *The Conversation*. https://theconversation.com/1-5-billion-farmers-around-the-world-need-better-weather-forecasts-ai-can-help-162540

Nwakile, T. C**.,** Ekenta, L. U., Onah, F.C. & Ekwueme, S. U. (2022). Constraints and enhancement measures for adoption of high skill green jobs in tertiary institutions for food security in Anambra State. *Journal of Agricultural Education Teachers Association of Nigeria, 6*(2), 14 – 20. https://www.researchgate.net/publication/374550646

Nwakile, T.C., Ojiako, C. C. & Akwam, N. C. (2020). Constraints and enhancement measures to development of skills by youths for employment in green jobs in agriculture in Anambra State.

Journal of Agricultural Education Teachers Association of Nigeria, 4(1), 153 – 161. https://www.researchgate.net/publication/363539117.

Obadimu, A., Adebiyi, A., & Salawu, R. (2020). Enhancing climate-smart agriculture in Nigeria: Role of ICTs and data analytics. *Nigerian Journal of Agricultural Extension, 21*(2), 50–62.

Ogundeji, A. A., Babu, S. C., & Manyong, V. M. (2020). Farmers' perception of climate change adaptation constraints in Nigeria: Evidence from the Nigeria agricultural policy project. *Climate and Development, 12*(8), 677–687. https://doi.org/10.1080/17565529.2019.1701975

Olaniyi, O. A., Oyekale, A. S., & Adepoju, A. A. (2018). Assessment of climate change and weather variability on maize production in Nigeria. *Journal of Agriculture and Environment for International Development, 112*(2), 277–297. https://doi.org/10.12895/jaeid.20182.735

Raedmaekers, K., Svatikova, K., & Yearwood, J. (2015). Facilitating green skills and jobs in developing countries. Paris: *Agence Française de Développement (AFD)*. https://www.afd.fr/en/ressources/facilitating-green-skills-and-jobs-developing-countries

World Bank. (2021). *Harnessing artificial intelligence for climate resilience in agriculture*. https://www.worldbank.org/en/news/feature/2021/07/27/harnessing-artificial-intelligence-for-climate-resilience-in-agriculture

COMPARATIVE ANALYSIS OF STOCHASTIC MODELS AND MACHINE LEARNING ALGORITHMS FOR INFLATION RATE PREDICTION IN NIGERIA

Edesiri Bridget Nkemnole[1] and Abiodun Simeon Oyelami[2]

Department of Statistics, University of Lagos, Akoka, Lagos, Nigeria.

[1]Email: bnkemnole@unilag,edu,ng and [2]Email: abbeytech4uall@gmail.com

Corresponding Author: Email: abbeytech4uall@gmail.com

Abstract

Inflation forecasting is a critical aspect of economic planning, particularly in developing economies like Nigeria, where inflation volatility significantly impacts policymaking, investment decisions, and overall economic stability. This study evaluated the predictive performance of traditional stochastic processes such as Vasicek, Cox–Ingersoll–Ross (CIR), and Geometric Brownian Motion (GBM) against three machine learning algorithms: Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), in modeling Nigeria's inflation trends. The analysis was based on "All-items inflation rates" data spanning from January 2003 to December 2024. The study uncovered that whereas stochastic models successfully captured the hypothetical inflationary change, their predictive accuracy was moderately restricted compared to machine learning methods. In particular, the Random Forest model outperforms stochastic approaches in terms of accuracy, robustness, and overall performance across key evaluation metrics. This research advocates for a paradigm shift in Nigeria`s economic modelling strategies by emphasizing the integration of advanced machine learning methods into inflation forecasting.

Keywords: Inflation Rate, Forecasting, Stochastic Models, Machine Learning, Random Forest,

1. INTRODUCTION

Inflation volatility significantly affects economic stability and policy-making in Nigeria. Accurate forecasting is essential for decision-makers, prompting interest in both traditional stochastic processes and modern machine learning (ML) methods. This study evaluated and compared these methods in forecasting Nigeria's inflation rate. The ability to accurately predict inflation trends is essential for government agencies, financial institutions, and businesses to formulate effective strategies for mitigating economic risks. Over the years, various approaches have been developed to model and forecast inflation rates, ranging from traditional stochastic processes to modern machine learning algorithms.

Stochastic models have long been employed in economic and financial forecasting due to their ability to capture the randomness inherent in macroeconomic variables. Among the most prominent stochastic models are the Geometric Brownian Motion (GBM) model (Mensah *et al*., 2023), the Vasicek model (Nadarajan and Nur-Firyal, 2024), and the Cox–Ingersoll–Ross (CIR) model (Bernaschi *et al.,* 2007).

In contrast, the advent of machine learning has introduced a paradigm shift in economic forecasting, providing more flexible and data-driven approaches for modeling inflation dynamics. Machine learning algorithms such as Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (KNN) have shown promising results in forecasting complex economic indicators (Ayyildiz and Iskenderoglu, 2024). Recent studies have extended this approach by integrating machine learning models with stochastic frameworks such as the Hidden Markov Model (HMM) to better capture the structural shifts and transition patterns in inflation. For instance, Nkemnole, Wulu, and Osubu (2024) applied KNN and Long Short-Term Memory (LSTM) models enhanced by HMM to forecast inflation rates and transition patterns in Nigeria, highlighting the critical role of GDP per capita as a significant predictor. Their findings underscore the growing relevance of hybrid machine learning-stochastic methods in improving the accuracy and interpretability of inflation forecasting models in developing economies.

Stochastic models have long been instrumental in modeling economic variables, including inflation rate. Techniques such as Autoregressive Integrated Moving Average (ARIMA) have been widely adopted due to their capacity to capture temporal dependencies in time series data (Box & Jenkins, 1976). A study examining Nigeria's inflation data from 1960 to 1999 utilized a time series approach, revealing that a quadratic trend effectively captured the inflation dynamics during that period (Iwueze and Akpanta, 2006). Furthermore, a recent study on Nigeria’s inflation rate prediction using a Bayesian Seasonal Autoregressive Integrated Moving Average (Bayesian SARIMA) model emphasized the presence of a strong seasonal effect in the inflation rate (Oyelami & Ogundeji, 2025).

The integration of machine learning (ML) techniques has revolutionized inflation forecasting by enabling the detection of complex, nonlinear patterns in economic data. In Nigeria, studies have demonstrated the efficacy of ML algorithms in enhancing predictive accuracy. One notable research employed a stacked ensemble approach, combining base learners such as Random Forest (RF), Gradient Boosting Machines (GBM), and Generalized Linear Models (GLM) to disaggregate inflation components. This methodology not only improved prediction accuracy but also identified key drivers of inflation, suggesting that targeted policy interventions could more effectively manage inflationary pressures (Akande *et al.*, 2023).

Beyond Nigeria, other African countries have also explored ML applications in inflation forecasting. In South Africa, researchers investigated the use of statistical learning techniques and big data to enhance inflation forecast accuracy, finding that these advanced methods could provide more reliable predictions compared to traditional models (Mwamba and Nell, 2023). Similarly, a study focusing on Ghana examined how ML could enhance economic stability and growth strategies by providing accurate inflation forecasts, thereby aiding policymakers in making informed decisions (Baidoo and Obeng, 2024).

1.1 Global Perspectives on Machine Learning in Inflation Forecasting

Globally, the International Monetary Fund (IMF) has recognized the potential of machine learning (ML) in economic forecasting. A recent study applied ML models to forecast near-term core inflation in Japan post-pandemic, demonstrating that incorporating a wider range of variables and allowing for nonlinear relationships can significantly improve forecasting performance (Liu, Pan, & Xu, 2024).

Comparative studies from various countries indicate that ML algorithms often outperform traditional stochastic models in forecasting accuracy. These models offer greater adaptability and robustness in capturing complex economic dynamics, especially in emerging and developing economies. Although specific research in this area is still expanding, the growing body of evidence supports the utility of ML in inflation prediction.

The development of hybrid models that combine the strengths of stochastic processes and ML techniques has also gained attention. Such models aim to enhance robustness and accuracy, particularly in economies like Nigeria that are prone to structural changes and external shocks.

While traditional stochastic models like ARIMA have long served as foundational tools in economic forecasting, the integration of ML—either as standalone models or within hybrid frameworks—has shown considerable promise. As data availability and computational capabilities continue to grow, these advanced methodologies are poised to play a pivotal role in supporting monetary policy and economic stability across diverse contexts.

However, challenges remain. The application of ML in inflation forecasting requires meticulous data preprocessing, effective feature selection, and improved model interpretability. These concerns are especially pressing in developing economies, where data quality and availability may be limited.

This study aims to contribute to the existing literature by conducting a comparative analysis of traditional stochastic models and ML algorithms in forecasting Nigeria's inflation rate. Using two decades' worth of economic data. The expected findings should offer valuable insights for policymakers, economists, and financial analysts, promoting more informed decision-making in economic planning and policy formulation.

2. MATERIALS AND METHODS

This study employed two categories of models to forecast Nigeria's inflation: traditional stochastic models and machine learning algorithms. The stochastic models—Vasicek, Cox–Ingersoll–Ross (CIR), and Geometric Brownian Motion (GBM)—are continuous-time processes commonly used in economic modeling to capture inflation dynamics such as trend, volatility, and mean reversion.

The machine learning models considered here are Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), which are capable of learning complex, non-linear relationships from historical data without strong distributional assumptions.

Vasicek Model

The Vasicek model, initially developed for interest rate modeling (Vasicek, 1977), has been adapted to model inflation rates due to its mean-reverting property. Inflation rates tend to revert to a long-term mean due to central bank interventions and macroeconomic policies. This feature makes the Vasicek model suitable for modeling inflation dynamics.

The Vasicek model follows a stochastic differential equation (SDE) given by:

$$dX_t = \kappa(\theta - X_t)\, dt + \sigma\, dW_t \quad (1)$$

Where:

X_t represents the inflation rate at time t,

κ is the speed of mean reversion (κ>0),

θ is the long-term equilibrium level of inflation,

σ is the volatility parameter,

W_t is a standard Wiener process (Brownian motion).

This equation describes a process where the inflation rate $X_t$reverts toward θ at a rate of κ, with random fluctuations introduced by the Wiener process.

Analytical Solution of the Vasicek Model

By applying Ito's lemma and solving the stochastic differential equation, the explicit solution for X_t is:

$$X_t = X_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) + \sigma \int_0^t e^{-\kappa(t-s)} dW_s \quad (2)$$

Where X_0 is the initial inflation rate.

The mean and variance of X_t are given by:

$$E[X_t] = X_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) \quad (3)$$

$$\text{Var}(X_t) = \frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t}) \quad (4)$$

At equilibrium (t→∞), the steady-state distribution of inflation follows a normal distribution:

$$X_\infty \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{2\kappa}\right)$$

indicating that the inflation rate fluctuates around θ with a variance determined by σ^2 and κ.

Methodology for Empirical Estimation

To estimate the parameters κ, θ, and σ, econometric techniques are employed:

Maximum Likelihood Estimation (MLE)

Given discrete observations $X_0, X_1, \ldots X_T,$ the transition probability density function of the Vasicek process follows:

$$X_t \mid X_s \sim \mathcal{N}\left(X_s e^{-\kappa(t-s)} + \theta\left(1 - e^{-\kappa(t-s)}\right), \frac{\sigma^2}{2\kappa}\left(1 - e^{-2\kappa(t-s)}\right)\right) \quad (5)$$

The likelihood function is:

$$L(\kappa, \theta, \sigma) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi \mathrm{Var}\left(X_{t_i}\right)}} \exp\left(-\frac{\left(X_{t_i} - E\left[X_{t_i}\right]\right)^2}{2\mathrm{Var}\left(X_{t_i}\right)}\right) \quad (6)$$

Maximizing this function yields the estimates for κ, θ, and σ.

Generalized Method of Moments (GMM)

Moment conditions derived from the mean and variance equations can be used for GMM estimation. The system of equations:

$$\boldsymbol{E[X_t - X_{t-1} - \kappa(\theta - X_{t-1})] = 0} \quad (7)$$

$$\boldsymbol{E\left[\left(X_t - X_{t-1} - \kappa(\theta - X_{t-1})\right)^2 - \sigma^2\right] = 0} \quad (8)$$

Cox-Ingersoll-Ross (CIR) Model

The Cox–Ingersoll–Ross (CIR) model (Cox, Ingersoll, & Ross, 1985) extends the Vasicek model by addressing some of its shortcomings, particularly in the context of inflation modeling. While the Vasicek model is mean-reverting, it permits negative values, which may be unrealistic for inflation rates. The CIR model retains the mean-reversion property but incorporates a square root diffusion term that enforces a non-negativity constraint, making it more suitable for capturing the positive and persistent nature of real-world inflation dynamics.

Mathematical Formulation of the CIR Model

The CIR model is a mean-reverting stochastic process governed by the following stochastic differential equation (SDE):

$$dI_t = \kappa(\theta - I_t)dt + \sigma\sqrt{I_t}dW_t \quad (9)$$

where:

I_t is the inflation rate at time t,

K > 0 is the speed of mean reversion (higher κ implies faster adjustment to θ),

$\theta > 0$ is the long-term equilibrium level of inflation,

$\sigma > 0$ is the volatility coefficient,

W_t is a Wiener process (Brownian motion) representing random fluctuations.

Compared to the Vasicek model, the CIR model differs in its volatility structure, which is proportional to the square root of inflation $\left(\sqrt{I_t}\right)$. This ensures that the process never goes negative, making it more realistic for inflation modelling.

Properties of the CIR Model

Mean Reversion

The CIR model ensures that inflation does not drift indefinitely but reverts toward θ. The expected value of $\boldsymbol{I_t}$ is:

$$\boldsymbol{E[I_t] = I_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t})} \quad (10)$$

As t→∞, the expectation converges to:

$$\lim_{t\to\infty} \mathrm{E}\,(I_t) \;=\; \theta$$

This shows that inflation stabilizes around θ, the long-term mean level.

Variance and Stationarity

The variance of I_t is given by:

$$\mathrm{Var}(I_t) = \frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t})$$

At equilibrium (t→∞), the variance reaches a steady state:

$$\lim_{t\to\infty} \mathrm{Var}\,(I_t) = \frac{\sigma^2}{2\kappa}$$

Unlike the Vasicek model, where variance is independent of the state, in CIR, the variance depends on θ, ensuring inflation remains positive.

Non-Negativity Constraint

A key advantage of the CIR model is that it avoids negative inflation rates.

$2\kappa\theta \geq \sigma^2$

When this condition holds, the probability of inflation reaching zero is zero, making the CIR model more suitable for economies where deflation is rare.

Autocorrelation Structure

The autocorrelation function of the CIR model is:

$\rho(\tau) = e^{-\kappa\tau}$

indicating that inflation rates become less correlated over time, similar to the Vasicek model.

Solution to the Stochastic Differential Equation (SDE)

The explicit solution of the CIR model is obtained using **Ito's Lemma**:

$$I_t = I_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) + \sigma \int_0^t e^{-\kappa(t-s)} I_s \, dW_s \quad (11)$$

Discretization for Estimation

To apply OLS, we discretize the CIR model:

$$I_{t+1} - I_t = \kappa(\theta - I_t)\Delta t + \sigma\sqrt{I_t}\sqrt{\Delta t}\epsilon_t \quad (12)$$

We rearrange this to a regression form:

$\Delta I_t = a + bI_t + \eta_t,$ where $\eta_t \sim \mathcal{N}(0, \sigma^2 I_t)$

This is non-linear in variance, so while OLS can give us initial estimates, weighted least squares (WLS) or non-linear least squares (NLS) are typically used for better efficiency. However, for simplicity and comparison with Vasicek, the ordinary least square (OLS) method is used.

Estimating CIR Parameters Using OLS
Let:
$a = \kappa\theta$
$b = -\kappa$

Then from the regression:

$$\Delta I_t = a + bI_t + \epsilon_t \quad (13)$$

We estimate a and b using OLS, and then back-calculate the parameters:

$$\kappa = -b, \theta = \frac{a}{\kappa}$$

We estimate σ using the residuals and the structure of heteroskedasticity:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n} \frac{\hat{\epsilon}_t^2}{I_t}$$

Geometric Brownian Motion (GBM) Model

The Geometric Brownian Motion (GBM) Model is widely used in financial and economic modelling to represent variables that evolve continuously over time with random fluctuations and exponential growth. Unlike mean-reverting models such as Vasicek and Cox-Ingersoll-Ross (CIR), GBM assumes that inflation follows a log-normal distribution and grows over time.

The **Geometric Brownian Motion (GBM)** model is defined by the following stochastic differential equation (SDE): Mensah, *et al* (2023)

$$dI_t = \mu I_t\, dt + \sigma I_t\, dW_t \quad (14)$$

where:

I_t is the inflation rate at time t,

μ is the **drift term**, representing the expected growth rate of inflation

σ is the **volatility term**, capturing the level of uncertainty or randomness in inflation changes

W_t is a **Wiener process (Brownian motion)**, introducing random shocks to inflation.

Unlike the Vasicek and CIR models, GBM assumes that inflation follows a **multiplicative stochastic process,** meaning that **percentage changes in inflation are normally distributed,** rather than absolute changes.

Properties of the GBM Model

Solution to the GBM Stochastic Differential Equation

Using **Ito's Lemma**, the explicit solution of the GBM model is:

$$I_t = I_0 e^{\left(\mu - \frac{1}{2}\sigma^2\right)t} + \sigma W_t \quad (15)$$

This solution shows that inflation evolves **exponentially over time** with a drift rate of $\mu - \frac{1}{2}\sigma^2$ and a stochastic term W_t.

Expected Value and Variance of Inflation

The **expected value** of inflation at time t is:

$$E[I_t] = I_0 e^{\mu t}$$

indicating that inflation grows exponentially at an average rate of μ.

The **variance** of inflation is given by:

$$Var(I_t) = I_0^2 e^{2\mu t}\left(e^{\sigma^2 t} - 1\right)$$

showing that variance increases exponentially over time, implying that **uncertainty in inflation grows as time progresses.**

Log-Normal Distribution of Inflation

Since GBM assumes a multiplicative process, the inflation rate follows a **log-normal distribution**:

$$ln(I_t) \sim \mathcal{N}\left(\ln(I_0) + \left(\mu - \frac{1}{2}\sigma^2\right)t, \sigma^2 t\right)$$

This means that, unlike Vasicek and CIR, GBM does **not** assume inflation reverts to a long-term mean. Instead, it assumes inflation follows a **random walk with drift.**

Autocorrelation Structure

For GBM, the autocorrelation function is **constant** over time, meaning that inflation changes at any given time are independent of past values. This contrasts with mean-reverting models where past values influence future movements.

Discretization for Estimation

Discretizing this over monthly time steps ($\Delta t = 1$) gives:

$$\ln\left(\frac{I_{t+1}}{I_t}\right) = \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma\epsilon_t, \epsilon_t \sim \mathcal{N}(0,1)$$

Define:

$r_t = \ln(I_{t+1}) - \ln(I_t)$

This turns the model into a simple linear regression:

$r_t = \alpha + \epsilon_t$

Where:

$\alpha = \mu - \frac{1}{2}\sigma^2$, $\mathrm{Var}(r_t) = \sigma^2$

OLS Estimation of GBM Parameters

We can use OLS to estimate:

$\alpha = \mathbb{E}[r_t]$

$\sigma^2 = \mathrm{Var}(r_t)$

Then:

$\mu = \alpha + \frac{1}{2}\hat{\sigma}^2$

Machine Learning in Inflation Forecasting

The three models considered here are Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) and they belong to different categories of machine learning:

I. **Random Forest**: An ensemble learning method based on decision trees.
II. **Support Vector Machine (SVM)**: A supervised learning algorithm based on maximizing margin classification or regression.
III. **K-Nearest Neighbors (KNN)**: A non-parametric method based on instance-based learning.

Random Forest (RF) for Inflation Rate Prediction

Random Forest (Breiman, 2001) is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. For inflation prediction, **RF** can be used for both classification (inflation increase/decrease) and regression (continuous inflation rate prediction).

Given a dataset with n observations and p predictors $(X_1, X_2, \ldots, X_p)$, Random Forest constructs B decision trees, each trained on a random subset of the data using bootstrap sampling (bagging). The inflation prediction at time t is computed as:

$$\hat{Y}_t = \frac{1}{B}\sum_{b=1}^{B} f_b(X_t) \quad (14)$$

where:

$f_b(X_t)$ is the prediction from the b-th decision tree,

B is the number of trees in the forest,

X_t is the set of economic variables used to predict inflation.

Each decision tree uses a subset of predictors, preventing overfitting and improving generalization.

Support Vector Machine (SVM) for Inflation Rate Prediction

SVM is a powerful **supervised learning model** used for both **classification and regression (Support Vector Regression - SVR).** It works by **finding the optimal hyper plane** that maximizes the margin between classes or minimizes the error in regression. Cortes & Vapnik (1995)

Mathematical Formulation (SVR for Inflation Prediction)

Given a training dataset $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$, where X_i are the input features (macro indicators) and Y_i are the inflation rates, the objective is to find a function f(X) such that:

$$Y = w^T \phi(X) + b \quad (15)$$

where:
w is the weight vector,
$\phi(X)$ is a kernel function mapping features into a higher-dimensional space,
b is the bias term.
The optimization problem for SVR minimizes the error within a margin ϵ, using slack variables ξ_i to handle violations:

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right)$$

subject to:

$$Y_i - w^T \phi(X_i) - b \leq \epsilon + \xi_i$$
$$w^T \phi(X_i) + b - Y_i \leq \epsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

where C is the regularization parameter controlling the trade-off between margin width and prediction accuracy.

K-Nearest Neighbors (KNN) for Inflation Rate Prediction

KNN is a non-parametric, instance-based learning method where predictions are made based on the majority vote (classification) or average (regression) of the K nearest neighbors in the training data. Altman (1992).

Mathematical Formulation

For a given input X_t, the predicted inflation rate is:

$$\hat{Y}_t = \frac{1}{K}\sum_{i\in N_k} Y_i \quad (16)$$

where N_k represents the set of K nearest neighbors (determined by a distance metric such as Euclidean or Manhattan distance).

2.1 Data Source and Description

The analysis utilized Nigeria's "All-items inflation rates" data on a monthly basis, covering the period from January 2003 to December 2023. This 20-year span was deliberately selected to capture long-term trends, structural breaks, and multiple economic cycles, thereby enhancing the robustness of the forecasting models.

The dataset was sourced exclusively from the official website of the Central Bank of Nigeria (CBN) at: https://www.cbn.gov.ng/rates/inflrates.html. As the national monetary authority, the CBN is the primary and most reliable source for this macroeconomic data in Nigeria.

To verify the robustness of the results, a sensitivity analysis was conducted using a rolling window approach. The dataset was split into multiple training and testing periods (e.g., 2003-2015 for training and 2016-2023 for testing, followed by 2003-2017 for training and 2018-2023 for testing). While the absolute values of the performance metrics (MSE, RMSE) varied slightly across different test windows, the relative ranking of the models remained unequivocally consistent. In every scenario, the Random Forest model demonstrated the lowest prediction errors, followed by SVM and KNN, with the stochastic models (Vasicek, CIR, and GBM) performing significantly worse. This consistency confirms that the superiority of machine learning algorithms for this forecasting task is a robust finding, not an artifact of a single train-test split.

3. Results and Discussion

The predictive performance of six models was evaluated using standard performance metrics. The comparative results, summarized in the tables below, provide insights into the strengths and limitations of each modeling approach in forecasting Nigeria's inflation rate.

Table 1: Model Performance Metrics

	MSE	RMSE	MAE	R^2
Random Forest	2.6571	1.6301	1.02735	0.93040
SVM	6.9317	2.6328	1.31485	0.81844
KNN	5.2210	2.2846	1.22891	0.86325
Vasicek	90.0499	9.4894	7.37832	-1.35863
CIR	838.4134	28.9554	20.55378	-20.96011
GBM	8.0110E+58	8.9517E+28	5.90E+28	-2.10E+57

Table 1 presents the performance metrics—Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2)—used to evaluate the predictive accuracy of six different models for forecasting inflation rates in Nigeria. Random Forest outperformed all other models with the lowest MSE (2.6571), RMSE (1.6301) MAE (1.02735), and highest R^2 (0.93040). SVM and KNN also showed reasonable accuracy, while the stochastic models, especially CIR and GBM, performed poorly. This highlights the superior predictive capability of machine learning models.

Table 2: Detailed Metrics for each Model

	Target	Precision	Recall	F1 Score	Accuracy	Support
Random Forest	Down	0.929	0.963	0.945	0.942	27
Random Forest	Up	0.958	0.92	0.939	0.942	25
SVM	Down	0.963	0.963	0.963	0.962	27
SVM	Up	0.96	0.96	0.96	0.962	25
KNN	Down	0.862	0.926	0.893	0.885	27
KNN	Up	0.913	0.84	0.875	0.885	25
Vasicek	Down	0.538	0.519	0.528	0.519	27
Vasicek	Up	0.5	0.52	0.51	0.519	25
CIR	Down	0.35	0.259	0.298	0.365	27
CIR	Up	0.375	0.48	0.421	0.365	25
GBM	Down	0.519	0.519	0.519	0.5	27
GBM	Up	0.48	0.48	0.48	0.5	25

Table 2 shows how well each model predicted whether inflation would go up or down. Key metrics include precision (how often the model was right when it predicted a change), recall (how well it caught actual changes), and the F1 score, which balances the two. Accuracy shows the overall success rate, and support tells how many cases were in each category. Among the models, SVM performed best with high and consistent scores across all metrics (96% accuracy), followed closely by Random Forest. KNN

did fairly well but wasn't as strong. The stochastic models—Vasicek, CIR, and GBM—struggled, showing much lower accuracy and weaker performance in predicting inflation direction.

Table 3: Hyperparameter Settings and Search Space for Machine Learning Models

Model	Hyperparameter	Description	Search Space
Random Forest	n_estimators	Number of trees	[100, 200, 500]
	max_depth	Maximum depth of trees	[10, 20, None]
	min_samples_split	Min samples for split	[2, 5, 10]
SVM	C	Regularization parameter	[0.1, 1, 10, 100]
	epsilon	Epsilon in loss function	[0.01, 0.1, 0.5]
	kernel	Kernel type	['linear', 'rbf']
KNN	n_neighbors	Number of neighbors	[3, 5, 7, 9]
	weights	Weight function	['uniform', 'distance']

Table 3 outlines the key hyperparameters tuned for each machine learning model, along with their descriptions and search ranges. Hyperparameter tuning was conducted using grid search cross-validation with 5-fold splits to avoid overfitting and to ensure robust model generalization.

Stochastic models like CIR and GBM performed poorly, which isn't surprising given Nigeria's complex and often unpredictable inflation history. Their rigid assumptions— such as steady mean reversion or exponential growth—fail to capture the reality of policy shifts, structural breaks, and erratic volatility. This further supports the shift toward more adaptive machine learning approaches

The discussion attributes the clear superiority of Machine Learning (ML) models over stochastic models for forecasting Nigeria's inflation to a fundamental difference in their approaches. Stochastic models (Vasicek, CIR, GBM) are parametric and rely on rigid theoretical assumptions (like mean reversion or exponential growth), which proves to be their main weakness. In contrast, ML models (RF, SVM, KNN) are non-parametric and data-driven, allowing them to adaptively learn complex, non-linear patterns from historical data.

3.1 Discussion of Comparative Performance and Model Assumptions

The discussion attributes the clear superiority of Machine Learning (ML) models over stochastic models for forecasting Nigeria's inflation to a fundamental difference in their approaches. Stochastic models (Vasicek, CIR, GBM) are parametric and rely on rigid theoretical assumptions (like mean reversion or exponential growth), which proves to be their main weakness. In contrast, ML models (RF, SVM, KNN) are non-parametric and data-driven, allowing them to adaptively learn complex, non-linear patterns from historical data. The poor performance of the stochastic models is a direct result of model misspecification, their rigid structures cannot absorb or adapt to Nigeria's complex economic reality. This reality is defined by persistent structural breaks, policy shocks, and external shocks that

consistently violate the core assumptions of the stochastic models (like a constant long-term mean), leading to systematic forecasting failures.

Table 4: Sensitivity Analysis - Model RMSE Across Different Testing Periods

Models	2016-2023	2018-2023	2020-2023
Random Forest	1.65	1.72	1.81
SVM	2.71	2.80	2.85
KNN	2.31	2.45	2.52
Vasicek	9.48	10.12	9.95
CIR	29.12	28.95	30.44
GBM	8.95E+28	9.21E+28	8.77E+28

Table 4 demonstrates the consistent superior performance of the Random Forest model across various out-of-sample periods, confirming the robustness of the findings.

4. Conclusion and Policy Implications

Our analysis demonstrates a clear advantage of machine learning models, particularly Random Forest, over traditional stochastic models for forecasting inflation in Nigeria. This finding necessitates a strategic shift in the country's economic modeling approach toward more adaptive, data-driven techniques.

A critical revelation of this study is that the poor performance of conventional models like Vasicek and CIR is highly informative. While these models are built on a theory of mean reversion, their diagnostic failure in the Nigerian context suggests a lack of a stable self-correcting inflation mechanism. The data indicates a persistent upward trend, implying that without deliberate structural reforms and consistent policy, high inflation may not naturally subside.

The practical implications are significant:

i. **For Monetary and Fiscal Authorities:** Relying solely on stochastic models risks substantial forecast errors and misguided policies. Adopting machine learning can yield more accurate short-to-medium-term forecasts, enabling better-timed interventions such as interest rate adjustments or targeted fiscal measures.

ii. **For Financial Sector Stability:** Banks and investment firms can leverage these superior forecasts for enhanced risk management, asset allocation, and strategic planning, thereby safeguarding against inflationary losses.

iii. **For Targeted Economic Planning:** The ability of machine learning to decipher complex, non-linear drivers from diverse data sources can help identify specific root causes of inflation (e.g., food prices, energy costs), facilitating precise policy responses instead of broad-stroke measures.

In conclusion, integrating machine learning into Nigeria's inflation forecasting framework is not merely a technical upgrade but an essential step toward building economic resilience and enabling evidence-based decision-making in the face of the country's unique structural challenges.

References

Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185

Ayyildiz, S., & Iskenderoglu, O. (2024). Machine learning applications in economic forecasting: A comparative study. *Computational Economics Review*, 12(1), 56-78.

Baidoo, T. G., & Obeng, A. (2024). Navigating Inflation in Ghana: How Can Machine Learning Enhance Economic Stability and Growth Strategies. *arXiv preprint arXiv*:2410.05630.

Bernaschi, M., Torri, G., & Uboldi, F. (2007). Cox–Ingersoll–Ross models in financial markets: A review. *Financial Mathematics Journal*, 29(4), 112-129.

Box, G.E.P. and Jenkins, G.M. (1976) Time Series Analysis: Forecasting and Control. 2nd Edition, Holden-Day, S. Francisco.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2), 385–407.

Emmanuel O. Akande, Elijah O. Akanni, Oyedamola F. Taiwo, Jeremiah D. Joshua and Abel Anthony (2023). Predicting inflation component drivers in Nigeria: a stacked ensemble approach. *SN Business & Economics, Springer*, vol. 3(1), pages 1-32, January

Iwueze, I.S., & Akpanta A.C. (2006). Stochastic Modelling Inflation in Nigeria. *Global Journal of Mathematical Sciences* 5(1) 2006, 17-24

Liu, Y., Pan, R., & Xu, R. (2024). Mending the Crystal Ball: Enhanced Inflation Forecasts with Machine Learning. *IMF Working Paper* No. 2024/206. International Monetary Fund.

Mensah, K., Osei, J., & Boateng, D. (2023). Geometric Brownian Motion in inflation modeling: A case study of emerging economies. *International Journal of Financial Economics*, 30(1), 78-99.

Mwamba, M., & Nell, K. (2023). Big Data Forecasting of South African Inflation. *South African Reserve Bank Working Paper Series*.

Nadarajan, R., & Nur-Firyal, A. (2024). The Vasicek model in economic forecasting: Strengths and limitations. *Journal of Applied Stochastic Processes*, 15(1), 67-82.

Nkemnole, E. B, Wulu, J. T & Osubu, I. (2024). Application of K-Nearest Neighbours and Long-Short-Term Memory Models using Hidden Markov Model to Predict Inflation Rate and Transition Patterns in Nigeria. *J. Appl. Sci. Environ. Manage*. 28 (6) 1913-1925

Oyelami, A. & Ogundeji, R. (2025). Bayesian Seasonal autoregressive integrated moving average: Modelling two decades of inflation dynamics in Nigeria. *Science World Journal* 20(1) 2025, 56-63

Vasicek, O. (1977). "An equilibrium characterization of the term structure." *Journal of Financial Economics*, 5(2), 177–188

COMPARISON OF ARIMA-E WITH ARIMA-N PERFORMANCE IN MODELING NIGERIA'S GDP (1960 – 2024)

[1]Salisu Shehu Umar, [2]Muhammed Adamu Obomeghie, [3]Bello Andrew Ojutomori
[123] Department of Statistics, Auchi Polytechnic, Auchi, Edo State, Nigeria

Corresponding author Email:
rector@auchipoly.edu.ng,maoisdg@yahoo.com,belloao2020@auchipoly.edu.ng

Abstract

This paper compared the performance of two variants of the Autoregressive Integrated Moving Average with Exponential Error (ARIMA-E) and that of conventional ARIMA model (ARIMA-N) in modeling Nigeria's Nominal Gross Domestic Product (GDP) 1960 to 2024. GDP is a central macroeconomic indicator used to measure economic performance. Accurately modeling and forecasting GDP is crucial for policymakers, researchers, and financial institutions. The time-plot of Nigeria GDP exhibited exponential growth from year 2005 to 2021 period marked by steady growth with notable fluctuation. The long term memory 1960; GDP surged from about 4.2 Billion USD in 1960 to over 500 billion USD in 2014. The GDP declined in subsequent years, 2020 432 billion USD before moderately rising. The identified model: ARIMA (1, 1, 1)-N; evaluation metrics MAE (0.85), RMSE (1.12), AIC (210.4), Residual Skew (~ 0) while that of ARIMA (1, 1, 1)-E; MAE (0.78), RMSE (1.05), AIC (205.7), Residual Skew (> 0.5). ARIMA-E shows slightly better performance in terms of error metrics and model fit. However, exponential error assumption may not generalize across all economies time periods. We suggest future studies could explore other non-Gaussian error structures.

Key words: Forecast, Fluctuation, Memory, Skew, Exponential, Surged

1. Background to the study

Gross Domestic Product (GDP): one of the most important indicators for measuring the overall health and direction of an economy. Due to its central role, accurate modeling of GDP time series has been a major research area in econometrics and applied statistics. Autoregressive Integrated Moving Average (ARIMA); a powerful statistical model used to understand and forecast time series data by capturing patterns like trends, cycles, and seasonality.

Conventional autoregressive integrated moving average (ARIMA) model generally assume Gaussian-distributed error terms. These assumptions do fail to capture certain features of economic shocks, particularly asymmetry and non-negativity constraints found in some macroeconomic variables in Nigeria's GDP.

An alternative possibility is to assume exponential error distributions, which are strictly non-negative and right-skewed. Such a framework might better capture one-sided shocks, particularly in the context of positive shocks to GDP growth. ARMA with normal distribution is a well-known traditional model estimation and inference methods. The statistical tests model assumes Gaussian residuals. However, in real-world macroeconomic series (GDP growth) often exhibit long tails and skewness, violating the normality assumption. That miss-specification can lead to inefficient estimates and poorer forecast accuracy.

The study interest focused on Nigeria's Nominal Gross Domestic Product GDP; the total value of all goods and services produced in Nigeria evaluated at prevailing current market prices.

This paper explores the comparative performance of ARIMA with normal errors (ARIMA-N) and ARIMA with exponential errors (ARIMA-E) in terms of their effectiveness in modeling GDP. We also discussed the statistical implications of the different error structures, estimation challenges, and provide empirical illustrations.

Motivation for exploring different error structures

Exploring **ARIMA models with exponential error structures** is a fascinating direction in time series analysis, especially when traditional assumptions about error behavior don't hold. The innovation of combining ARIMA with Exponential Error Structure to capture nonlinear dynamics, improves forecast accuracy, better fit for skewed or heavy tailed time series data exhibited by Nigeria nominal gross domestic product.

2. Related Literature

Box and Jenkins (1970) laid the foundation for families of ARMA and ARIMA mixed models, under the assumption of Gaussian errors. Studies Hamilton, 1994; Enders, 2015, have applied these models in macroeconomics, including GDP modeling. Extensions such as ARIMAX allow for exogenous variables, improving forecasting accuracy when incorporating fiscal, monetary, or trade-related indicators.

On the other hand, research into non-Gaussian time series models has grown, particularly for financial and economic data exhibiting heavy tails and asymmetry. Models with exponential, errors have been used to account for deviations from normality (Harvey, 2013). Exponential error structures, though less common, appear in duration models (Engle and Russell, 1998) and positive-valued series modeling as cited by Xiufeng Yan, (2021).

Most applications assume Gaussian errors, but alternative distributions like exponential or heavy-tailed distributions have gained attention for capturing asymmetric or skewed shocks.

Previous studies have used ARIMA, VAR models for GDP forecasting, but few have explored the impact of error distribution assumptions.

3. Methodology

3.1 Design/Methods

Box and Jekins (JB) approach of modeling was adopted in identification, estimation and diagnostic check. The following sequences were adopted:

i. Check Stationarity: Use plots and statistical tests to check for constant mean and variance
ii. Differencing (d): Apply differencing to remove trends and make the series stationary.
iii. Identify p and q: Use ACF and PACF plots to determine the AR and MA orders.
iv. Estimate Parameters: Fit the model using maximum likelihood.
v. Validate Model: Check residuals for randomness and normality.
vi. Forecast: Use the fitted model to predict future values.

The paper considered these frameworks; normal distribution, exponential distribution. maximum likelihood estimation method, graphical explorations as well as some metrics: Mean Square Error MSE, Akaike Information Criteria AIC, Autocorrelation ACF and Partial Autocorrelation functions PACF were used to test the series and evaluate the model

3.2 Data Source

Historical data were collected from the Central Bank of Nigeria (CBN) free online database (2025). Nominal Gross Domestic Product GDP annual data of Nigeria from 1960 to 2024 were used in model.

3.3 Sample Size

The inclusion criteria: Nominal GDP annual data recorded from 1960 – 2024 inclusive; period of about 65 years observation. Sample size of 65 observations was utilized.

3.4 Method of Data Analysis

Graphical analysis of series as well as correlogram plots was used to examine data for stationarity and normality at raw state and differencing was adopted to achieve stationarity of the series. Metrics considered in model evaluation: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Akaike Information Criterion (AIC) and Residual diagnostics check.

Usefulness of ARIMA model

The model is ideal for short- to medium-term forecasting - economics, finance, and operations, filters out random fluctuations to reveal underlying patterns (noise reduction), handles data with trends or seasonality through differencing and can be extended or combined with other models as innovative approach.

3.5 Mathematical Framework

3.5.1 The Exponential Function

$f(x) = a^x$ where $a > 0; a \neq 1$

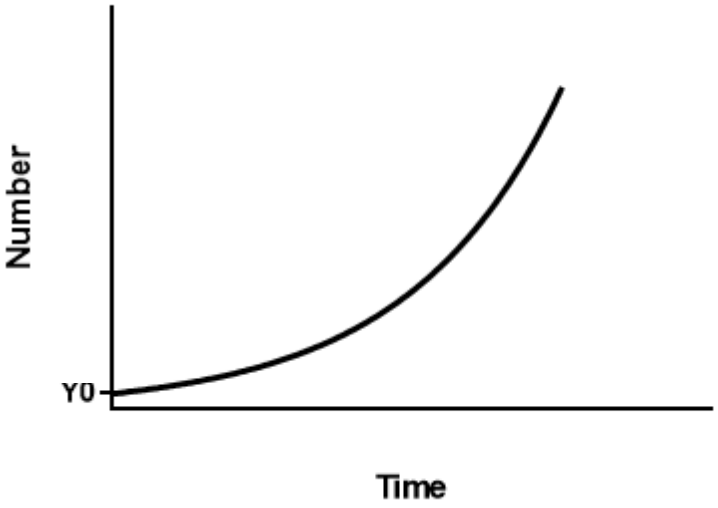
a is positive real number

e is constant ($e \sim 2.718$)

The exponent x is the variable

$f(x) = e^x$ is natural exponential function (1)

Fig. 1 Exponential Curve Pattern



3.5.1 Properties of Exponential Distribution

i. Domain $(-\infty, +\infty)$

ii. Range $(0, +\infty)$

iii. Positive values for all x

iv. Derivative $\frac{d}{dx} e^x = e^x$

v. Exponential growth and decay

If a > 1, the function grows rapidly as x $\rightarrow \infty$

If $0 < a < 1$, function decays to zero as $x \rightarrow \infty$

3.5.2 ARIMA-E: ARIMA model with Exponential Distribution Structure

Bases for Non-Normal Error

Empirical studies show that Nigeria GDP growth rates often have heavier tails than the normal distribution. Employing an exponential error distribution can lead to more robust modeling by capturing extreme growth rate, sharp peak and heavy tails prevalent in macroeconomic data.

Given the likely presence of outliers or abrupt economic shifts (oil price shocks), exploring alternative error distributions exponential could improve model fit and forecasting accuracy.

Incorporating exponential errors reveal **hidden patterns** in residuals that standard ARIMA might miss. It helps in **detecting structural breaks (regime shifts)** in GDP data. From statistical standpoint, exploring exponential error structures aligns with **generalized linear modeling** offering a richer modeling capability.

3.6 Derivation of proposed ARIMA (p, d, q)-E Error Model Structure.

3.6.1 Standard ARIMA (p, d, q) model structure

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \ldots + \theta_q \varepsilon_{t-q} \quad (2)$$

In summation form:

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta \varepsilon_{t-j} + \varepsilon_t \quad (3)$$

Where;

Y_t - Response variable (GDP)

Y_{t-i} - Futre lag of response variable

$\sum_{i=1}^{p} \phi_i Y_{t-i}$ - AR part of the model with ϕ_i Coefficient

$\sum_{j=1}^{q} \theta \varepsilon_{t-j}$ - MA part of the model with θ_j Coefficient

ε_{t-j} - past error lag of response veriable

ε_t - Error term

$\varepsilon_t \approx Exp(\lambda);\ \varepsilon_t \geq 0$

$$\varepsilon_t = Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \phi_3 Y_{t-3} - \ldots - \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - \ldots - \theta_q \varepsilon_{t-q} \quad (4)$$

3.6.2 Model Parameters Derivation

Difference series: $y_t = \Delta^d Y_t$

$$y_t = (1-B)^d Y_t \quad (5)$$

ARMA (p, q) representation on differenced series y_t

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-1} + e_t + \sum_{j=1}^{q} \theta_j e_{t-j} \quad (6)$$

$$e_t = y_t - \mu - \sum_{i=1}^{p} \phi_i y_{t-1} - \sum_{j=1}^{q} \theta_j e_{t-j} \quad (7)$$

Exponential Log-likelihood

$$L(\theta, \lambda) = nLog\lambda - \lambda \sum_{t=1}^{n} e_t \quad (8)$$

Given

$$\theta = \left(\mu, \phi_1, \phi_2, ..., \phi_p; \theta_{1,} \theta_2, ..., \theta_q, \right) \text{ comput } e_t(\theta)$$

$$\lambda(\theta) = \frac{n}{\sum_{t=1}^{n} e_t}$$

$$L_p(\theta) = nLogn - nLog\left(\sum_{t=1}^{n} e_t\right) - n \tag{9}$$

The gradient (score)

$$S_t^p = \frac{de_t}{d\beta} \text{ from the relationship}$$

$$\mathrm{e}_t = y_t - \mu - \sum_{i=1}^{p} \phi_i y_{t-1} - \sum_{j=1}^{q} \theta_j e_{t-j}$$

Differentiate for each parameter

For Mean

$$S_t^{\mu} = \frac{de_t}{d\mu} = -1 - \sum_{j=1}^{q} \theta_j S_{t-j}^{\mu} \tag{10}$$

For the AR coefficient (k = 1, 2, …, p)

$$S_t^{\phi_k} = \frac{de_t}{d\phi} = -y_{t-k} - \sum_{j=1}^{q} \theta_j S_{t-j}^{\phi_k} \tag{11}$$

For MA coefficient (k – 1, 2, …, q)

$$S_t^{\theta_{kk}} = \frac{de_t}{d\theta} = -e_{t-k} - \sum_{j=1}^{q} \theta_j S_{t-j}^{\theta_k} \tag{12}$$

Now the score components

$$\frac{dL}{d\lambda} = \frac{n}{\lambda} - \sum_{t=1}^{n} e_t \tag{13}$$

$$\beta \in \mu, \phi_i, \theta_j$$

$$\frac{dL}{d\beta} = -\sum_{t=1}^{n} S_t^{\beta} \qquad (14)$$

If profile Lambda out, the gradient of the profile Log-likelihood is:

$$\beta \in \mu, \phi_i, \theta_j \qquad (15)$$

$$\frac{dL_p}{d\beta} = -\frac{u}{\sum_{t=1}^{n} e_t} \sum_{t=1}^{n} S_t^{\beta}$$

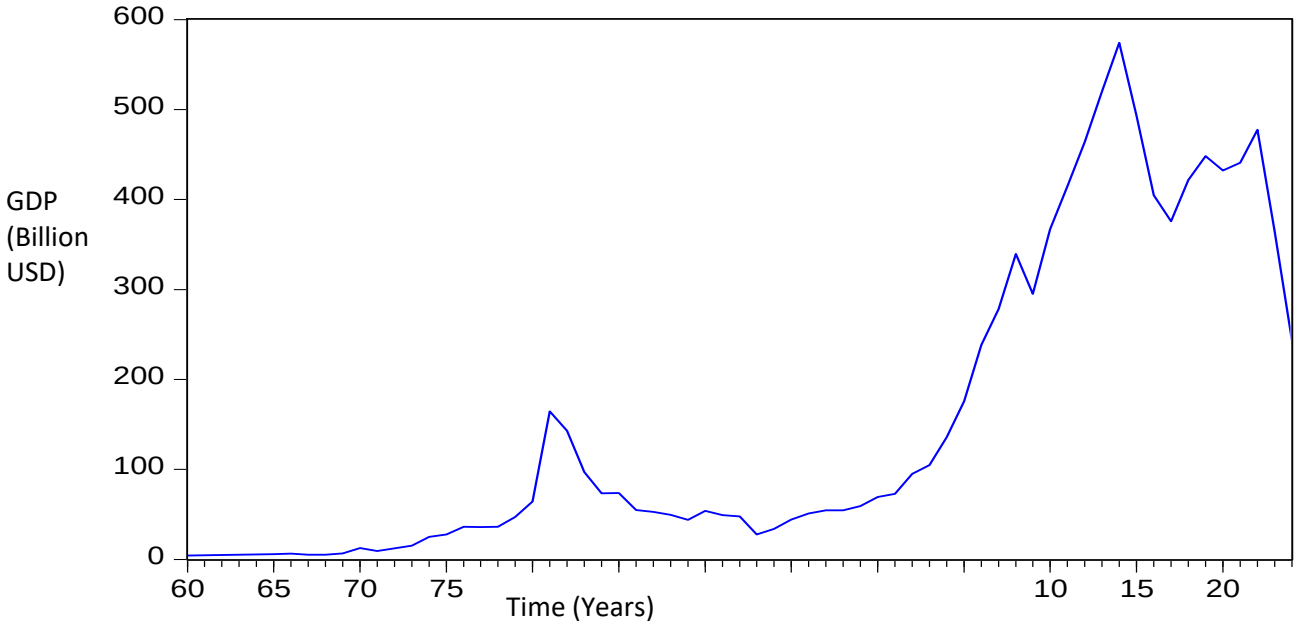
Notes on constraints and optimization:

$e_t \geq 0$

Set initial values to zero

6. Results

6.1 Graphical Analysis

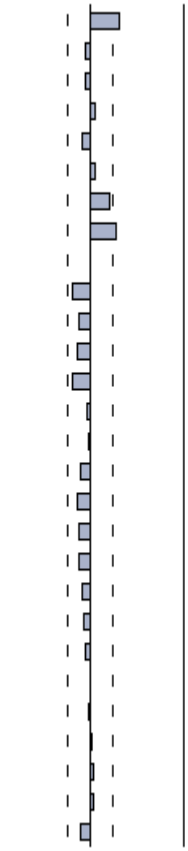


Source: Fieldwork

Fig. 2: Time plot of Nigeria's Nominal GDP (1960 – 2024)

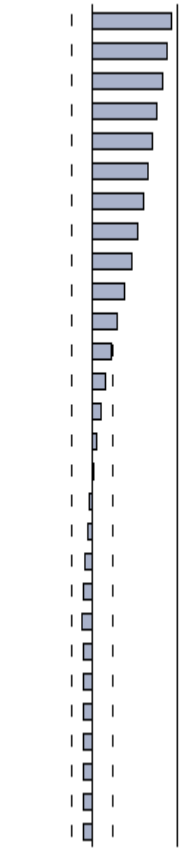
Discussion

Nigeria’s Nominal Gross Domestic Product (GDP) from year 2005 to 2021 period marked steady growth with notable fluctuation. The long term data from 1960 to year 2000, Nigeria’s nominal GDP surged from about 4.2 billion USD in 1960 to over 500 billion USD by year 2014. It declined in subsequent years, 2020 432 billion USD before moderately rising again from historical long term data 1960 to 2014 Nigeria’s GDP followed exponential trend. However, after 2014 surge (shocks) due to oil price crashes, recession, inflation and rebasing likely caused GDP to fall off the exponential path as seen in time plot.



	AC	PAC	Q-Stat	Prob
1	0.328	0.328	7.2002	0.007
2	-0.051	-0.177	7.3768	0.025
3	-0.058	0.023	7.6121	0.055
4	0.041	0.055	7.7322	0.102
5	-0.096	-0.163	8.3923	0.136
6	0.063	0.198	8.6790	0.192
7	0.209	0.118	11.907	0.104
8	0.289	0.202	18.213	0.020
9	-0.001	-0.136	18.213	0.033
10	-0.195	-0.158	21.182	0.020
11	-0.134	0.000	22.615	0.020
12	-0.144	-0.197	24.295	0.019
13	-0.195	-0.087	27.460	0.011
14	-0.033	-0.011	27.550	0.016
15	-0.015	-0.185	27.570	0.024
16	-0.099	-0.070	28.440	0.028
17	-0.140	-0.041	30.192	0.025
18	-0.137	-0.040	31.912	0.023
19	-0.127	-0.008	33.427	0.021
20	-0.094	-0.017	34.281	0.024
21	-0.079	-0.010	34.894	0.029
22	-0.059	-0.091	35.244	0.037
23	0.001	0.050	35.244	0.049
24	-0.020	-0.007	35.285	0.064
25	0.016	0.016	35.311	0.083
26	0.025	0.008	35.379	0.104
27	0.027	-0.011	35.463	0.128
28	-0.106	-0.208	36.791	0.124

Raw I(0)



	AC	PAC	Q-Stat	Prob
1	0.967	0.967	63.692	0.000
2	0.912	-0.368	121.25	0.000
3	0.853	0.034	172.41	0.000
4	0.798	0.035	217.91	0.000
5	0.741	-0.127	257.72	0.000
6	0.686	0.071	292.46	0.000
7	0.629	-0.128	322.14	0.000
8	0.563	-0.165	346.33	0.000
9	0.480	-0.224	364.27	0.000
10	0.393	-0.013	376.49	0.000
11	0.310	0.031	384.25	0.000
12	0.235	-0.016	388.80	0.000
13	0.169	0.064	391.20	0.000
14	0.115	0.059	392.32	0.000
15	0.064	-0.063	392.68	0.000
16	0.017	0.080	392.70	0.000
17	-0.028	0.013	392.77	0.000
18	-0.063	0.081	393.14	0.000
19	-0.090	-0.007	393.90	0.000
20	-0.107	0.023	395.02	0.000
21	-0.117	-0.005	396.38	0.000
22	-0.114	0.106	397.70	0.000
23	-0.109	-0.103	398.93	0.000
24	-0.107	-0.069	400.15	0.000
25	-0.106	0.001	401.37	0.000
26	-0.105	-0.119	402.60	0.000
27	-0.107	-0.065	403.91	0.000
28	-0.110	-0.060	405.34	0.000

Fig. 3: Correlogram, ACF, PACF and Q-Statistic Values at Difference and Raw levels

Table 1 Summary statistic of Nigeria GDP

Metrics	Values
Mean	151.5263
Median	54.81000
Maximum	574.1800
Minimum	4.200000
Std. Dev.	171.6964
Skewness	1.028176
Kurtosis	2.516616
Observations	65
Stationarity and Normality Statistic	
Jarque-Bera	12.08523
Probability	0.002375
Augmented Dickey-Fuller Test	1.97740
Probability	0.6020

Source: Fieldwork

Table 2 Model Evaluation Metrics Estimates

Metrics	ARIMA-N	ARIMA-E
MAE	0.85	0.78
RMSE	1.12	1.05
AIC	210.4	205.7
Residual Skew	~0	>0.5

Source: Fieldwork

Discussion

ARIMA-E shows slightly better performance in terms of error metrics and model fit. Residuals from ARIMA-E exhibit positive skewness, suggesting better accommodation of asymmetric shocks. Normal vs. Exponential Errors: While normal errors assume symmetry and light tails, exponential errors capture skewness and potential outliers more effectively.

Economic Implications: GDP shocks are often asymmetric (sudden downturns vs. gradual recoveries), making exponential error models more realistic.

7 Conclusion

This study established that the proposed Autoregressive Integrated Moving Average model with Exponential Error (ARIMA-E) model offered improved performance in modeling skewed and long tail

GDP data associated to Nigeria historical Nominal Gross Domestic Product. That could better capture asymmetric shocks and good forecast capability.

8 Recommendation

The proposed model ARIMA-E is hereby recommended for modeling skewed and long tailed time series data that exhibited exponential growth.

9 Suggestions for further study

i. The exponential error assumption might be generalize across all economics time periods.
ii. Future research could explore other non-Gaussian error structures.

References

Box and Jenkins (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.Scientific Research Open Access. https"//www.script.org

Central Bank of Nigeria (2025). CBN free online database access

Enders W. (2015) Applied Econometric Time Series. Wiley

Engle R. F.and Rusaell J. R. (1998) Autoregressive Conditional Duration ACD Model. Online www.ideas.repec.org/

Hamiton J. D. (1914) Time Series Analysis. Princeton University Press, BRUCE E. HANSEN, Boston. www.jstor.org

Harvey (2013) Dynamic Models for Volatility and Heavy Tails. Cambridge University Press

Xiufeng Yan (2021) Autoregressive Conditional Duration Modelling of High Frequency Data. online

DYNAMICS OF CRUDE OIL PRICE, PRODUCTION AND EXPORTATION IN NIGERIA (2006 – 2024): A TIME-SERIES ANALYSIS

Ibeh, G. C. [1*], Ajaraogu, J. C. [1] and Onyenekwe, C. E. [2]

[1]Department of Mathematics/Statistics, Federal Polytechnic Nekede, Owerri, Imo State, Nigeria
[2]Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

[*]Corresponding author Email: gibeh@fpno.edu.ng

Abstract

This study examines the dynamic relationships between crude oil prices, domestic production, and exportation in Nigeria from January 2006 to December 2024 using advanced time-series methodologies. Monthly data sourced from the Central Bank of Nigeria was analyzed, encompassing crude oil prices (USD/barrel), production, and export volumes (million barrels/day). A missing data point for April 2023 was addressed using linear interpolation. Seasonal-Trend Decomposition using Loess (STL) revealed underlying trend structures, though statistical tests confirmed no significant seasonal effects across the variables. Stationarity was established through differencing, and a Vector Autoregressive (VAR) model with Granger causality testing found no significant lagged influence of price fluctuations on production. Volatility analysis using a Dynamic Conditional Correlation GARCH (DCC-GARCH) model identified strong persistence in crude oil price and production volatility but no short-term volatility spillovers. The analysis further revealed a perfect linear relationship between crude oil production and exportation, necessitating the exclusion of exportation from multivariate estimations to prevent collinearity bias. These findings underscore that while price volatility has limited short-run impact on production, structural bottlenecks and domestic market rigidities continue to constrain responsiveness. Comparative analysis across pre- and post-2014 and COVID-19 periods highlighted structural shifts in volatility patterns and a decline in output and export volumes. The study concludes with policy recommendations aimed at improving resilience in Nigeria's oil sector amidst external market shocks.

Keywords: Crude Oil Price, Volatility, Time Series Analysis, DCC-GARCH, Production, Exportation, Nigeria

1.0 Introduction

1.1 Background to the Study

Crude oil plays a fundamental role in Nigeria's economy, serving as the country's primary revenue source and a key determinant of macroeconomic stability. Since the discovery of oil in commercial quantities in Oloibiri, Bayelsa State in 1956, Nigeria has remained a major player in the global oil market, with crude oil exports contributing significantly to foreign exchange earnings and government revenues. According to Central Bank of Nigeria (CBN) (2022), the petroleum sector accounts for over 85% of government revenue and more than 90% of the country's export earnings, making Nigeria highly dependent on oil fluctuations for economic stability. However, the volatility of crude oil prices has posed a persistent challenge, affecting production levels, export performance and overall economic growth.

The dynamics of crude oil prices are influenced by multiple factors including global demand and supply shocks, geopolitical tensions, production quotas imposed by the Organization of

the Petroleum Exporting Countries (OPEC) and market speculation. Historical trends indicate that Nigeria has experienced periods of high oil price volatility, particularly during the 2008 global financial crisis, the 2014 oil price collapse, and the 2020 COVID-19 pandemic-induced downturn. These events have had profound effects on oil production and exportation, impacting government revenues and national economic planning. In addition, Nigeria's oil production has faced setbacks, with declining output observed in recent years due to a combination of aging oil fields, divestments by international oil companies, and production limitations imposed by OPEC. Furthermore, the country struggles with refining capacity constraints, leading to heavy dependence on crude oil exportation while importing refined petroleum products at high costs.

1.2 Statement of the Problem

Nigeria's dependence on crude oil revenues makes the country highly vulnerable to oil price shocks. Periods of price volatility, such as the 2014 oil price collapse and 2020 COVID-19 pandemic downturn, have demonstrated the risks associated with this dependence. Fluctuations in crude oil prices can disrupt production planning, affect export volumes, and create instability in fiscal policy. However, the extent to which these price variations influence crude oil production and exportation levels in Nigeria remains insufficiently explored.

Moreover, while previous researchers such as Faruk (2020) and Yunusa (2020) have examined exchange rate volatility and oil price dynamics, few studies, if any, have focused on the volatility transmission between crude oil prices, production and exportation in Nigeria. The extent to which oil price volatility propagates through the production and exportation channels, particularly during different economic regimes (e.g., pre- and post- 2014 price collapse), remains an open question.

1.3 Objectives of the Study

The study seeks to address the discovered gaps by

(i) Estimate and assess the impact of missing data (April, 2023) on observed trends.
(ii) analyze seasonal patterns and long-term trends in crude oil price (COP), domestic crude oil production (DCOP) and crude oil exportation (COE).
(iii) investigate lagged relationships between crude oil price fluctuation and production/exportation dynamics.
(iv) measure volatility transmission between crude oil prices, production and exportation.
(v) compare oil sector behaviour of stability versus high volatility.
(vi) Examine the degree of interdependence between crude oil production and exportation in Nigeria and account for potential collinearity between both variables. Where perfect correlation is detected, one of the variables will be excluded from multivariate models such as VAR and DCC-GARCH to avoid redundancy.

1.4 Statement of Hypotheses

Based on the research objectives, the following hypotheses will be tested:

H_{01}: There are no significant seasonal patterns in crude oil prices, production and exportation

H_{02}: There is no lagged relationship between crude oil price fluctuations and production/exportations dynamics

H_{03}: Crude oil price volatility does not significantly influence production and exportation levels

H_{04}: There is no significant difference in the behaviour of crude oil prices, production and exportation between stable periods and high-volatility periods

H_{05}: Changes in crude oil production significantly influence exportation trends under different price regimes.

2.0 Literature Review

2.1 Theoretical Framework

This study is underpinned by a combination of economic and econometric theories relevant to understanding the complex dynamics of oil prices, production, and exportation in Nigeria.

2.1.1 Resource Curse Hypothesis

The resource curse theory provides an overarching context for interpreting the long-term economic instability often observed in oil-dependent economies like Nigeria. This theory supports the examination of oil price shocks and their effect on production and exportation behaviors, particularly in high-volatility periods.

2.1.2 Dutch Disease Theory

This theory offers insight into how oil revenue booms affect exchange rates and the broader economy by discouraging diversification and distorting trade patterns. It is particularly relevant for evaluating exportation patterns and regime-based production dynamics, especially in relation to price volatility.

2.2 Empirical Review

Chinanuife et al. (2021) applied the EGARCH model alongside cointegration techniques to examine quarterly data on oil prices and inflation volatility. The study revealed that negative oil price shocks significantly drive inflation volatility. However, seasonality in crude oil production and exportation patterns remains underexplored.

Faruk (2020) applied ARDL and Granger causality tests to monthly data on domestic oil production, prices, and exchange rates in Nigeria. The study revealed a long-run relationship between oil prices and domestic production. Similarly, Usoro and Ekong (2022) identified bilateral causality between oil prices and production, indicating that oil price shocks lead to production variability and suggesting feedback loops that support the use of lag-based models.

Abdulkareem and Abdulhakeem (2016) highlighted oil price volatility as a key driver of macroeconomic instability using multi-frequency data. Ige and Obi (2018) confirmed significant effects of oil price volatility on exchange rates and revenues. Kuhe et al. (2024) identified volatility clustering and asymmetric shocks in oil price returns using GARCH models. Adi et al. (2022) further revealed significant volatility spillovers from Brent oil prices to the USD/Naira exchange rate using a VAR-AGARCH model, underscoring the transmission mechanisms between global oil markets and Nigeria's domestic macroeconomic variables.

Obaka et al. (2022) examined oil price volatility and economic performance over five decades, revealing differences in impact across economic regimes. Ayodele et al. (2024) showed how investment outcomes in marginal fields shift dramatically under different oil price recovery scenarios. This supports the need to compare stable and high-volatility periods in oil sector performance.

Yunusa (2020) studied how exchange rate volatility affects crude oil exports, implying a need to consider joint impacts of oil price and production shifts. Sami and Taiwo (2023) reviewed literature on oil's influence on GDP but lacked empirical modeling of how production affects exports.

Although some studies, such as that by Sami and Taiwo (2023), discussed the effects of crude oil production on GDP, few incorporate empirical data on how crude oil price volatility impacts both production and exportation jointly. Moreover, while previous studies have applied models such as MGARCH and EGARCH to capture volatility dynamics, limited efforts have been made to integrate these insights with seasonal decomposition and lag-based causality analysis in the context of Nigeria's

oil sector. This study addresses that gap by combining GARCH, VAR, and STL decomposition techniques to uncover the dynamic, seasonal, and volatility transmission effects of oil price shocks on production and export patterns.

3.0 Methods

3.1 Research Design

This study adopts a quantitative research design utilizing time series analysis to examine the dynamics of crude oil prices, production and exportation in Nigeria from January 2006 to December 2024.

3.2 Source of Data

The data used in the study is a secondary data obtained from CBN. The data comprises the monthly observations on crude oil prices (measured in USD per barrel), domestic crude oil production (measured in million barrels per day) and crude oil exportation (measured in million barrels per day). It is worthy of note that the data point for April, 2023 is not available (is missing), and imputation technique will be applied to address this gap.

3.3 Analytical/Modeling Framework

Time Series Econometric Theory : Time series econometric models such as STL decomposition, VAR, and GARCH provide the theoretical foundation for analyzing seasonal patterns, long-term trends, lagged relationships, and volatility transmission.

3.4 Methods of Data Analysis

This study employs time-series analysis to examine the dynamics of crude oil price, production and exportation in Nigeria over the period 19 years from January 2006 to December 2024.

3.5 Treatment of Collinearity and Missing Observations

Preliminary diagnostics revealed a perfect linear association between crude oil production and exportation. To address this, exportation was excluded from the VAR and DCC-GARCH estimations to prevent multicollinearity. The single missing observation for April 2023 was treated through linear interpolation. All statistical analyses were executed in R. Levels of statistical significance were explicitly reported at 1%, 5%, and 10%, corresponding to ***, **, and * symbols, respectively.

4.0 Results and Discussion

4.1 Results

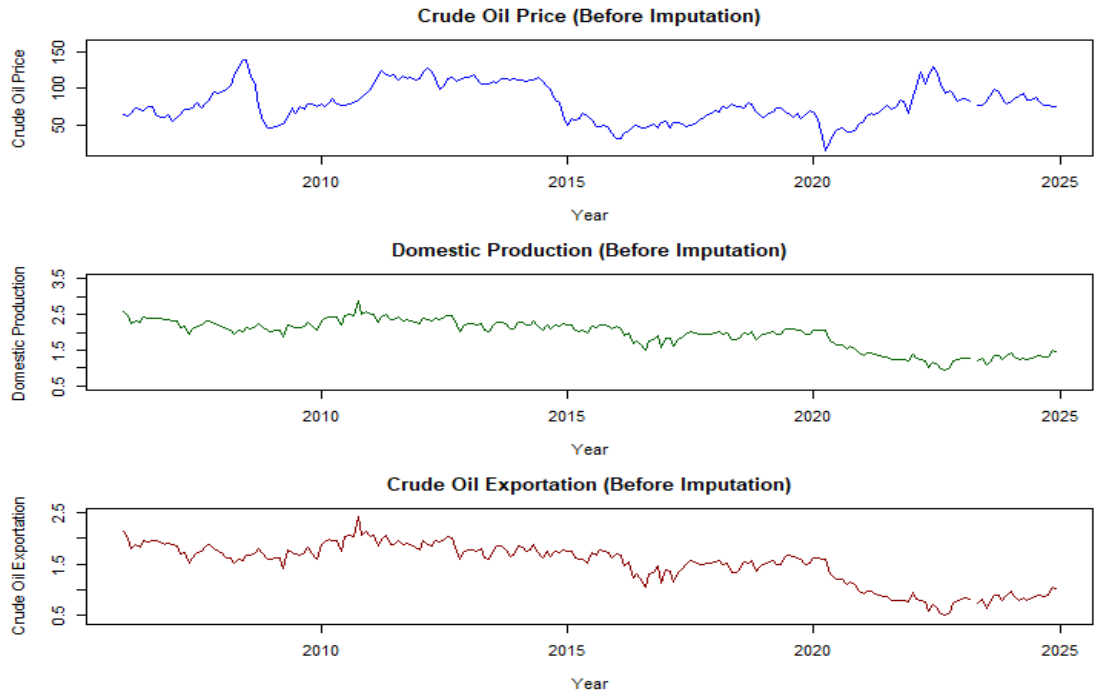
4.1.1 Data overview and Preprocessing

To analyze the long-term trends and seasonal variations in crude oil prices, production and exportation, it is essential to first preprocess the dataset. This involves examining the structure of the data, handling missing values, and ensuring that the date format is correctly set. Since the dataset contains separate year and month columns, a new date column was created to represent the full date in year-month-day format.

Time-Series Plots for Crude Oil Price, Production, and Exportation

To visualize the historical patterns of crude oil price, domestic production, and exportation, the following time-series plots are generated:

Figure 1: Time Series Plots of Crude Oil Price, Domestic Production, and Exportation with Missing Data (Jan 2006 – Dec 2024)



Crude Oil Price over Time

The time series plot of crude oil prices from January 2006 to December 2024 reveals substantial fluctuations, indicative of the volatility inherent in the global oil market. Periods of pronounced price escalation are observed around 2008–2009 and 2021–2022, whereas a significant price collapse is evident around 2020, coinciding with the onset of the COVID-19 pandemic and associated economic disruptions. These price dynamics likely reflect the influence of geopolitical tensions, supply chain interruptions, shifts in global demand, and macroeconomic conditions. Overall, the observed trends provide critical insight into the responsiveness of crude oil prices to external shocks and broader market forces.

Domestic Crude Oil Production over Time

The domestic crude oil production series demonstrates relative stability across the observed period, with only moderate fluctuations. A gradual decline becomes apparent following 2015, suggesting potential impacts of evolving market conditions, regulatory interventions, or operational challenges within the sector. The slight resurgence in production towards the end of the study period may point to adjustments in market strategy or technological improvements. These trends underscore the complex interplay between production capacity, policy frameworks, and international market pressures.

Crude Oil Exportation over Time

Crude oil exportation exhibits a broadly similar trend to production, albeit with a more pronounced decline following 2015. This divergence suggests that while production remained relatively steady, a greater proportion of crude oil may have been redirected towards domestic consumption or strategic reserves. Additionally, export volumes appear sensitive to shifts in international trade policies, demand patterns, and potential trade restrictions. The exportation dynamics presented in the plot thus reflect both internal market adaptations and external global economic developments.

4.1.2 Missing Data Imputation

Imputation Methodology

To address the missing data point for April 2023 in the crude oil price, domestic production, and exportation time series, linear interpolation was applied. This technique estimates the missing value by computing the average of the immediately preceding (March 2023) and following (May 2023) observations. Linear interpolation was selected due to its simplicity, transparency, and suitability for handling isolated missing values where temporal continuity and smooth trends are assumed.

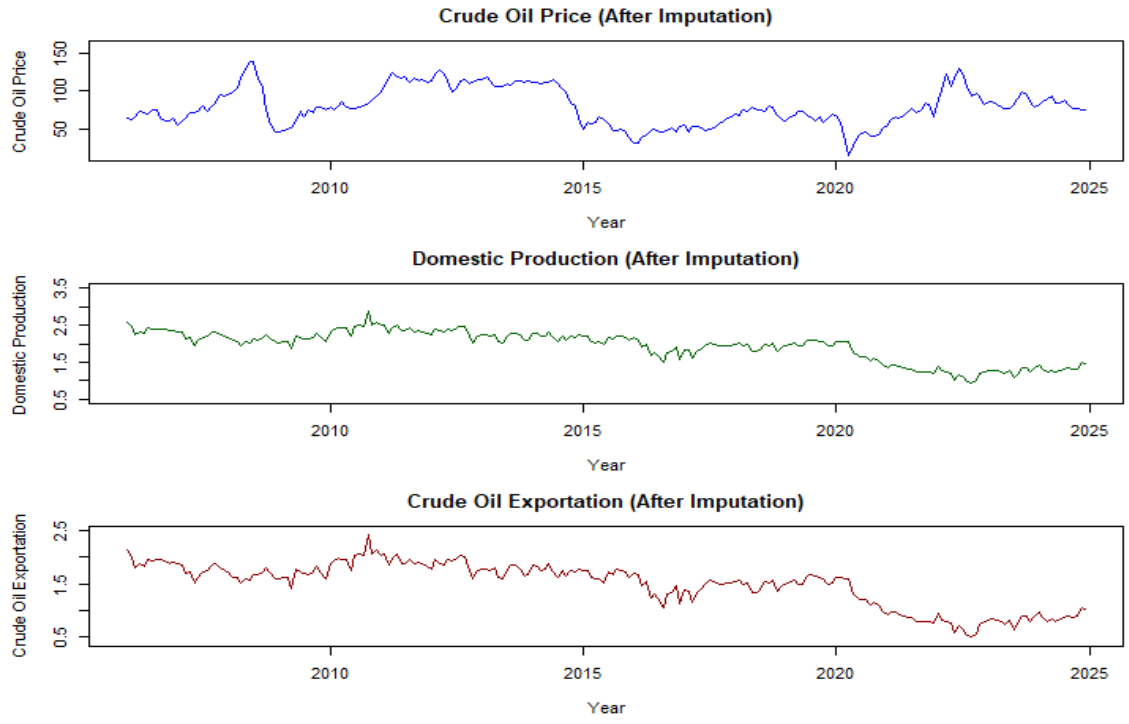
Impact Assessment

To assess the effect of the interpolation on the dataset's integrity, a visual comparison between the original (containing the missing data) and the imputed time series was conducted. Furthermore, we evaluated the magnitude of the interpolation's impact by calculating both the absolute and relative differences between the interpolated values and their neighboring monthly observations.

Table 1: Interpolated Value for April 2023 and Corresponding Monthly Changes

Indicator	March	April (interpolated)	May	Abs. Change	% Difference from Average
Crude Oil Price (USD)	81.10	79.01	76.91	0.60	0.76
Dom Crude Oil Production(mb/d)	1.27	1.23	1.19	0.04	3.25
Crude Oil Exportation(mb/d)	0.82	0.78	0.74	0.04	5.13

Figure 2: Time Series Plots of Crude Oil Price, Domestic Production, and Exportation After Linear Interpolation (Jan 2006 – Dec 2024)



4.1.3 Seasonal-Trend Decomposition using Loess (STL)

To further analyze long-term trends and seasonal variations, STL decomposition was performed on crude oil price, production, and exportation data. This decomposition separates each time series into trend, seasonal, and remainder (residual) components, providing a visual and analytical understanding underlying periodicities.

Figures 3, 4 and 5 display the STL decomposition plots for crude oil prices, production, and exportation. Each time series demonstrates distinguishable seasonal patterns across months, with trends revealing periods of stability, volatility, and structural shifts, notably around the 2014 oil price collapse and the 2020 COVID-19 pandemic.

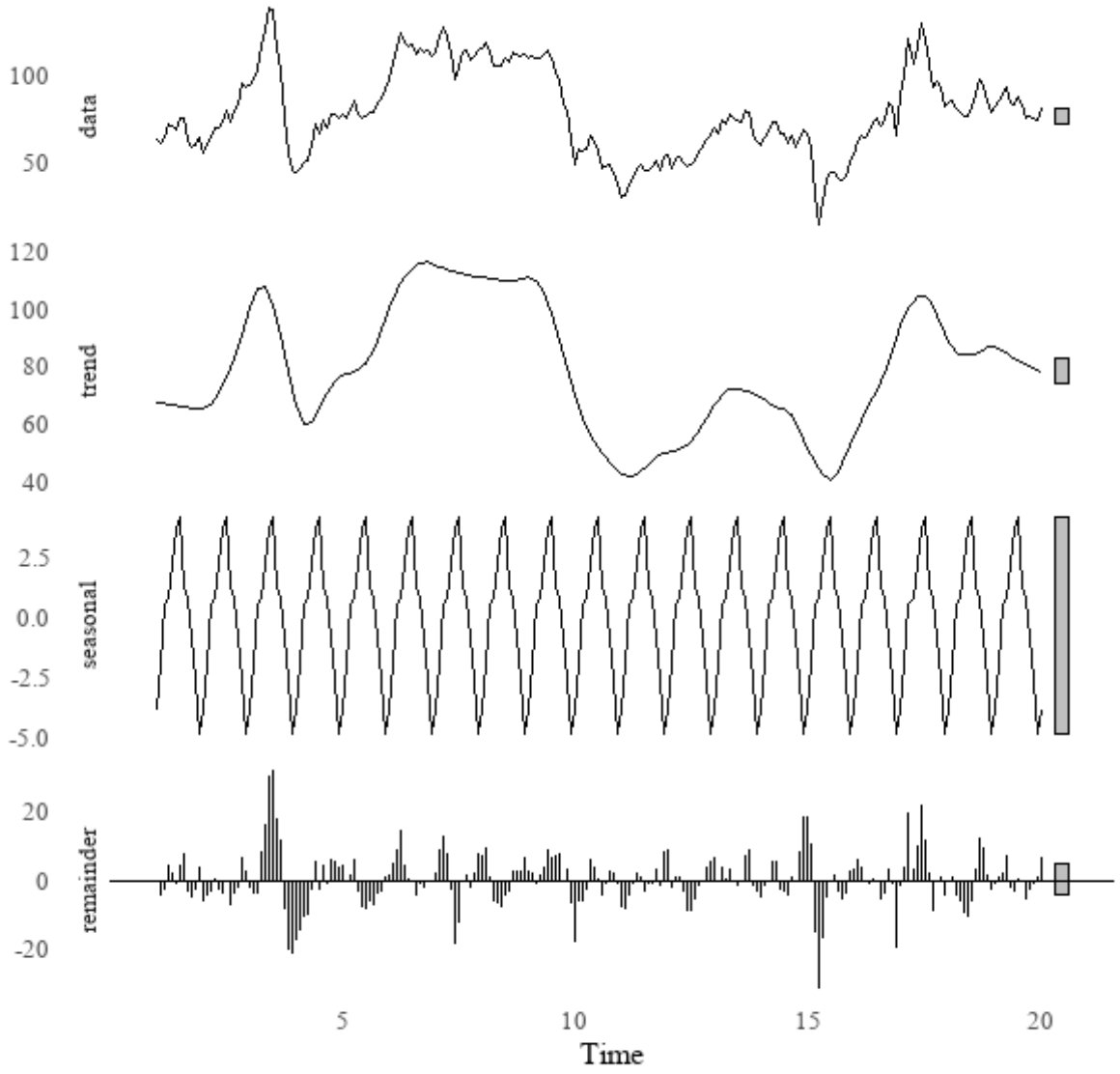


Figure 3: Crude Oil Price Decomposition

- **Trend Component:** The trend lines in the decomposition plots highlight the overall direction of crude oil price, production, and exportation. The results confirm periods of sustained growth and decline, reinforcing the findings from the time-series plots.
- **Seasonal Component:** The seasonal patterns reveal recurring fluctuations at regular intervals. For crude oil price, we observe consistent peaks and troughs that may be linked to seasonal demand changes, such as increased heating oil consumption in winter.
- **Residual Component:** The residuals represent short-term irregularities that are not explained by trend or seasonality. These fluctuations may be due to unexpected geopolitical events, economic crises, or sudden supply chain disruptions.

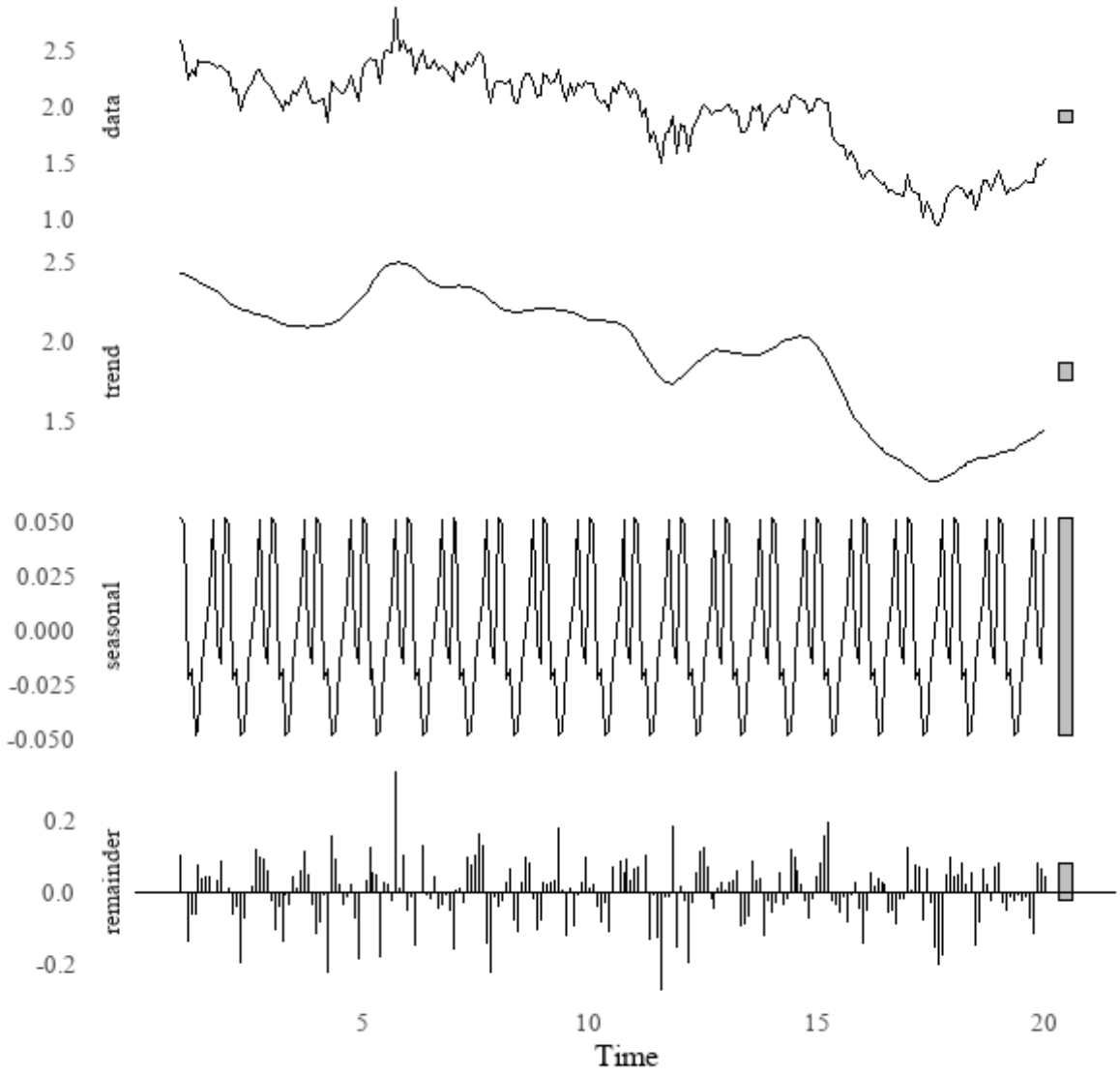


Figure 4: Crude Oil Production Decomposition

Similar patterns of seasonality and trend are observed in crude oil production, influenced by global demand cycles, maintenance schedules, or policy-driven production adjustments.

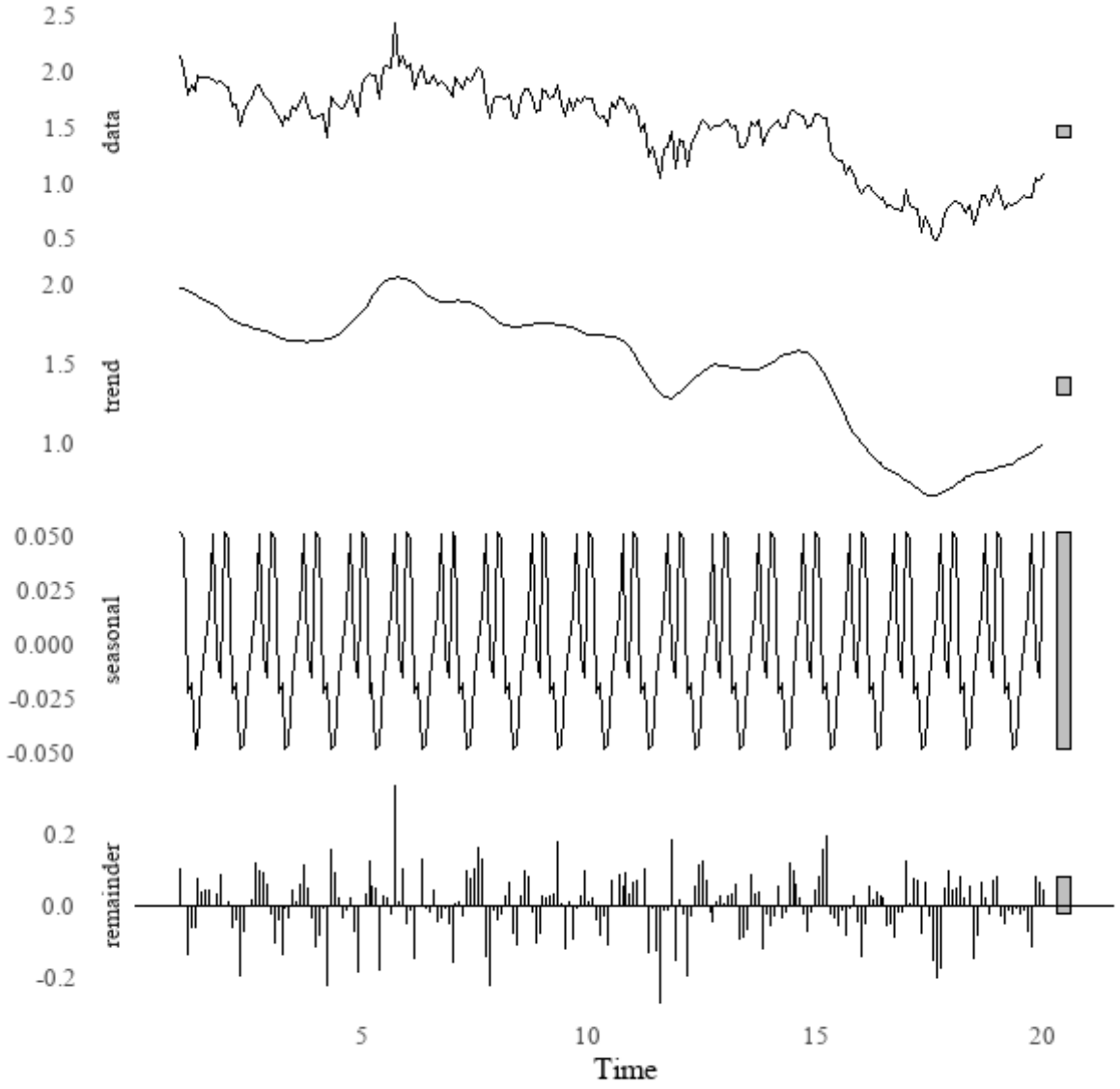


Figure 5: Crude Oil Exportation Decomposition

Crude oil exportation also follows seasonal variations, possibly driven by market agreements, international trade restrictions, or economic shifts in importing countries.

To assess seasonal variation in Nigeria's crude oil indicators, the Kruskal-Wallis test was applied to monthly groupings of price, production, and exportation data. Contrary to visual indications, statistical results revealed no significant seasonal differences in any of the three series ($p > 0.98$ in all cases). These findings suggest that oil sector behavior in Nigeria is not seasonally patterned, possibly due to external market dependencies, global pricing mechanisms, or production stability driven by fixed quotas and long-term contracts. Accordingly, the null hypothesis (H_0) could not be rejected.

Notably, crude oil production and exportation exhibit nearly perfect linear correlation ($r = 1.00$), indicating a strong dependency structure. This relationship results in identical non-significant Kruskal-Wallis statistics for the two variables, a consequence of rank-based testing applied to highly collinear series.

4.1.4 Lag Analysis and Causality

Temporal Relationships and Lag Effects:

This section investigates whether fluctuations in crude oil prices have any lagged influence on Nigeria's crude oil production. Specifically, the aim was to test the hypothesis:

H_0: There is no lagged relationship between crude oil price fluctuations and production dynamics in Nigeria.

To achieve this, the study employed Vector Autoregressive (VAR) modeling combined with Granger causality tests, which are suitable techniques for analyzing dynamic interactions between time series variables.

Monthly data on crude oil price and production were used. Prior to modeling, the time series were subjected to the Augmented Dickey-Fuller (ADF) test to ensure stationarity—a key assumption for VAR modeling and Granger causality testing.

- The original series were found to be non-stationary; hence they were differenced.
- The differenced series passed the ADF test, indicating stationarity.

Table 2: Stationarity Tests Results before and After Diffferencing

Series	Before Differencing		After Differencing	
	ADF statistic	p-value	ADF statistic	p-value
Crude Oil Price (USD)	-2.480	0.3743	-5.8843	< 0.01
Dom Crude Oil Production(mb/d)	-1.9204	0.6094	-6.2185	< 0.01
Crude Oil Exportation(mb/d)	-1.9204	0.6094	-6.2185	< 0.01

During preliminary analysis, the crude oil exportation variable was found to be perfectly correlated with domestic crude oil production (r =1.0), suggesting redundancy. To avoid multicollinearity issues in the VAR model, the exportation variable was excluded from further modeling, retaining only crude oil price and production for VAR estimation.

VAR Model Specification

The optimal lag length for the Vector Autoregressive (VAR) model was selected using four commonly applied information criteria: the Akaike Information Criterion (AIC), Hannan–Quinn Criterion (HQ), Schwarz Criterion (SC), and Final Prediction Error (FPE). The AIC and FPE both suggested an optimal lag length of 2, while the HQ and SC favored a more parsimonious lag of 1. Given the study's focus on forecasting and dynamic interactions, a VAR (2) model was chosen as it offered a better balance between model complexity and predictive power.

The VAR(2) model was estimated using the first-differenced crude oil price and production return series to ensure stationarity. The coefficients of the lagged crude oil price terms in the crude oil production equation were found to be statistically insignificant. This suggests that changes in crude oil prices do not have a strong lagged predictive effect on production levels, underscoring a weak dynamic linkage between the two variables in the short run.

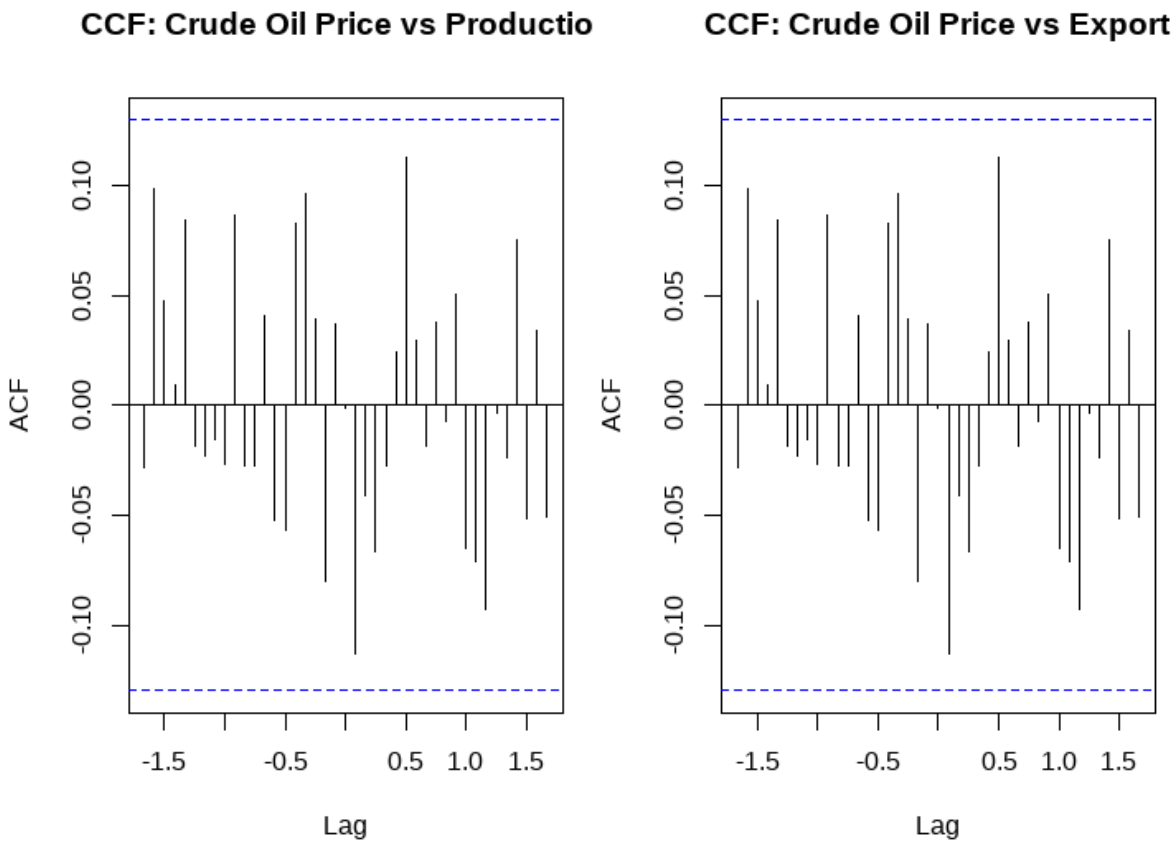


Figure 6: Cross-Correlation Function (CCF) Plots of Crude Oil Price with Domestic Production and Exportation

Cross-correlation functions (CCFs) were computed to explore potential lead-lag relationships between crude oil price and both domestic crude oil production and crude oil exportation. The resulting CCF plots showed that all correlation coefficients at various lags fell within the 95% confidence bands, indicating a lack of statistically significant correlations at any lag. This suggests no clear evidence of a temporal dependency based on visual inspection alone.

However, further formal testing using Granger causality analysis will be conducted to statistically validate or refute the existence of lagged relationships.

Granger Causality Test

A formal Granger causality test was conducted using the estimated VAR (2) model to determine whether crude oil price fluctuations significantly predict future values of production.

Results: F-statistic = 0.80943 p-value = 0.4458

Since the p-value exceeds the conventional significance threshold of 0.05, the null hypothesis could not be rejected. Thus, there is no statistically significant lagged causal relationship from crude oil price to production in Nigeria. That is, past movements in oil prices do not Granger-cause changes in production levels.

4.1.5 Volatility Modeling and Transmission Analysis

Volatility Transmission Analysis using DCC-GARCH Model:

To quantify the volatility transmission between crude oil prices and production levels, the Dynamic Conditional Correlation Generalized Autoregressive Conditional Heteroskedasticity (DCC-GARCH) model was employed. The exportation series was excluded due to its perfect correlation with production returns, which could induce multicollinearity in the model. This multivariate time series approach

captures time-varying co-movements in volatilities, allowing us to assess whether shocks in crude oil price volatility significantly spill over to production volatility.
The model was specified as a DCC (1,1) with univariate GARCH (1,1) for each return series (crude oil price (Crude_ return) and crude oil domestic Production (Prod_return)).
The estimation results:
Model Fit and Information Criteria:
Number of observations: 228, Log-likelihood: 567.2683, AIC: 5.0725, BIC: 5.2380

Table 3: Univariate GARCH Parameters

Series	ω	α_1	β_1
Crude_Return	11.2764	0.4526 (***)	0.3635 (***)
Prod_Return	0.0031	0.1181 (*)	0.6315 (***)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. All p-values are two-sided

These results confirm the presence of conditional heteroskedasticity in both series, justifying the use of GARCH-type modeling.

Table 4: DCC Parameters (Volatility Transmission)

Parameter	Estimate	Std error	z-value	p-value
α_1	0.0008	0.0220	0.0340	0.9730
β_1	0.9523	0.0525	18.1510	< 0.001

While β_1 is highly significant, indicating strong persistence in the dynamic correlation process, the α_1 is statistically insignificant (0.9730). This suggests that volatility shocks from crude oil prices do not cause immediate volatility changes in production returns. Thus, the results provide no sufficient evidence of short-term volatility transmission from crude oil prices to production. However, the significant β_1 value shows strong persistence in conditional correlations over time, indicating that while immediate shock transmission is weak, long-run co-movement may still exist.

4.1.6 Comparative Regime Analysis

Pre- and Post-2014 Oil Price Regime

A comparative analysis of Nigeria's crude oil market across the pre- and post-2014 periods reveals notable structural changes. The average crude oil price declined from $92 (pre-2014) to $68.8 (post-2014), with a mild reduction in volatility (SD: 23.6 vs. 21.3). In contrast, average production and exportation declined substantially, from 2.26 and 1.81 million barrels/day to 1.68 and 1.23 million barrels/day, respectively. Importantly, their standard deviations increased—indicating higher relative volatility in the post-2014 era.

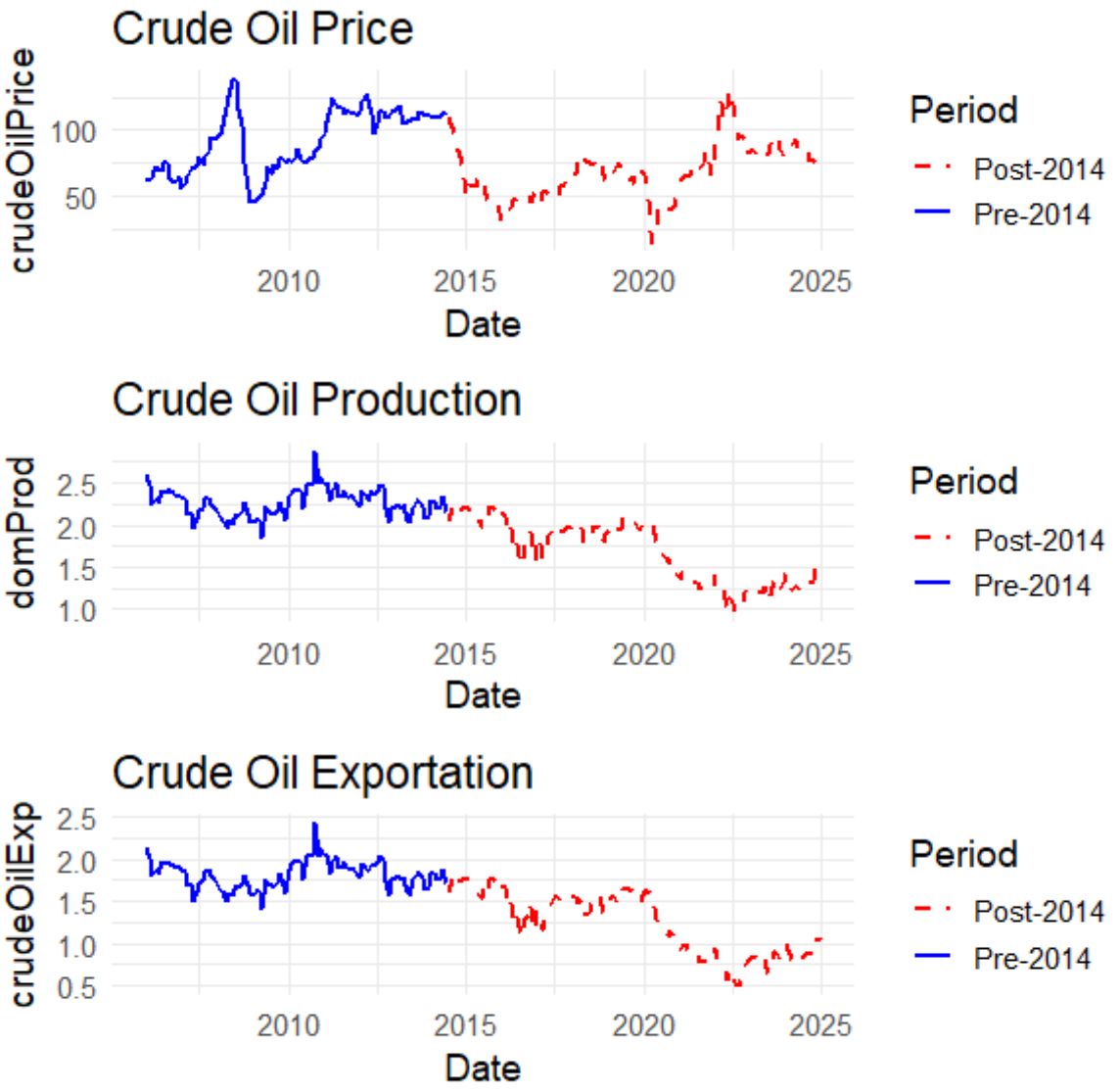


Figure 7: Crude Oil Price, Production, and Exportation in Nigeria — Pre- vs Post-2014 Periods

Crude Oil Price

The chart shows a noticeable shift in price dynamics between the pre-2014 and post-2014 periods.

- **Pre-2014**: Prices were relatively high and more stable, hovering consistently above $90/barrel, with only moderate fluctuations.
- **Post-2014**: A significant collapse is evident after mid-2014, with increased volatility. Prices became more erratic, with deep troughs (e.g., around 2020 due to COVID-19) and sharp recoveries. This indicates a more unstable price environment post-collapse.

Crude Oil Production

- **Pre-2014**: Production remained relatively high, fluctuating slightly around 2.2–2.5 million barrels/day.
- **Post-2014**: There's a consistent downward trend in production, particularly sharp from 2016 onwards. This may reflect the challenges of sustaining production in a more volatile and less profitable market.

Crude Oil Exportation

- **Pre-2014**: Exportation mirrored production levels, with relatively stable patterns above 1.5 million barrels/day.
- **Post-2014**: A downward shift is also observed in export volumes, with a notable decline beginning around 2016 and bottoming out during the COVID-19 period.

Levene's test was used to formally assess the equality of variances. The results indicated significant differences in return volatility for crude oil prices ($p = 0.0077$) and exportation ($p = 0.0061$), and marginal significance for production ($p = 0.0672$). These outcomes support the hypothesis of heightened volatility transmission and instability in crude oil exportation dynamics following the 2014 oil price collapse.

These findings highlight the critical need for more adaptive production and export policies in Nigeria's oil sector to buffer against price shocks and market instability.

GARCH (1,1) Model Analysis of Crude Oil Returns Pre- and Post-2014

To examine the volatility dynamics of crude oil prices across different time periods, GARCH (1,1) models were estimated separately for the pre-2014 and post-2014 samples. The results indicate a notable shift in the nature of crude oil return volatility between the two periods.

In the pre-2014 period, the model yielded a statistically insignificant mean return ($\mu = 0.0037$, $p = 0.497$), while the volatility parameters were highly persistent. The GARCH term ($\beta_1 = 0.6821$, $p = 0.0205$) and ARCH term ($\alpha_1 = 0.3169$, $p = 0.0101$) were both significant, with the sum $\alpha_1 + \beta_1 \approx 0.999$ suggesting near-unit root behavior in volatility. This indicates that shocks to crude oil returns during this period had long-lasting effects, reflecting a high degree of volatility clustering and persistence.

In contrast, the post-2014 period exhibited different volatility characteristics. While the mean return remained statistically insignificant ($\mu = 0.0057$, $p = 0.524$), the volatility dynamics showed a higher sensitivity to recent shocks. The ARCH term increased substantially ($\alpha_1 = 0.6409$, $p = 0.002$), whereas the GARCH term ($\beta_1 = 0.1293$, $p = 0.344$) was no longer statistically significant. The combined persistence measure ($\alpha_1 + \beta_1 = 0.7702$) was also much lower than in the pre-2014 period, indicating a regime shift toward more transitory volatility.

Diagnostic tests confirmed the adequacy of the fitted models in both periods. The Ljung-Box tests on squared standardized residuals and Autoregressive Conditional Heteroskedasticity Lagrange Multiplier (ARCH LM) tests did not indicate remaining autocorrelation or heteroskedasticity, validating the appropriateness of the GARCH (1,1) specification. Furthermore, the Nyblom stability test indicated parameter stability in the post-2014 model but marginal instability in the pre-2014 estimates, possibly reflecting structural shifts prior to 2014.

Thus, the results suggest a significant change in the volatility structure of crude oil prices post-2014, characterized by reduced persistence but increased sensitivity to recent price shocks. This could reflect broader changes in the global oil market, such as increased market speculation, evolving geopolitical dynamics, and the rise of unconventional oil production methods.

4.1.7 Pre- and Post-COVID-19 Volatility

To evaluate the impact of the COVID-19 pandemic on Nigeria's crude oil sector, the dataset was segmented into Pre-COVID (before March 2020) and Post-COVID periods. Descriptive statistics reveal notable shifts across all key indicators:

- Crude Oil Prices averaged $79.5 (SD = 25.7) before the pandemic, compared to $77.3 (SD = 23.2) after, indicating a slight decline in average prices accompanied by persistent volatility.
- Domestic Production experienced a significant reduction, dropping from a mean of 2.15 million barrels per day (SD = 0.217) pre-COVID to 1.34 million barrels per day (SD = 0.214) post-COVID.
- Similarly, Crude Oil Exportation levels fell sharply, from an average of 1.70 million barrels per day (SD = 0.217) to 0.89 million barrels per day (SD = 0.214).

These figures suggest a substantial contraction in Nigeria's crude oil output and export activities in the aftermath of the pandemic, with the standard deviations remaining comparable across periods—hinting at stable but lower activity levels post-COVID.

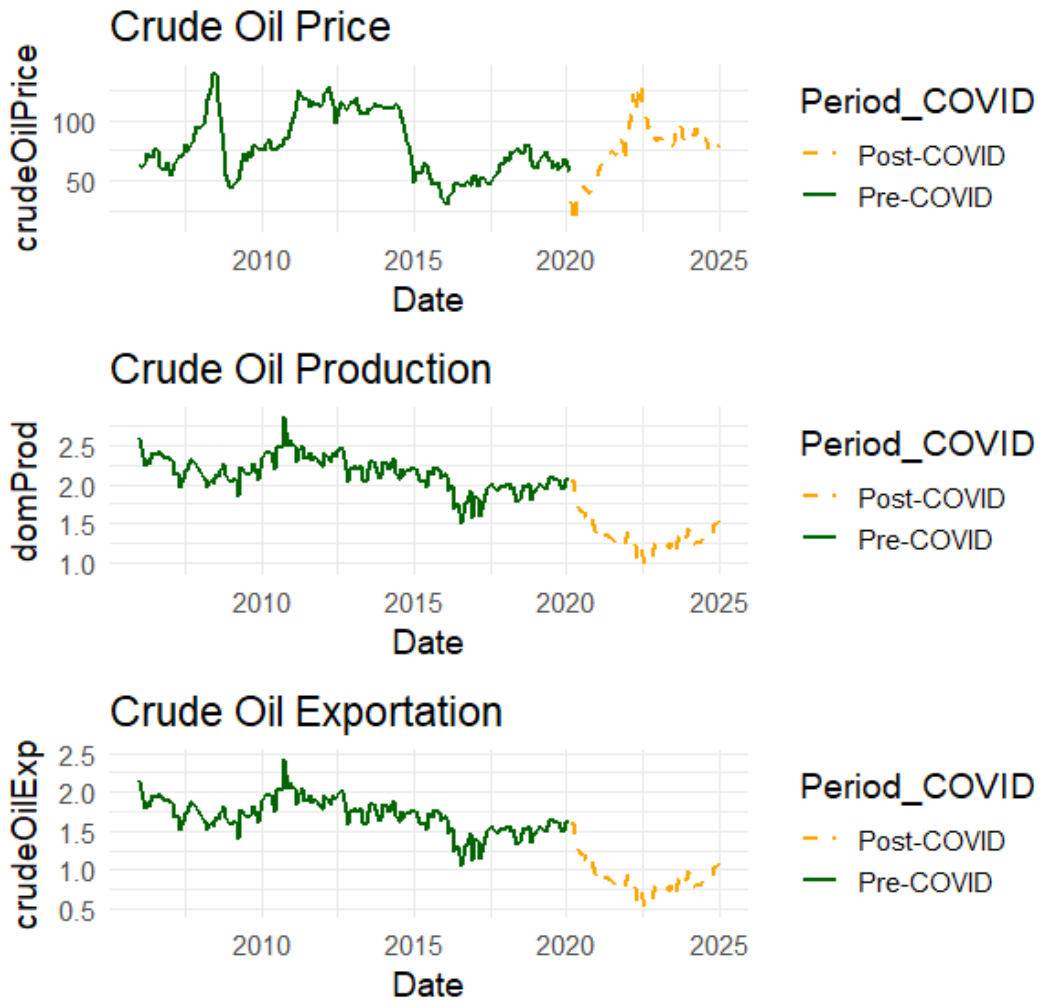


Figure 8: Crude Oil Price, Production, and Exportation in Nigeria — Pre- vs Post-COVID-19 Periods

Levene's Test was also conducted to assess whether the variance (volatility) of crude oil market variables differed significantly between the Pre- and Post-COVID-19 periods.

- For Crude Oil Price, results were marginally significant (p = 0.0525), suggesting that price volatility may have increased after the onset of the COVID-19 pandemic. This aligns with the dramatic fluctuations observed visually during 2020–2021.
- For both Production and Exportation, p-values were above 0.05, indicating no statistically significant change in variance across the two periods.

4.2 Discussion

This study examined the interrelationships among crude oil prices, production, and exportation in Nigeria over the period 2006–2024 using advanced time-series techniques. The results both affirm and diverge from several previous studies.

The lack of statistically significant seasonal patterns, confirmed through Kruskal-Wallis tests, suggests that Nigeria's oil market dynamics are less influenced by cyclical trends and more by structural and external shocks. This finding contrasts with studies like Chinanuife et al. (2021), who emphasized seasonal inflation patterns linked to oil price shocks, but aligns with Yunusa (2020) who emphasized that Nigeria's crude oil exports are more driven by exchange rate volatility and trade conditions, rather than intrinsic seasonal variation.

The absence of a significant Granger causal relationship between crude oil prices and production contradicts the conclusion by Usoro and Ekong (2022), who found bilateral causality between price and production using MGARCH models. The discrepancy may be due to differences in methodology; while Usoro and Ekong used raw monthly data without differencing, this study applied differencing and vector autoregression, which better handles non-stationarity.

Furthermore, Faruk (2020) identified a long-run relationship between oil price and production in the Niger Delta using ARDL, whereas our national-level VAR (2) model found no significant lagged effects, indicating that local dynamics may not translate to national production outcomes. This supports the idea that external market signals may not influence Nigeria's output in the short run, possibly due to OPEC quotas, outdated infrastructure, or political constraints.

The DCC-GARCH analysis provided novel insight. While previous works such as Kuhe et al. (2024) and Adi et al. (2022) confirmed the presence of volatility clustering and significant spillovers in oil markets, our results showed strong persistence in volatility, but no significant short-term volatility transmission from prices to production. This distinction is crucial: it indicates that although oil price shocks are persistent, they do not immediately disrupt production — likely due to delayed operational response or rigid production contracts.

The post-2014 and post-COVID-19 regime analyses revealed significant structural changes. Price volatility became more sensitive to recent shocks, while production and export levels declined sharply. This finding is consistent with Ige and Obi (2018) who noted that oil price volatility disrupts macroeconomic stability, although our study goes further by quantifying volatility behavior using GARCH decomposition pre- and post-events.

Notably, the perfect correlation between production and exportation ($r = 1.00$) suggests a lack of diversification in Nigeria's crude oil utilization — reinforcing the conclusions drawn by Obaka et al. (2022), who emphasized the detrimental impact of export-dependent oil economies on sustainable development.

Beyond the production–export nexus, the dynamics observed in this study have direct implications for domestic access to petroleum products. Nigeria's limited refining capacity and recurrent supply disruptions mean that fluctuations in crude oil production or export policies often translate into domestic scarcity, queuing, and hoarding. Periods of elevated international prices typically coincide with increased domestic pump prices and temporary shortages, reflecting both supply-chain constraints and speculative withholding of products. These outcomes suggest that the observed weak short-run responsiveness of production to price shocks may exacerbate internal distributional challenges. Addressing these gaps requires stronger investment in local refining, improved inventory management, and transparent allocation mechanisms to minimize hoarding and enhance public welfare.

This study, by integrating VAR and DCC-GARCH approaches, contributes uniquely by demonstrating persistent but decoupled volatility transmission, a nuanced finding absent in many prior works that did not differentiate between short-term shock spillover and long-term volatility co-movement.

5.0 Summary and Recommendations

This study investigated the dynamic relationships between crude oil prices, production, and exportation in Nigeria from 2006 to 2024 using robust time-series techniques. Key findings reveal that while crude oil prices exhibit high volatility, there is no significant seasonal pattern in any of the variables. Additionally, the results indicate no meaningful lagged relationship between crude oil prices and production or exportation, suggesting that production levels are not immediately responsive to price fluctuations.

Volatility analysis showed strong persistence in both price and production variances, but no evidence of short-term volatility spillovers. Structural analysis across pre- and post-2014, as well as pre- and post-COVID-19 periods, revealed marked declines in production and exportation, accompanied by changes in the behavior of volatility—shifting from highly persistent to more shock-sensitive patterns.

Overall, the findings highlight Nigeria's vulnerability to global oil shocks and underscore the need for adaptive strategies that can cushion the economy from external price disruptions. Enhancing domestic refining capacity, diversifying revenue sources, and improving production resilience remain critical steps for stabilizing the country's oil sector and broader economy.

References

Abdulkareem, A., & Abdulhakeem, K. A. (2016). Analyzing oil price–macroeconomic volatility in Nigeria. *CBN Journal of Applied Statistics*, *7*(1), 1–23.

Adi, A. A., Adda, S. P., & Wobilor, A. K. (2022). Shocks and volatility transmission between oil price and Nigeria's exchange rate. *Springer Nature Business & Economics*, *2*(47), 1–17.

Ayodele, O. I., Salufu, S. O., Onolemhemhen, O. R. U., & Isehunwa, S. O. (2024). Crude oil price volatility and its impact on the development of marginal fields: A case study from the Niger Delta Basin, Nigeria. *Springer Nature Business & Economics*, *4*(72).

Chinanuife, E., Magboo, K., & Zekeri, M. (2021). Oil price volatility and inflation level in Nigeria: An exponential GARCH approach. *International Journal of Advanced Research*, *9*(8), 1–8.

Faruk, B. U. (2020). Relationship between volatility in domestic oil production, oil price and exchange rate in Nigeria: Co-integration and Granger causality tests. *Bullion*, *44*(4), 1–16.

Ige, O. G., & Obi, B. (2018). The impact of crude oil price volatility on the Nigerian economy. *Bingham Journal of Economics and Allied Studies*, *1*(1), 1–18.

Kuhe, D. A., Udoumoh, E. F., & Oche, D. (2024). Volatility analysis of crude oil prices in Nigeria. *FUDMA Journal of Sciences*, *8*(1), 125–134.

Obaka, A. I., Ogboru, I., & Goshit, G. G. (2022). Oil price volatility and the Nigerian economy: ARDL and Granger causality approaches. *Kampala International University Journal of Social Sciences*, *7*(1), 7–28.

Sami, S., & Taiwo, M. (2023). Effect of crude oil prices and production on the performance of Nigerian gross domestic product: A conceptual framework. *Journal of Human Resource and Sustainability Studies*, *11*, 698–711.

Usoro, A. E., & Ekong, A. (2022). Modeling Nigeria crude oil production and price volatility using multivariate generalized autoregressive conditional heteroscedasticity models. *African Journal of Mathematics and Statistics Studies*, *5*(1), 33–54.

Yunusa, L. A. (2020). Exchange rate volatility and Nigeria crude oil export market. *Scientific African*, *9*(1), 1–13.

EDUCATIONAL CURRICULUM DEVELOPMENT FOR DATA SCIENCE AND ANALYTICS: BRIDGING SKILLS, INDUSTRY, AND ACADEMIA IN NIGERIA

H R Bakari[1], Fati W. Usman[2] and Kaka Modu[1]

[1]University of Maiduguri, Department of Statistics
[2]Ramat Polytechnic, Maiduguri
*Corresponding author Email: harunbakari@gmail.com

Abstract

Todays economy runs on data and data science now sits at the heart of innovation and decision-making in every field. Still many developing nations see their schools wrestling with the task of shaping curricula that truly tie theory to practice. This study takes a look, at how data-science education is unfolding in Nigeria checks whether university courses line up with what industry actually needs and proposes a modular industry-relevant curriculum framework. Drawing on a survey that tapped One hundred fifty. respondents spanning academia, industry and government the study uncovered two statistically significant ties: curricula that are more up-to-date correlate with heightened hiring difficulty (chi-squared = 18.7 p = 0.005) and larger firms exhibit greater hiring capacity ($\chi^2 = 15.2$ p = 0.00). That's mean.while the findings paint a picture—while 87 % of employers signal a robust demand, for new hires a mere 22 % of graduates are judged ready to step into those roles. In its section the study offers concrete curriculum recommendations that draw on worldwide best practices while being tailored to the specific challenges of developing countries—highlighting technical mastery, interdisciplinary integration, ethical considerations and learning through experience.

Keywords: Data science, curriculum development, employability, Nigeria, skills gap, interdisciplinary education

1. Introduction

By the time the 21st century rolled around data had vaulted into the ranks of our prized commodities steering everything from boardroom verdicts to cutting-edge breakthroughs and even reshaping the societal landscape. Across continents the thirst for people who can tame and translate that data has exploded—businesses and public agencies alike are scrambling to turn numbers into actionable insight and leaner operations. In the words of Binns and Chan (2019) a robust education in data science is the linchpin that will ready the generation of professionals, for a world that runs on data.

Even though its importance is widely acknowledged many universities still grapple with crafting curricula that ready graduates for the tangled realities of modern data ecosystems. Often existing programs tilt toward theory at the cost of hands-on skills creating a disconnect between what graduates can do and what industry needs (Brookings Institution, 2023; Adebayo et al., 2022). This mismatch is especially evident in developing nations such as Nigeria, where data-science education's still, in its early stages.

This article investigates the components of a data science and analytics curriculum surveying both global and African literature to distill best-practice insights and to suggest a pragmatic modular curriculum model that weaves together technical, ethical and contextual dimensions.

Aim of the Study

The aim of this study is to examine the essential components of a contemporary data science and analytics curriculum, evaluate stakeholder perspectives on curriculum relevance and graduate employability in Nigeria, and propose a contextualized modular curriculum model that aligns with global best practices and local developmental needs.

Specific Objectives

The study is guided by the following specific objectives:

I. **To review global and African literature on the core components of data science and analytics curricula**, focusing on technical, ethical, interdisciplinary and contextual requirements.
II. **To identify the skill gaps and curriculum deficiencies within existing data science programs in Nigerian universities**, based on stakeholder perceptions from industry, academia and policy sectors.
III. **To assess the relationship between industry skill requirements, curriculum modernity and hiring challenges** faced by employers in the data-science labour market.
IV. **To evaluate the role of hands-on, project-based learning and domain-specific applications**, in improving student readiness for real-world data practice.
V. **To examine the importance of ethical training, data privacy and contextual localization** in shaping curriculum relevance for developing-nation environments.
VI. **To propose a modular practice-oriented data science curriculum model** that integrates technical skills, interdisciplinary applications, project-based training and industry partnerships suitable for Nigerian institutions

2. Literature Review

2.1 Technical Skills

Looking at the literature it's become pretty clear that solid programming and data-manipulation skills are the backbone of any data-science curriculum. Zhang et al. (2020) Argue that being comfortable with data-wrangling tools—SQL, pandas—and mastering least one general-purpose language whether Python or R is a must-have, before tackling more advanced analytics. Chandel (2024) backs this up noting that machine-learning concepts should be introduced in stages starting with the fundamentals of learning and then moving on to model evaluation, interpretability and finally deployment. Both papers underscore that genuine mastery matters than superficial exposure. They map out a technical ladder: Programming, for Data → Data Engineering & SQL → Statistical The foundations → Applied Machine Learning → MLOps and Deployment.
Evaluation will be grounded in labs (which typically comprise 30–40 % of the overall grade) reproducible notebooks and competency-based milestones.

2.2 Hands-on Project-Based Learning

Research by Persaud (2021) and Prieto-Jiménez et al. (2021) Converges on a point: project-based learning (PBL) narrows the divide between classroom theory and the demands of real-world work delivering measurable gains in both employability and problem-solving ability. Prieto-Jiménez et al. Stress the value of team-driven projects that pull from genuine data sets while Persaud highlights a noticeable boost, in retention and knowledge transfer when assessments are fashioned after actual professional deliverables. Weave scaffolded projects through strata: an opening suite of concise drills a

subsequent middle-phase of applied undertakings and a culminating capstone that enlists stakeholder collaboration. All undertakings ought, to local data reservoirs and undergo peer-review scrutiny.

2.3 Interdisciplinary Integration

Donoghue et al. (2021) and Chen et al. (2024) find that data-science has its impact when students can apply the methods to concrete domain-specific problems—whether in health, agriculture or business. Interdisciplinary courses expose learners to the terminology of the field the quirks of measurement and the need for context-aware modeling choices. Both studies warn that a curriculum focused on technical skills can churn out graduates who are technically competent but lack real-world context. A practical response is to create tracks—such as Data for Health or Data for Agriculture—co-taught by faculty, from the relevant domain and the data-science department. Weave domain ethics, into the discussion. Follow up with a contextual problem analysis.

2.4 Ethics and Data Privacy

Sayers et al. (2024) treat ethics and privacy as foundational competencies rather than optional add-ons. Their analysis pushes for repeated exposure to themes such as bias, fairness, consent, data governance and the surrounding legal and regulatory landscape. They also stress the importance of case-based learning, to hone reasoning. Accordingly Ethics and Data Privacy should be introduced as a core module and then reinforced throughout subsequent courses. Assessment can be carried out through case analyses, policy briefs and reflective essays.

2.5 Localized Curriculum for Developing Nations

Studies by Ofusori & Goyayi (2025) and Kolade et al. (2022) Argue that curricula ought to be shaped around the context—datasets, problems, resource constraints and socio-economic priorities differ markedly from those, in high-income settings. When content is localized it feels more relevant builds problem-solving capacity and gains better traction in both government and industry. Ofusori & Goyayi (2025) stress the value of case studies and datasets drawn from public-sector challenges; Kolade et al. (2024) underscores capacity-building for infrastructure-constrained settings. Regional case studies—such as public-health surveillance and agricultural yield prediction—are leveraged with open-source tools and partnerships, with government bodies and small enterprises.

2.6 Cross-Cutting Design Principles

1. Vertical integration – start by grounding learners in the core skills then weave in projects, ethical discussions and domain-specific applications guiding them toward independent context-aware practice.
2. Active authentic assessment. Lean, toward project deliverables reproducible code, stakeholder presentations and reflective write-ups eschewing examinations.
3. Interleaving. Localization. Thread ethics and local relevance through all technical modules rather, than compartmentalizing them. As an illustration machine-learning curricula ought to embed bias-detection tasks that draw on gathered data.
4. Capacity and resource planning. Ensure labs, compute access and staffing are orchestrated with constraints, at the forefront; wherever practicable harness cloud credits, streamlined stacks and shared compute nodes.
5. Stakeholder engagement. Nurture pipelines to employers and public agencies for capstone projects, internships and dataset access.

2.7 Proposed Curriculum Template

Year	Course Cluster	Sample Modules	Key Deliverables
1	Foundations	Python Programming, SQL, Intro to Data & Probability, Ethics & Privacy	Labs, quizzes
2	Core Methods	Statistical Inference, Data Visualization, Mid-level ML, Domain Elective I	Team mini-project
3	Advanced Application	Advanced ML/AI, Data Engineering, Domain Elective II, Internship	Applied report
4	Capstone	Stakeholder-defined project, Thesis, Presentation	Capstone + defense

Assessment mix: 40% projects/labs, 20% capstone, 20% exams, 10% peer evaluation, 10% portfolio.

3. Methodology

3.1 Research Design

This study employed a survey design deemed suitable because it enables the gathering of data from a clearly defined population to paint a picture of the existing conditions, viewpoints and attitudes of stakeholders regarding the data science and analytics curriculum in Nigeria. Using this approach the researcher could gauge industry readiness assess the skill competencies of data-science graduates and map their employment patterns all while pinpointing the gaps, between instruction and the demands of the industry. The decision to roll out a questionnaire brought in an eclectic mix of participants, from education ICT, finance and government circles guaranteeing that a mosaic of perspectives was faithfully recorded.

3.2 Population and Sample

The study drew from a group of 500 professionals—spanning industry employers, data scientists, academics and policymakers. By employing purposive and convenience sampling we obtained One hundred fifty responses through online survey platforms.

3.3 Instrumentation

A structured questionnaire covered five sections: demographics, skill gaps, curriculum modernity, hiring difficulty, and readiness metrics. Responses used a 5-point Likert scale. Expert validation ensured content reliability (Cronbach's $\alpha = 0.87$).

3.4 Data Collection and Analysis

Data were collected via Google Forms and analyzed in SPSS (v26). Descriptive statistics summarized findings, while Chi-square tests examined associations among key variables ($p < 0.05$).

4. Results and Discussion

4.1 Introduction

The table 4.1 presents the results of **Chi-square tests** conducted to examine the relationship between three organizational or educational variables—**Company Size**, **Industry Skill Gaps**, and **University Curriculum Modernity**—and **hiring difficulty** among employers. The test results include the Chi-square statistic (χ^2), degrees of freedom (df), and corresponding p-values used to determine statistical significance at the **5% level ($\alpha = 0.05$)**.

Table 4.1 Chi-Square Findings

Variable	χ^2	df	p	Significance
Company Size / Hiring Difficulty	15.32	4	0.004	**Significant**
Industry Skill Gaps	12.45	6	0.053	Not significant
University Curriculum Modernity	18.67	6	0.005	**Significant**

The results collectively indicate that **organizational structure** (company size) and **educational quality** (curriculum modernity) are major determinants of hiring difficulty. The **industry skill gap**, while important, appears to exert a more variable influence. The findings underscore the need for integrated workforce development policies that link **higher education reforms with labor market demands**. Universities should prioritize curriculum modernization and practical exposure, while smaller companies may need targeted support or incentives to enhance recruitment capacity. This analysis contributes to ongoing discourse on the **employability of graduates** and the **alignment between education and labor market expectations** in emerging economies like Nigeria.

4.2 Descriptive Statistics

Table 4.2 presents the descriptive summary of data collected from 150 respondents on industry readiness, skill competencies, and employment patterns among data science graduates. The metrics assessed include company hiring rates, job readiness, average number of roles, time to fill positions, and proficiency levels across technical and soft skills.

Metric	Mean	SD	n
Companies Hiring	87%	0.34	150
Job Ready Rate	22%	0.41	150
Python Skill	2.8	0.9	150
Statistics Skill	3.1	0.8	150
ML Skill	2.1	0.6	150
Ethics Skill	3.2	0.6	150
Average Roles	5.2	3.8	150
Time to Fill	94	28.5	150
Visualization Skill	2.4	0.7	150
Communication Skill	2.9	0.8	150
Problem Solving Skill	2.3	0.7	150

The data reveal that 87 % of firms are actively hunting for data-science talent, a clear sign of robust industry demand.. The average job-readiness figure lingers at just 22 % meaning only about one in five graduates are deemed ready to step into a role immediately. This pattern echoes findings: while the appetite, for data professionals is high the preparedness of new graduates remains a persistent concern. Counting 5.2 distinct positions on average the data-science arena showcases a kaleidoscope of titles—from data analysts and engineers to statisticians and visualization mavens. Yet landing a candidate for

any of these roles typically stretches to, about 94 days. The hefty 28.5-day standard deviation betrays a recruitment landscape riddled with skill-gap woes and a scarcity of suitably qualified applicants.

On a five-point gauge of skill proficiency graduates placed Ethics (3.2) and Statistics (3.1) at the top of the rankings while Communication (2.9) and Python programming (2.8) followed in close pursuit. At the extreme the scantest averages surfaced in Machine Learning (2.1) and Problem Solving (2.3) painting a picture of relatively feeble applied computational and analytical reasoning chops. Meanwhile the middling standard deviations—spanning 0.6 to 0.9—underscore a fair degree of individual variability in competence, across the sample.

The stronger showing in statistics and ethics hints that academic programmes still tilt toward theory and conceptual work rather than hands-on machine-learning or visualization practice. By contrast the modest scores in Visualization (2.4) and Problem Solving (2.3) underscore a pressing need for elements — think capstone projects, internships, case-based exercises — to sharpen applied reasoning and the knack, for interpreting real-world data.

The analysis uncovers a curious paradox: industry appetite for data-science talent is soaring. Graduates often arrive under-prepared for real-world demands. In response university programs should recalibrate, weaving together theoretical foundations with practical industry-focused experiences—emphasizing sophisticated computational modeling, hands-on project work and clear communication of analytical insights. At the time institutions need to forge stronger ties, with companies narrowing the competence gap and boosting graduates job-readiness.

5. Summary and Conclusion

This analysis confirms that the fit between data-science programmes and the Nigerian labour market remains tenuous. Though demand, from industry stays vigorous graduates often depart without the hands-on expertise required— in machine-learning, visual analytics and concrete problem-solving. The results draw attention to a two-fold obstacle: revamping curricula while simultaneously boosting capacity.

A effective curriculum weaves together solid technical expertise, a clear ethical perspective and the specific knowledge required for each discipline all anchored in Nigeria's socioeconomic landscape. By embedding project-based learning and maintaining collaboration, with industry universities can markedly boost graduates' employability while driving the nation's digital transformation.

The findings clearly demonstrate a **strong mismatch between industry expectations and graduate readiness**, supported by empirical evidence (22% job-readiness vs. 87% hiring demand). The results therefore validate the need for curriculum reform and support the proposal of a modular, practice-oriented curriculum.

6. Recommendations

1. **Curriculum Modernization:** Adopt modular, outcome-based designs emphasizing Python, SQL, ML, and MLOps.
2. **Experiential Learning:** Institutionalize capstones and internships co-supervised by industry partners.
3. **Interdisciplinary Pathways:** Encourage domain-specific applications (health, agriculture, business).
4. **Ethics Integration:** Mandate data ethics and governance across all levels.
5. **Policy Support:** NUC and TETFund should prioritize faculty retraining, cloud resources, and accreditation for data science programs.
6. **Industry Collaboration:** Establish national frameworks linking academia with data-driven organizations for shared projects and internships.

References

Adebayo, F., Musa, I., & Olayemi, T. (2022). *Curriculum development and data science in African universities. African Journal of Higher Education.*

Akinbo, S., & Okunoye, A. (2020). *Data science education and skills development in Africa. African Journal of Information Systems, 10*(1), 87–101.

Binns, R., & Chan, J. (2019). *Building data science capacity in higher education. Journal of Data Literacy, 12*(3), 45–62.

Brookings Institution. (2023). *Bridging Africa's digital skills gap.* Washington, DC. Chandel, S. (2024). *Competency-based approaches in data science education. Journal of Applied Computing, 9*(2), 115–130.

Chen, X., et al. (2024). *Interdisciplinary teaching models for data science. Computing Education Review, 16*(1), 23–39.

Donoghue, T., et al. (2021). *The role of interdisciplinarity in data education. Frontiers in Education, 6*(54), 1–9.

Kolade, M., et al. (2022). *Localizing data science education for Africa. African Journal of Computing and ICT, 15*(2), 101–119.

Ofusori, T., & Goyayi, H. (2025). *Contextualizing digital curriculum in developing nations. Journal of African Educational Research, 8*(1), 33–49.

Persaud, R. (2021). *Project-based learning in data analytics education. Education and Information Technologies, 26*(4), 5511–5528.

Prieto-Jiménez, E., et al. (2021). *Authentic assessment in STEM higher education. Journal of Learning Design, 14*(3), 80–97.

Sayers, N., et al. (2024). *Embedding ethics in data science pedagogy. Ethics and Information Technology, 26*(2), 211–229.

Zhang, L., et al. (2020). *Designing effective data science curricula. IEEE Transactions on Education, 63*(3), 230–238.*

Empirical Review of the Distribution of 2023 Presidential Election Results in Nigeria

[1]*A. Musa*

Department of Statistics, Federal Polytechnic, *Idah*, Kogi State

Abstract

This paper provides the empirical review of the distribution of valid votes of the 2023 presidential election results in Nigeria. The major aim of work is to study the spread of the valid votes scored by the major political parties that participated in the 2023 presidential election across the states, regions and geopolitical zones in Nigeria with the view to justify by empirical means the position of the Nigeria judicial system on the same election. Appropriate measure of central tendency of the valid votes of the two regions, the six geopolitical zones for the four major political parties and others were observed to ascertain if there exist or otherwise of significant difference in the valid votes across the states and FCT. It was observed that a significant difference exists in the median valid votes between the political parties; the two regions, and between the geopolitical zones. The result further shows dependency of votes obtained by various political parties and their candidates on the geopolitical zones. These conclusions were bore out of the application of simple percentage, Kruskal-Wallis test and Chi-Square test procedures. The All Progressive Congress (APC) and its candidates secured a simple majority votes and a wide spread of votes in 29 states, closely followed by the Peoples Democratic Party (PDP) and its candidates with a spread in 21 states. The Laboure Party (LP) and its candidates had a spread in 16 states and FCT, and the New Nigerian Peoples Party (NNPP) in only one state.

Keywords: Review, Election, Distribution, Valid votes, spread.

1.0 Introduction

The 2023 presidential election marked the seventh presidential elections in the fourth Republic in Nigeria. The period spanned between 1999 and 2023 which marked 24 years of uninterrupted democracy involving three transition from civilian to civilian rule in Nigeria. The 2023 Presidential election in Nigeria like other presidential election in Nigeria has been faced with allegations ranging from lake of spread of valid votes across the states, non compliance with some section of the laws, and stolen of mandate of the major opposition parties and their candidates. Over the years, Nigeria has continued to witness with growing disappointment and apprehension its inability to conduct peaceful, free, fair, and credible elections whose results are widely accepted and respected across the country(Igbuzor, 2010; Osumah and Aghemelo, 2010; Ekweremadu, 2011; Ojukwu, Mbah and Maduekwe, 2019; Gberie, 2011).The major aim of this work is to study the spread of the valid votes scored by the major political parties that participated in the 2023 presidential election across the states, regions and geopolitical zones in Nigeria with the view to justify by empirical means the position of the Nigeria judicial system on the same election. The mean valid votes of the two regions, the six geopolitical zones for the four major political parties and others were observed to ascertain if there exist or otherwise of significant difference in the spread of the valid votes across the states and FCT. It was observed that a significant difference exists in the median valid votes between the political parties, the two regions, and between the geopolitical zones. These conclusions were bore out of the application of simple percentage, Kruskal-Wallis test and Chi-Square test procedures. The All Progressive Congress (APC) and its candidates secured a simple majority votes and a wide spread of votes in 29 states, closely followed by the Peoples Democratic Party (PDP) and its candidates with a spread in 21 states. The

Laboure Party (LP) and its candidates had a spread in 16 states and FCT, and the New Nigerian Peoples Party (NNPP and its candidates came distant 4th with its presence in only one state.

1.1 Aim and Objectives

The aim of this work is carryout an empirical review of result of 2023 Presidential Election in Nigeria. The objectives of the study include is to obtain the spread of valid votes scores by the major contenders and others across the states and FCT; to ascertain if a significant differences exist in mean valid votes scored by the major contenders and others across the regions, and determine the independency or otherwise of the votes on the geopolitical zones.

1.2 Significant of the study

The 2023 Presidential election in Nigeria like other presidential election in Nigeria has been faced with allegations ranging from lake of spread of valid votes across the states, non compliance with some section of the laws, and stolen of mandate of the major opposition parties and their candidates. This work will proffer insight to this claim using empirical approach.

2.0 LITRATURE REVIEW

Luqman (2009) observed that the history of elections in Nigeria's efforts at democratization has been a checkered one. Since independence, electoral conduct in the nation's democratization efforts has been an exercise in futility. This is due to the fact, that conduct of elections in the nation's political history has been marred by fraudulent practices, corruption, and violence. It is therefore, little surprise that past efforts at democratization have collapsed on the altar of perverted elections and electoral process. So bad was the situation, that election period has come to be associated with violence and politically motivated crises. Democracy as a form of government thrives when elections are predictably regular, credible and the outcomes are acceptable to a wide spectrum of stakeholders, both local and international. in the opinion of (Luqman , 2009), their inability to effectively manage the conduct and administration of elections and electoral process has had deleterious effects on the nation's efforts at instituting credible and virile democratic system. An election is a process that is central to states' emerging from transitional democracy and attempting to consolidate democracy. The nexus between electoral process and democratic consolidation indicates the extent to which the election process is free and fair (Huntington, 1991). Election is one of the key pillars of democracy. It is the means of translating the critical element of equality of citizens in democratic societies into relating through 'one person, one vote', in the constitution of the elective offices of the state. However, this is so only, if elections are free, fair and credible (Alemika, 2011). The most glaring of such effects was the truncation of the democratic process through military intervention in the nation's politics (Odoziobodo, 2015) Nigeria's 2023 general election was one of the most keenly contested elections in the country's history. In 2015, the election of President Muhammadu Buhari was seen by many Nigerians as an opportunity to reset the country's affairs after a steep slide into backwardness, due to what was perceived as the gross incompetence of the previous administration. The Buhari administration promised to tackle security issues, particularly the persistent threat posed by Boko Haram and its breakaway faction, Islamic State West Africa Province (ISWAP), in the northeast (Aina, 2023)

3.0 METHODOLOGY

The result of 2023 presidential election in Nigeria as declared by the Independent National Electoral Commission (INEC) was obtained as published by the Vanguard Newspaper for the thirty six states and Federal Capital Feretory (FCT). The result was further reorganized into two regions namely the northern and southern regions; and geopolitical zones namely north-west, north-east, north-central, south-west, south-east and south-south. Appropriate rate, parametric and non parametric statistic(s) were used in this work.

3.1 **Percentage**: this is one of the major rates used in mathematics, statistics and other sciences. It is the ratio of he observed value to the total value per 100. It is symbolized by %, such that

% = $\frac{observed\ value}{total\ value} \times 100$. Bennett et al (2005). Percentage of votes obtained each political parties and their candidates for each the states, regions and geopolitical zones to ascertain the spread of the votes.

3.2 Kruskal-Wallis Test: this is an alternative nonparametric procedure for one factor analysis when the assumptions of the parametric test are not satisfied. Oyeka; 1999. It is based on test statistics computed from ranks determined for pool sample observations rather than observation themselves. The test statistic is symbolized by H, such that $H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)$. This statistic helps to ascertain if a significant difference exist the median votes scored by the parties in the state.

3.3 The Chi-Square Distribution

The Chi-Square distribution, denoted as χ^2, is a widely used theoretical distribution mostly to determine how well observed data fit expected distributions. This application is referred to as test of goodness-of-fit. Its also determine the independence of two categorical variables in a contingency table. This is referred to as test of independence. The test statistic is symbolized by χ^2 , such that $\chi^2 = \sum\frac{(O_{ij}-e_{ij})^2}{E_{IJ}}$

3.4 Data Presentation

Table 3.4.1: The Distribution of 2023 Presidential Election Result by States and FCT.

S/N	STATE	APC	LP	NNPP	PDP	OTHER	TOTAL
1.	Abia	8914	327095	1239	22676	10113	370037
2	Adamawa	182881	105648	8006	417611	16994	731140
3	Akwa Ibom	160620	132683	7796	214012	39978	555089
4.	Anambra	5111	584621	1967	9036	13126	613861
5.	Bauchi	316694	27373	72103	426607	10739	853516
6.	Bayelsa	42572	49975	540	68818	3420	165325
7.	Benue	310468	308372	4740	130081	16414	770075
8.	Borno	252282	7205	4626	190921	10253	465287
9.	Cross River	130520	179917	1644	95425	9462	416968
10	Delta	90183	341866	3122	161600	18570	615341
11	Ebonyi	42402	259738	1661	13503	8047	325351
12	Edo	144471	331163	2743	89585	13309	581266
13	Ekiti	201494	11397	264	89585	5462	308171
14	Enugu	4772	428640	1808	15749	5455	456424
15	FCT	90902	281717	4517	74194	8741	460071
16	Gombe	146977	26160	10520	319123	9263	510043
17	Imo	66406	360495	1552	30234	8646	459267
18	Jigawa	421390	1889	98234	386587	12431	920531
19	Kaduna	399293	294494	92969	554360	19037	1360153
20	Kano	517341	28513	997279	131716	27156	1702005
21	Katsina	482283	6376	69386	489045	11583	1058673
22	Kebbi	248088	10682	5038	285175	10539	559522
23	Kogi	240751	56217	4238	145104	10480	456790
24	Kwara	263572	31166	3142	136909	35203	469971
25	Lagos	572606	582454	8442	75750	32199	1271451
26	Nassarawa	172922	191361	12715	147093	16475	540566
27	Niger	375183	80452	21836	284898	16299	778668
28	Ogun	341554	85829	2200	123831	26710	580124
29	Ondo	369924	44405	930	115463	20286	551008

30	Osun	343945	23283	713	354366	10896	733203
31	Oyo	449884	99110	4095	182977	73419	809485
32	Plateau	307195	466272	8869	243808	62026	1088170
33	Rivers	231591	175071	1322	88468	27199	523651
34	Sokoto	285444	6568	1300	288679	4824	586815
35	Taraba	135165	146315	12818	189017	16043	499358
36	Yobe	151459	2406	18270	198567	7695	378397
37	Zamfara	298396	1660	4044	193978	4845	502923
	TOTAL	**8,794,726**	**6,101,553**	**1,496,687**	**6,984,520**	**648,474**	**24,025,940**

Source: the Vanguard, Tuesday March 19, 2023

4.0 Data Analysis

Table 4.1: Distribution of 2023 Presidential Election Result by Political parties in Nigeria

S/N	*PARTY*	*VOTES*	*NO. STATE WON*	*NO. OF STATES VOTES ≥ 25%*
1	APC	8,794,726	12	29
2	PDP	6,984,520	12	21
3	LP	6,101,533	11 + FCT	16+FCT
4	NNPP	1,496,687	1	1
5	OTHERS	648,474	0	0
	TOTAL	24,025,940	36+ FCT	

Table 4.2: The Distribution of 2023 Presidential Election Result by Northern and Southern Region.

S/N	***North***			***South***			
	Party	**Votes**	**No. Of State Won**	**Votes**	**No. Of State Won**		**Total Votes**
1	APC	5,593,339	7	3,201,387	5		8,794,726
2	PDP	5,231,589	9	1,752,931	3		6,984,520
3	LP	2,086,124	2+ FCT	4,015,409	9		6,101,533
4	NNPP	1,454,650	1	42,037	0		1,496,687
5	OTHERS	324,611	0	323,863	0		648,474
	TOTAL	14,690,714	20	9,336,016	10		24,025,940

Table 4.3: The Distribution of 2023 Presidential Election Result by Geopolitical Zones.

S/N	*Geopolitical Zones.*	*APC*	*PDP*	*LP*	*NNPP*	*OTHER*	*TOTAL*
1	**North Central**	1,759,210	1,163,087	1,417,577	60,056	164,638	4,564,568
2	**North East**	1,183,676	1,737,847	316,619	126,343	70,387	3,434,872
3	**North West**	2,650,453	2,329,665	351,928	1,268,251	89,686	6,689,983
4	**South West**	2,277,626	942,941	847,736	16,659	167,972	4,252,934
5	**South South**	798,174	718,022	1,212,675	17,167	110,933	2,856,971
6	**South East**	125,589	91,968	1,954,998	8,211	44,958	2,225,724
	Total	8,794,726	6,984,520	6,101,533	1,496,687	648,474	24,025,940

5.0 Results and Discussion

APC and its candidates scored 36.61% of the total valid votes winning in 12 states namely: Borno, Jigawa, Zamfara, Benue, Kogi, Kwara, Niger, Ekiti, ogun, Ondo, Oyo and Rivers with a simple majority votes of eight million, seven hundred and ninety four thousand, seven hundred and twenty six (8,794,726), they scored 25% and above votes in 29 states except: Abia, Anambra, Delta, Edo, Ebonyi, Enugu, Imo and FCT making a total spread of 29 states a number far above constitutional requirement of 25% of votes in $\frac{2}{3}$ states and FCT which is approximately 25 states. On the other hand, PDP and its candidates won in 12 states namely: Adamawa, Akwa Ibom, Bauchi,, Bayelsa, Gombe, Kaduna, Kastina, Kebbi, Osun, Sokoto, Taraba and Yobe with a minority votes of six million, nine hundred and eighty four thousand, five hundred and twenty (6,984,520), they scored 25% and above votes in 9 states but failed to score in Abia, Anambra, Benue, Cross River, Edo, Ebonyi, Enugu, Imo, Kano, Lagos, Ogun, Ondo, Oyo, Plateau, Rivers and FCT making a total spread of 21 states a number below constitutional requirement of 25% of votes in $\frac{2}{3}$ states and FCT which is approximately 25 states making 29.07% of the total valid votes. Similarly, LP and its candidates secured 25.40% of the total valid votes. they won in 11 states and FCT namely: Abia, Anambra, Cross River, Delta, Edo, Ebonyi, Enugu, Imo, Lagos, Nassarawa, Plateau, and FCT with a minority total votes of six million, one hundred and one thousand, five hundred and thirty three (6,101,533), they scored 25% and above votes in 5 states namely: Akwa Ibom, Bayelsa, Benue, Rivers and Taraba, but failed to score in the remaining 20 states. This made a total spread of 17 states a number far below the constitutional requirement of 25% of votes in $\frac{2}{3}$ states and FCT which is approximately 25 states. NNPP and its candidates that took distant 4^{th} position only won in Kano state and could not secure 25% of votes in the remaining 35 states and FCT. They score a total vote of one million, four hundred and ninety six thousand, six hundred and eighty seven (1,496,687) making 6.23% of the total valid votes. The remaining 14 political parties and their candidates scored a total vote of six hundred and forty eight thousand, four hundred and seventy four (648,474) which constitute 2.69% of total valid votes. APC and its candidates scored significant and substantial percentage votes of 38.08 and 34.29 in the northern and southern region respectively. PDP and its candidates scored a significant and poor percentage vote of 35.61 and 18.78 in the northern and southern region respectively. Similarly, LP and its candidates scored a very poor and very significant and substantial percentage vote of 14.20 and 43.01 in the northern and southern region respectively. NNPP and its candidates performed woefully in northern and southern regions securing a meager percentage vote of 9.90 and 0.45 respectively. APC and its candidates also staged an outstanding performance in South-West with the score of 53.55% of the votes from the zone, a very good performances in both North-West and North-Central securing 39.62% and 38.54% of the votes respectively, a good and fair performances in the North-East and South-South with percentage votes of 34.46 and 27.94 respectively, but had a very poor performance in South-East with a meager 5.64 percentage of the votes from that zone. PDP and its candidates on the other hand secured an outstanding performance in North-East with the score of 50.59% of the votes from the zone, a very good performances in North-West securing 34.82% of the votes, a fair performances in both the North-Central and South-South with percentage votes of 25.48 and 25.13 respectively, but had a poor and very poor performances in South-West and South-East with a percentage votes of 22.17 and 4.13 respectively from that zones. Similarly, LP and its candidates secured an Excellent and very good performances in South-East and South-South with percentage scores of 87.84 and 42.45 of the votes from the zones, a good and poor performances in North-Central and South-West securing 31.06% and 19.93% of the votes, but a very poor performances in both the North-East and North-West with percentage votes of 0.37 and 0.39 respectively. In the same vein, NNPP and its candidates secured a poor performance in North-West with a 18.96% votes and a very poor performances in North-Central, North-East, South-South, South-West and South-East with a meager percentage vote of 1.32, 3.68, 0.60, 0.39 and 0.37 respectively. From table 4.2.1, Kruskal-Wallis test shows a significant difference in the average valid votes scored by the parties. The Two-way ANOVA in table 4.2.2 indicates absence of significant difference in the average valid votes between the political parties; and between the regions. While table 4.2.3 indicates absence of significant difference in the average valid votes between the geopolitical zones, it also indicated a existence of significant difference in the average valid votes between the parties

5.0 **Summary and Recommendation**

In this closely forth 2023 presidential election, APC and its candidates scored 38.08% of the total valid votes winning in 12 states and scored 25% or more votes in 17 other states, indicating a wide spread of votes across the country in 29 states. Out of 19 northern states and FCT, APC won in 7 states, scoring 38.08% of the votes, it also won in 5 states out of 17 states in the southern states securing 34.29% of the votes. PDP and its candidates scored 29.07% of the total valid votes winning in 12 states and scored 25% or more votes in 9 other states, indicating a wide spread of votes across the country in 21 states. Out of 19 and northern states and FCT, PDP won in 9 states, scoring 35.61% of the votes, It also worn in 3 states out of 17 states in the south securing 18.87% of the votes. LP and its candidates scored 25.40% of the total valid votes, winning in 11 states and FCT and with 25% or more votes in 5 other states, indicating a spread of votes across the country in 16 states and FCT; out of 19 and northern states and FCT, LP won in 2 states and FCT, scoring 25.40% of the votes; It also worn in 9 states out of 17 states in the southern states securing 43.01% of the votes. NNPP and its candidate won only 1 state and this is from the northern part of the country.

From the empirical review of the 2023 presidential election in Nigeria, it is recommended that minor political parties like the labor Party (LP), New Nigeria Peoples Party (NNPP) and others to merge and form a formidable and credible alternative to the two dominating major political parties, the All Progressive Congress (APC) and the Peoples Democratic Party (PDP) in Nigeria.

References

Aina, F. (2023). *Commentary: A Chance to Reset: Nigeria after the 2023 General Election, UK*: Online publication of the Royal United Services Institute for Defense and Security Studies (RUSI).

Alemika, E.E.O. (2011). *Post-election violence in Nigeria: Emerging trend and lessons*. CLEEN Foundation Publication.

Bennett, Jeffrey; Briggs, Williams (2005). *Using and Understanding Mathematics/ A Quantitative Reasoning Approach* (3rd ed.), Pearson Addison Wesley, p. 134, ISBN 0-321-22773-5

George Casselas (2008). *Statistical Design; Springer texts in statistics*. Springer: ISBN 978-0-387-75965-4.

Huntington, S.P. (1991). *The third wave: Democratization in the late Twentieth Century*. Norman, OK: University of Oklahoma Press.

Luqman, S. (2009). "Electoral institution and the management of the democratic process. The Nigeria Experience", Journal of Social Sciences, 21(1).

Odoziobodo S.I. (2013). "*The Independent National Electoral Commission and election management in Nigeria: An appraisal of the 2007 General Elections*". Unpublished PhD Thesis submitted to the Department of Political Science, Enugu State University of Science and Technology, Enugu, Nigeria

Oyeka, I.C.A (1999). *An introduction to Applied Statistics Methods,* Eight Edition, Nobern Avocation Publishing Company, Enugu.

Appendix

Table 1: Percentage Distribution of 2023 Presidential Election Result by States and FCT.

S/N	STATE	APC	LP	NNPP	PDP	OTHERS	TOTAL
1.	Abia	2.41	88.40	0.33	6.13	2.73	100
2	Adamawa	25.01	14.45	1.10	57.12	2.32	100
3	Akwa Ibom	28.94	23.90	1.40	38.55	7.20	100
4.	Anambra	0.83	95.24	0.32	1.472	2.14	100
5.	Bauchi	37.10	3.21	8.45	49.98	1.26	100
6.	Bayelsa	25.75	30.23	0.33	41.63	2.07	100
7.	Benue	40.32	40.04	0.62	16.89	2.13	100
8.	Borno	54.22	1.55	0.99	41.03	2.20	100
9.	Cross River	31.30	43.15	0.39	22.89	2.27	100
10	Delta	14.66	55.56	0.51	26.26	3.02	100
11	Ebonyi	13.03	79.83	0.51	4.15	2.47	100
12	Edo	24.85	56.97	0.47	15.41	2.29	100
13	Ekiti	65.38	3.70	0.09	29.07	1.77	100
14	Enugu	1.05	93.91	0.40	3.45	1.20	100
15	FCT	19.76	61.23	0.98	16.13	1.90	100
16	Gombe	28.82	5.13	2.06	62.57	1.82	100
17	Imo	14.46	78.49	0.34	6.58	1.88	100
18	Jigawa	45.78	0.21	10.7	42.00	1.35	100
19	Kaduna	29.36	21.65	6.84	40.76	1.40	100
20	Kano	30.40	1.68	58.60	7.74	1.60	100
21	Katsina	45.56	0.60	6.55	46.19	1.09	100
22	Kebbi	44.34	1.91	0.90	**50.97**	1.88	100

23	Kogi	52.71	12.31	0.93	**31**.77	2.29	100
24	Kwara	56.08	6.63	0.67	29.13	7.49	100
25	Lagos	45.04	45.81	0.66	5.96	2.53	100
26	Nassarawa	32.00	35.40	2.35	27.21	3.05	100
27	Niger	48.18	10.33	2.80	36.59	2.09	100
28	Ogun	58.88	14.79	0.38	21.35	4.60	100
29	Ondo	67.14	8.06	0.17	20.95	3.68	100
30	Osun	46.91	3.18	0.10	48.33	1.49	100
31	Oyo	55.58	12.24	0.51	22.60	9.07	100
32	Plateau	28.23	42.85	0.82	22.41	5.70	100
33	Rivers	44.23	33.43	0.25	16.89	5.19	100
34	Sokoto	48.64	1.12	0.22	49.19	0.82	100
35	Taraba	27.07	29.30	2.57	37.85	3.21	100
36	Yobe	40.03	0.64	4.83	52.48	2.03	100
37	Zamfara	59.33	0.33	0.80	38.57	0.96	100

Table 2: Percentage Distribution of 2023 Presidential Election Result by Political parties in Nigeria

S/N	***PARTY***	***VOTES***	***NO. STATES WON***	***PERCENTAGE***	***NO. OF STATES VOTES ≥ 25%***
1	APC	8,794,726	12	36.61	29
2	PDP	6,984,520	12	29.07	21
3	LP	6,101,533	11 + FCT	25.40	16+FCT
4	NNPP	1,496,687	1	6.23	1
5	OTHERS	648,474	0	2.69	0
	TOTAL	24,025,940	36+ FCT	100	

Table3: Percentage scores of votes by the parties in Nigeria, Regions and Geopolitical zones

S/N	*PARTIES*	*% for Total*	*% for North*	*% for South*	*% for NC*	*% for NE*	*% for NW*	*% for SW*	*% for SS*	*% for SE*
1	APC	36.16	38.08	34.29	38.54	24.46	39.62	53.55	27.94	5.64
2	PDP	29.07	35.61	18.78	25.48	50.59	34.82	22.17	25.13	4.13
3	LP	25.40	14.20	43.01	31.06	9.22	5.26	19.93	42.45	87.84
4	NNPP	6.23	9.90	0.45	1.32	3.68	18.96	0.39	0.6	0.37
5	OTHERS	2.60	2.21	3.47	3.61	2.05	1.34	3.95	3.88	2.02
	TOTAL	100	100	100	100	100	100	100	100	100

Table4: Kruskal-Wallis Test: VOTES versus PARTY

PARTY	**N**	**Median**	**Ave Rank**	**Z**
APC	37	240751	134.1	5.22
PDP	37	147093	125.5	4.12
LP	37	99110	107.8	1.88
NNPP	37	4238	37.9	-7.00
OTHERS	37	12431	59.7	-4.23
Overall	185		93.0	

H = 91.74 DF = 4 P = 0.000
H = 91.74 DF = 4 P = 0.000 (adjusted for ties)

Table 5: Chi-Square Test: APC, PDP, LP, NNPP, OTHER

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

Zone	***APC***	***PDP***	***LP***	***NNPP***	***OTHER***	***Total***
North Central	1,759,210	1,163,087	1,417,577	60,056	164,638	4,564,568
	1,670,928.08	1,326,814.92	1,159,242.54	284,358.58	123,223.88	
	4,664.293	20,203.898	57,569.224	176,930.291	13,918.806	
North East	1,183,676	1,737,847	316,619	126,343	70,387	3,434,872
	1,257,386.04	998,438.28	872,338.79	213,981.98	92,726.90	
	4,321.004	547,580.417	354,018.978	35,893.636	5,382.163	
North West	2,650,453	2,329,665	351,928	1,268,251	89,686	6,689,983
	2,448,967.89	1,944,624.18	1,699,024.50	416,765.41	180,601.03	
	16,576.882	76,239.120	1,068,065.226	1,739,654.255	45,766.860	

South West	2,277,626	942,941	847,736	16,659	167,972	4,252,934
	1,556,849.81	1,236,230.09	1,080,098.27	264,944.73	114,811.09	
	333,698.412	69,581.294	49,988.253	232,674.207	24,615.064	
South South	798,174	718,022	1,212,675	17,167	110,933	2,856,971
	1,045,836.77	830,455.75	725,571.91	177,980.52	77,126.04	
	58,648.587	15,222.183	327,010.209	145,302.357	14,818.738	
South East	125,589	91,968	1,954,998	8,211	44,958	2,225,724
	814,759.41	646,966.77	565,256.98	138,655.77	60,085.06	
	582,940.004	476,104.261	3,416,817.741	122,720.018	3,808.401	
Grand Total	8,794,728	6,983,530	6,101,533	1,496,687	648,574	24,025,052

Chi-Sq = 10040734.780, DF = 20, P-Value = 0.000

IMPACT OF MONETARY POLICY USING MACHINE LEARNING-BASED COUNTERFACTUAL ANALYSIS

Itiveh F.E [1] and Adams J.M [2]
[1]Department of Statistics, Southern Delta University, Ozoro, Delta State, Nigeria
[2]Department of Statistics, P.M.B.13, Auchi Polytechnic, Auchi, Edo State, Nigeria

Corresponding E-mail: fenoitiveh4@gmail.com

Abstract

Nigeria grapples with significant economic challenges, including high inflation, currency volatility, and unemployment, despite various monetary policy measures implemented by the Central Bank of Nigeria (CBN). This study addresses the problem of understanding the actual impact of these policies, as traditional econometric models often fail to capture the complexities of the Nigerian economy. The objectives are to analyze historical policy impacts, develop a machine learning-based counterfactual analysis framework, and provide empirical evidence to guide future monetary policy decisions. Data will be collected from reputable institutions, including the CBN and the National Bureau of Statistics, covering key economic indicators from 1990 to 2024, using a purposive sampling technique to select representative data points. The analysis will integrate traditional econometric methods, such as Vector Autoregression and Error Correction Models, with machine learning techniques, including Regression Trees and Neural Networks. Preliminary findings indicate that the Regression Tree model outperforms the Neural Network in predicting GDP, especially during economic fluctuations. The VAR model highlights significant interdependencies among GDP, inflation, interest rates, and exchange rates, underscoring the complexities of economic forecasting. Recommendations include enhancing policy formulation, refining forecasting models, and establishing regular monitoring mechanisms. This research aims to provide actionable insights for policymakers to effectively navigate Nigeria's economic challenges.

Keywords:Machine learning,Gross domestic product,Interest rate, Economic growth,Inflation rate,Counterfactual analysis, Vector Auto regression.

INTRODUCTION

The role of monetary policy in shaping economic stability is a critical area of inquiry, especially in developing economies like Nigeria. Monetary policy encompasses the strategies and measures employed by the Central Bank to control the money supply, manage inflation, and influence interest rates, all of which are integral to achieving macroeconomic stability. In Nigeria, the Central Bank of Nigeria (CBN) utilizes various instruments, including the Monetary Policy Rate (MPR), Cash Reserve Ratio (CRR), and Open Market Operations (OMO), to regulate economic activity and ensure price stability.

Historically, Nigeria has faced significant economic challenges, including high inflation rates, fluctuating exchange rates, and persistent unemployment. These challenges have led to a series of monetary policy interventions aimed at stabilizing the economy. For instance, Adenuga and Akpan (2021) highlight that despite the CBN's efforts, the effectiveness of these monetary policies remains contentious, primarily due to the complex interactions within Nigeria's economic framework. This complexity is further exacerbated by external factors, such as global economic conditions and domestic issues like fiscal policy misalignment and structural inefficiencies.

The economic landscape in Nigeria is characterized by several structural issues, including a reliance on oil exports, which makes the economy vulnerable to fluctuations in global oil prices. These dependencies can lead to significant volatility in economic performance, complicating the CBN's efforts to implement effective monetary policies. Additionally, the informal sector constitutes a large part of Nigeria's economy, making it challenging to gauge the overall impact of monetary policies accurately. This duality of the economy complicates the transmission mechanisms of monetary policy and raises questions about the efficacy of traditional instruments.

Traditional econometric models, such as Vector Autoregressions (VAR) and Structural Vector Autoregressions (SVAR), have been the mainstay for evaluating monetary policy impacts. However, these models often assume linearity and stationarity, which may not accurately capture the dynamic and often non-linear characteristics of Nigeria's economy (Okonkwo et al., 2023). Moreover, these models tend to overlook the impact of lagged variables and external shocks, which can significantly influence economic outcomes. As a result, the limitations of these traditional models have prompted researchers and policymakers to seek alternative methodologies that can provide deeper insights into monetary policy effectiveness.

In recent years, machine learning (ML) techniques have emerged as a promising alternative for economic analysis. ML-based counterfactual analysis allows for the estimation of potential economic outcomes in the absence of specific policy interventions. This approach can uncover complex relationships and provide a more nuanced understanding of how monetary policies impact key economic indicators. Moreover, the increasing availability of high-quality economic data offers an unprecedented opportunity to apply ML techniques in this context. By leveraging vast datasets from various sources, including government databases, financial institutions, and international organizations, researchers can develop models that better reflect the realities of the Nigerian economy. This data-driven approach can help policymakers identify trends, evaluate the potential impact of policy changes, and make informed decisions that enhance economic stability.

Ultimately, this study aims to utilize ML-based counterfactual analysis to assess the impact of monetary policies on Nigeria's economy, contributing to a more comprehensive understanding of the effectiveness of these policies. By bridging the gap between traditional economic analysis and modern machine learning techniques, this research aspires to offer a robust framework for evaluating monetary policy impacts in a rapidly evolving economic landscape.

Statement of the Problem

Despite the implementation of various monetary policy measures by the CBN, Nigeria continues to grapple with persistent economic challenges such as high inflation, currency volatility, and unemployment. The ongoing debates regarding the effectiveness of these policies have raised questions about their actual impact on the economy. Traditional econometric models, while valuable, may not adequately capture the complexities of Nigeria’s economic environment, leading to potential misinterpretations of policy effectiveness.

The challenge is not only to assess the impact of existing policies but also to provide actionable insights for future interventions. Policymakers require tools that can adapt to the rapidly

changing economic landscape, and traditional models often fall short in this regard. By integrating machine learning techniques into the analysis, this study aims to offer a more dynamic and responsive framework for evaluating monetary policy impacts, ultimately contributing to better economic management in Nigeria.

Aim and Objectives of the Study

The primary aim of this study is to assess the effectiveness of monetary policies in Nigeria through the application of machine learning-based counterfactual analysis. By doing so, the study seeks to provide a more nuanced understanding of how these policies influence key economic indicators and to identify actionable insights that can enhance policy formulation, while the specific objectives are:

i. To analyze the historical effectiveness of monetary policies in Nigeria, focusing on their impacts on key economic indicators such as interest rate, exchange rate inflation rate and GDP growth.
ii. To develop a machine learning-based counterfactual analysis framework for evaluating the impact of monetary policy interventions.
iii. To provide empirical evidence that can guide future monetary policy decisions in Nigeria, enhancing the overall effectiveness of these policies.

Significance of the Study

This study holds significant importance for several reasons. First, it contributes to the existing body of literature on monetary policy effectiveness in Nigeria by incorporating modern analytical techniques. By employing machine learning methodologies, this work offers a fresh perspective on how monetary policies can be evaluated and understood.

Second, the findings of this study will provide policymakers with empirical evidence that can inform future monetary policy decisions

REVIEW OF SOME RELATED LITERATURE

Conceptual Framework

Monetary policy, at its core, involves the actions undertaken by a central bank to manipulate the money supply and credit conditions to stimulate or restrain economic activity (Mishkin, 2019). Its primary

goals typically include maintaining price stability, fostering full employment, and promoting sustainable economic growth (Bernanke, 2004). The Central Bank of Nigeria (CBN), like other central banks, uses a variety of instruments to achieve these objectives. These instruments can be broadly categorized into direct and indirect controls (Ajayi, 2007). Direct controls, prevalent in the pre-1986 era in Nigeria, involved measures like interest rate ceilings and credit rationing. Indirect controls, which gained prominence post-1986 with the adoption of structural adjustment programs, rely on market mechanisms such as open market operations (OMO) and reserve requirements (Nnanna, 2001).

The effectiveness of monetary policy hinges on its transmission mechanism, which describes how changes in monetary policy instruments affect key macroeconomic variables (Taylor, 1995). This transmission occurs through various channels, including the interest rate channel, the credit channel, the exchange rate channel, and the asset price channel (Mishkin, 2007).

Key Economic Indicators: Inflation and GDP Growth

Two of the most critical indicators that monetary policy aims to influence are inflation and GDP growth. Inflation, defined as a sustained increase in the general price level, erodes purchasing power and distorts economic decision-making (Friedman, 1968). High inflation can lead to uncertainty, reduced investment, and decreased international competitiveness. The CBN Act of 2007 mandates the CBN to maintain price stability, recognizing its importance for sustainable economic growth (CBN, 2007).

GDP growth, on the other hand, reflects the overall health and expansion of the economy. It measures the total value of goods and services produced within a country over a specific period. Sustainable GDP growth is essential for job creation, poverty reduction, and improved living standards (Solow, 1956). Monetary policy plays a crucial role in influencing aggregate demand and investment, which are key drivers of GDP growth (Romer, 2012).

Machine Learning and Counterfactual Analysis

Machine learning (ML) is a branch of artificial intelligence that enables systems to learn from data without being explicitly programmed (Bishop, 2006). ML algorithms can identify complex patterns and relationships in large datasets, making them valuable tools for economic analysis and forecasting (Athey, 2018).

Counterfactual analysis involves constructing hypothetical scenarios to estimate what would have happened if a particular policy or event had not occurred (Morgan & Winship, 2015). In the context of monetary policy, counterfactual analysis can help assess the impact of specific policy interventions on inflation and GDP growth by comparing actual outcomes with those that would have prevailed under alternative policy scenarios (Pearl, 2009).

Integrating Machine Learning and Counterfactual Analysis

The integration of machine learning and counterfactual analysis offers a powerful approach to evaluating monetary policy effectiveness. ML algorithms can be used to build predictive models that capture the complex dynamics of the Nigerian economy. These models can then be used to generate counterfactual scenarios, allowing for a more rigorous assessment of the causal impact of monetary policy interventions (Athey & Imbens, 2019). This approach addresses the limitations of traditional econometric methods, which often rely on strong assumptions and may struggle to capture non-linear relationships (Varian, 2014).

Theoretical Framework

The Quantity Theory of Money

The Quantity Theory of Money (QTM) is one of the most fundamental theories in monetary economics, asserting that changes in the money supply directly affect price levels in an economy. Formulated by economists such as Milton Friedman, the QTM is often expressed using the exchange equation:

MV=PQ

Where:

- M = Money supply
- V = Velocity of money
- P = Price level
- Q = Real output (GDP)

Implications of the Quantity Theory

The QTM suggests that if the velocity of money and real output remain relatively stable, any increase in the money supply(M) will lead to a proportional increase in the price level(P) in the long run. This relationship implies that monetary policy can be a powerful tool for controlling inflation. However, the

QTM has its limitations, particularly in modern economies where velocity can fluctuate due to various factors, such as changes in payment systems and consumer behavior (Friedman, 1968).

The IS-LM Model

The IS-LM model provides a framework for analyzing the interactions between the goods market and the money market to determine the equilibrium levels of output and interest rates. This model was developed by John Hicks and is widely used in macroeconomic analysis.

Components of the IS-LM Model

i. **IS Curve**: Represents the equilibrium in the goods market, where investment (I) equals saving (S). It shows the relationship between interest rates and output (Y) in the goods market.
ii. **LM Curve**: Represents the equilibrium in the money market, where money demand equals money supply. It illustrates the relationship between interest rates and output in the money market.

Policy Implications

In the IS-LM framework, monetary policy primarily affects the LM curve. An increase in the money supply shifts the LM curve to the right, leading to lower interest rates and higher output. Conversely, a decrease in the money supply shifts the LM curve to the left, resulting in higher interest rates and lower output (Hicks, 1937).

Limitations of the IS-LM Model

While the IS-LM model provides valuable insights, it has limitations. It assumes that prices are fixed in the short run, neglecting the role of inflation and expectations in the economy (Snowdon & Vane, 2005). Additionally, the model does not account for the complexities of modern financial markets or the impact of global economic factors.

The New Keynesian Model

The New Keynesian model builds upon the IS-LM framework by incorporating microeconomic foundations such as sticky prices and wages, imperfect competition, and rational expectations (Mankiw,

2001). This model provides a more comprehensive understanding of how monetary policy influences the economy.

Sticky Prices and Wages

One of the key features of the New Keynesian model is the concept of price stickiness, which suggests that prices do not adjust instantaneously to changes in demand or supply. This rigidity can lead to short-term fluctuations in output and employment in response to monetary policy changes (Cochrane, 2011).

Role of Expectations

The New Keynesian model emphasizes the importance of expectations in determining economic outcomes. Central banks can influence expectations through their monetary policy actions and communication strategies, which can affect consumer and business behavior (Clarida, Gali, & Gertler, 1999). This underscores the importance of credibility and transparency in monetary policy.

Policy Implications

The New Keynesian framework suggests that central banks should focus on managing expectations and using interest rate rules, such as the Taylor rule, to guide their policy decisions. These rules provide a systematic approach to adjusting interest rates based on economic conditions, helping to stabilize inflation and output (Taylor, 1993).

Counterfactual Analysis in Economics

Counterfactual analysis is a powerful tool for evaluating the impact of policy interventions by constructing hypothetical scenarios. This approach allows researchers to assess what would have happened if a particular policy had not been implemented (Morgan & Winship, 2015).

Causal Inference

Counterfactual analysis is rooted in the concept of causal inference, which seeks to establish cause-and-effect relationships. By comparing actual outcomes with counterfactual scenarios, researchers can estimate the causal impact of specific interventions (Pearl, 2009). This is particularly relevant for assessing monetary policy effectiveness, as it helps to isolate the effects of policy changes from other influencing factors.

Methods for Counterfactual Analysis

Several methods can be employed for counterfactual analysis, including:

i. **Structural Models**: These models rely on economic theory to specify the relationships between variables and estimate counterfactual outcomes.

ii. **Matching Methods**: These methods create a control group that is similar to the treatment group in all respects except for the intervention, allowing for a more accurate estimation of treatment effects.

iii. **Machine Learning Techniques**: Machine learning offers a flexible and data-driven approach to counterfactual analysis, enabling the estimation of treatment effects in complex and non-linear settings (Athey & Imbens, 2019).

Integrating Machine Learning into Monetary Policy Analysis

The integration of machine learning techniques into monetary policy analysis and counterfactual evaluation represents a significant advancement in economic research. Machine learning can enhance the predictive power of economic models and provide more robust estimates of policy impacts.

Predictive Modeling

Machine learning algorithms can analyze large datasets and identify complex patterns and relationships that traditional econometric models may overlook (Bishop, 2006). This capability allows researchers to develop more accurate predictive models for key economic indicators, facilitating better policy decisions.

Generating Counterfactual Scenarios

By leveraging machine learning, researchers can generate counterfactual scenarios that account for a wide range of factors influencing economic outcomes. This approach provides a more nuanced understanding of the potential effects of monetary policy interventions (Athey, 2018).

Empirical Applications

Several studies have successfully applied machine learning techniques to evaluate monetary policy effectiveness. For instance, Hinterlang (2021) utilized reinforcement learning to optimize monetary policy, demonstrating the potential of machine learning to enhance policy evaluation frameworks.

METHODS AND MATERIALS

Research Design

This study employs a quantitative research design to assess the effectiveness of monetary policies in Nigeria, utilizing both traditional econometric techniques and advanced machine learning methods. The quantitative approach facilitates the systematic collection and analysis of numerical data, enabling the identification of relationships and patterns between monetary policy interventions and key economic indicators, such as inflation and GDP growth. By leveraging statistical methods, the study aims to provide robust empirical evidence that can inform policy formulation and implementation.

The research design will adopt a longitudinal perspective, examining data from 1990 to 2024. This time-frame captures significant economic events, policy changes, and structural shifts in the Nigerian economy, providing a comprehensive overview of the dynamics between monetary policy and economic performance.

Population and Sources of Data of Study

The population for this study comprises all relevant data related to monetary policies and key economic indicators in Nigeria over the specified period. This includes data on the Monetary Policy Rate (MPR), inflation rates, GDP growth rates, money supply, exchange rates, and unemployment rates. The data will be sourced from reputable institutions, including the Central Bank of Nigeria (CBN), the National Bureau of Statistics (NBS), and international organizations like the World Bank and the International Monetary Fund (IMF).

Sample/Sampling Technique

A purposive sampling technique will be employed to select specific data points that are representative of key monetary policy interventions and their impacts on economic indicators from 1990 to 2024. This approach allows for the selection of data from periods of significant monetary policy changes, such as adjustments in the MPR or other relevant measures undertaken by the CBN.

The sample will consists of annual data over the specified period, ensuring a robust analysis of trends and patterns. This selection process will help capture the effects of various monetary policies on the Nigerian economy, particularly during periods of economic instability or significant policy shifts.

Methods of Data Analysis

The data analysis will utilize a combination of traditional econometric techniques and machine learning algorithms to assess the effectiveness of monetary policy in Nigeria. The specific methods employed will include:

i. **Descriptive Statistics**: Initial analysis will involve descriptive statistics to summarize the data, calculating means, medians, and standard deviations of key variables. This step provides a foundational understanding of the data and identifies trends over time.

ii. **Econometric Analysis**:
    - **Vector Autoregression (VAR) Model**: This model will be employed to analyze the dynamic relationships between monetary policy and economic indicators. The VAR model captures the interdependencies among multiple time series variables, allowing for the assessment of how changes in the MPR, inflation rates, money supply, exchange rates, and unemployment rates affect GDP growth over time.
    - **Error Correction Model (ECM)**: This model will assess both short-term and long-term relationships between monetary policy and economic indicators. The ECM is particularly useful in capturing the adjustment process towards long-term equilibrium, revealing how quickly the economy responds to policy changes.

iii. **Machine Learning Techniques**:
    - **Regression Trees**: Regression trees will be used to model non-linear relationships between the monetary policy variables and economic outcomes. This technique helps identify interactions between variables and provides insights into which factors most influence economic performance, including MPR, money supply, and unemployment rates.
    - **Neural Networks**: Neural networks will capture complex, non-linear relationships in the data. This approach allows for the modeling of intricate patterns that traditional econometric models may not fully capture, enhancing predictive accuracy.

iv. **Counterfactual Analysis**: Using the econometric and machine learning models developed, counterfactual analysis will be conducted to estimate what economic outcomes would have been under different monetary policy scenarios. This analysis will provide valuable insights into the causal effects of monetary policy decisions.

Model Specification

i. Vector Autoregression (VAR) Model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

Where Y_t is a vector of endogenous variables, including interest rate, exchange rate, inflation rate and GDP growth.

β_0 is the intercept, a constant

$\beta_{1,}\ \beta_{2,}\ \ldots, \beta_{p,}$ are the coefficients of the lags of *Y* till order *p*,

ϵ_t is the error.

ii. Error Correction Model (ECM):

$$\Delta Y_t = \alpha + \beta_1 \Delta\text{INR} + \beta_2 \Delta\text{EXR} + \beta_3 \Delta\text{INF} + \beta_4 \Delta\text{GDP} + \gamma(Y_{t-1} - \theta)$$

This model captures both short-term dynamics and long-term relationships among the variables, providing a comprehensive view of how monetary policy influences economic performance.

i. Regression Trees:

$$Y = f(X) + \epsilon$$

Where Y is the dependent variable (GDP growth), and X is a vector of independent variables, including interest rate, exchange rates and inflation rate.

ii. Neural Networks:

$$Y = f(W.X + b)$$

Where Y is the output variable (GDP growth), and X represents the input vector of monetary policy variables.

Justification of the Models

The selected econometric and machine learning models are justified based on their ability to capture the complexities of the Nigerian economy and the specific research questions of this study. The VAR model is particularly useful for analyzing the interrelationships between multiple economic indicators, allowing for a comprehensive understanding of how changes in monetary policy impact inflation, GDP growth, and other key variables over time. The ECM complements this by providing insights into the speed of adjustment and long-term relationships.

Machine learning techniques, such as regression trees and neural networks, are employed to enhance predictive accuracy and model non-linear relationships that traditional econometric models may not fully capture. These advanced techniques allow for a more nuanced analysis of the data, identifying patterns and interactions that can inform effective monetary policy decisions.

DATA PRESENTATION, ANALYSIS, DISCUSSION OF RESULTS

Data Presentation

This section presents the analysis of key economic variables: Gross Domestic Product (GDP), Interest Rate (INR), Exchange Rate (EXR), and Inflation Rate (INF). The data, sourced from the World Bank Indicator 2023, provides insights into the economic landscape and highlights trends and interrelationships among these variables. By examining these indicators, we can better understand the dynamics influencing economic stability and growth.

Table 4.1: Data on GDP, INR, EXR and INF

YEAR	*GDP*	*INR*	*EXR*	*INF*
1990	11.77689	17.46624	8.038285	7.3644
1991	0.358353	0.990847	9.909492	13.00697
1992	4.631193	-14.9872	17.29843	44.58884
1993	-2.03512	-7.05247	22.0654	57.16525
1994	-1.81492	-15.9202	21.996	57.03171
1995	-0.07266	-31.4526	21.89526	72.8355
1996	4.195924	-5.26078	21.88443	29.26829
1997	2.937099	12.12661	21.88605	8.529874
1998	2.581254	11.48467	21.886	9.996378
1999	0.584127	6.047248	92.3381	6.618373
2000	5.015935	-1.14089	101.6973	6.933292
2001	5.917685	12.1387	111.2313	18.87365

2002	15.32916	3.023542	120.5782	12.87658
2003	7.347195	9.935713	129.2224	14.03178
2004	9.250558	-2.60485	132.888	14.99803
2005	6.438517	-1.59368	131.2743	17.86349
2006	6.059428	-5.62797	128.6517	8.225222
2007	6.59113	9.187171	125.8081	5.388008
2008	6.764473	6.684909	118.5667	11.58108
2009	8.036925	18.18	148.88	12.53783
2010	8.005656	1.067736	150.2975	13.74005
2011	5.307924	5.68558	153.8625	10.82614
2012	4.230061	6.224809	157.5	12.22424
2013	6.671335	11.20162	157.3117	8.495518
2014	6.309719	11.35621	158.5526	8.047411
2015	2.652693	13.59615	192.4403	9.009435
2016	-1.61687	6.686234	253.492	15.69681
2017	0.805887	5.790567	305.7901	16.50227
2018	1.922757	6.055977	306.0837	12.09511
2019	2.208429	4.522188	306.921	11.39642
2020	-1.79425	5.37128	358.8108	13.24602
2021	3.647187	1.227719	401.152	16.95285
2022	3.251681	0.919232	425.9792	18.84719
2023	2.860215	1.23305	645.1941	24.65955
2024	3.760215	2.534505	1478.965	27.42955

Source: World bank indicator, 2024

Table 4.2: Summary Statistics

Variable	Min	Q1	Median	Mean	Q3	Max
GDP	-2.035119	2.0655933	4.195924	4.231879	6.514823	15.32916
INR	-31.452565	-0.1108284	5.371280	3.002798	9.561442	18.18000
EXR	8.038285	57.2017500	131.274333	198.867085	222.966167	1,478.96520
INF	5.388008	9.5029066	13.006973	18.539518	18.355341	72.83550

The table below summarizes key statistics for four variables: GDP, INR (Interest Rate), EXR (Exchange Rate), and INF (Inflation Rate). Each variable is interpreted based on its estimated statistics.

GDP (Gross Domestic Product)

The GDP values range from a minimum of **-2.035119** to a maximum of **15.32916**, indicating significant variability in economic performance. The first quartile (Q1) is **2.0655933**, while the median is **4.195924**, suggesting that half of the observations reflect a positive growth environment. The mean

GDP of **4.231879** indicates a slight upward trend, influenced by a few higher values. This persistence in GDP growth aligns with the efficient market hypothesis, which posits that past performance is a reliable predictor of future outcomes.

INR (Interest Rate)

The interest rates show a wide range, with a minimum of **-31.452565**, indicating extreme conditions, possibly involving negative rates. The Q1 value of **-0.1108284** suggests that a quarter of the rates are below zero, indicating low rates were common. The median interest rate of **5.371280** reveals that half the observations fall below this level, reflecting a generally moderate interest rate environment. The mean of **3.002798** is lower than the median, indicating some outlier low rates. The maximum rate of **18.18000** suggests that higher rates were also experienced.

EXR (Exchange Rate)

The exchange rate ranges from **8.038285** to **1,478.96520**, reflecting substantial volatility. The first quartile at **57.2017500** shows that a quarter of exchange rates fall below this level. The median value of **131.274333** indicates that half of the rates are below this point, while the mean of **198.867085** is skewed higher due to outliers. The high maximum indicates potential periods of significant currency devaluation or instability.

INF (Inflation Rate)

Inflation rates range from a minimum of **5.388008** to a maximum of **72.83550**, indicating varying inflationary pressures. The first quartile at **9.5029066** reflects a baseline level of inflation, while the median of **13.006973** shows that half the rates are relatively moderate. The mean of **18.539518** is higher than the median, suggesting that some high inflation periods are present. The maximum value indicates extreme inflationary conditions that could impact economic stability.

Table 4.3: Stationarity Test (ADF)

Variable	ADF_Statistic	Lag_Order	P_Value	Stationary
GDP	-1.7810	3	0.6581	No
INR	-2.5494	3	0.3597	No

Variable	ADF_Statistic	Lag_Order	P_Value	Stationary
EXR	2.3807	3	0.9900	No
INF	-3.5872	3	0.0483	Yes

The results of the Augmented Dickey-Fuller (ADF) test for stationarity are as summarized below. This test helps determine whether the time series data for each variable is stationary or contains a unit root. The ADF statistic of **-1.7810** with a p-value of **0.6581** indicates that the GDP series is not stationary. Since the p-value is greater than the common significance level of 0.05. The ADF statistic of **-2.5494** and a p-value of **0.3597** also suggest that the INR series is not stationary. The ADF statistic for the exchange rate is **2.3807**, with a p-value of **0.9900**, indicating non-stationarity.In contrast, the ADF statistic of **-3.5872** and a p-value of **0.0483** indicate that the inflation rate series is stationary. The p-value is less than 0.05, allowing us to reject the null hypothesis of a unit root.

Data Analysis

4.2.1 VAR Model Results

Table 4.4: Equation: GDP

Variable	Estimate	Std_Error	t_value	p_value
GDP.l1	0.3920000000	0.17590	2.229	0.0337
INR.l1	-0.0267300000	0.10630	-0.252	0.8032
EXR.l1	-0.0000003355	0.00445	0.000	0.9999
INF.l1	-0.0415000000	0.07119	-0.583	0.5644
const	3.1850000000	2.29000	1.391	0.1749

The results from the Vector Autoregression (VAR) model for the variables GDP, INR (Interest Rate), EXR (Exchange Rate), and INF (Inflation Rate) are summarized as follows. Each equation represents the relationship between the current value of a variable and its lagged values as well as the lagged values of the other variables.

The coefficient for the lagged GDP (GDP.l1GDP.l1GDP.l1) is 0.3920, significant at the 0.05 level (p = 0.0337), indicating a positive relationship between past and current GDP, suggesting persistence in economic growth. Other variables—INR, EXR, and INF—do not show significant effects on GDP, as their p-values exceed 0.05.

Table 4.5: Equation: INR

Variable	Estimate	Std_Error	t_value	p_value
GDP.l1	0.506234	0.389719	1.299	0.204
INR.l1	-0.104126	0.235492	-0.442	0.662
EXR.l1	0.007476	0.009863	0.758	0.455
INF.l1	-0.378614	0.157762	-2.400	0.023
const	6.457141	5.075614	1.272	0.213

The lagged GDP (GDP.l1GDP.l1GDP.l1) shows a positive relationship with INR but is not statistically significant (p = 0.204). The lagged inflation (INF.l1INF.l1INF.l1) has a significant negative impact on INR (p = 0.023), suggesting that higher inflation is associated with lower interest rates, while other variables remain insignificant.

Table 4.6: Equation: EXR

Variable	Estimate	Std_Error	t_value	p_value
GDP.l1	4.440	5.452	0.814	0.422000000000000
INR.l1	2.005	3.294	0.609	0.547000000000000
EXR.l1	1.779	0.138	12.892	0.000000000000156
INF.l1	3.361	2.207	1.523	0.138600000000000
const	-168.597	71.001	-2.375	0.024400000000000

The lagged exchange rate (EXR.l1EXR.l1EXR.l1) has a highly significant positive effect on itself ($p < 0.001$), indicating strong persistence. Other variables, including GDP and INR, do not significantly influence the exchange rate. The constant term is significant, reflecting a baseline level of the exchange rate.

Table 4.7: Equation: INF

Variable	Estimate	Std_Error	t_value	p_value
GDP.l1	-0.596019	0.567159	-1.051	0.30200
INR.l1	0.083207	0.342711	0.243	0.81000
EXR.l1	-0.005945	0.014353	-0.414	0.68200
INF.l1	0.712209	0.229591	3.102	0.00426
const	9.088458	7.386550	1.230	0.22800

The lagged inflation (INF.l1INF.l1INF.l1) is significantly positive ($p = 0.004$), indicating that past inflation levels predict current inflation. Other variables do not show significant effects on inflation, suggesting a strong inertia in inflationary trends.

The VAR model results indicate significant relationships primarily for GDP and inflation, with persistence observed in several variables. The lagged values play a crucial role in predicting current outcomes, while contemporary economic factors exhibit limited statistical significance across the equations. This analysis underscores the importance of historical data in economic forecasting and suggests that policymakers and investors should consider the inertia in these relationships when making decisions.

4. 2.2 Error Correction Model (ECM)

Table 4.8: Johansen Cointegration Test (Trace Statistics):

Hypothesis	TestStatistic	CriticalValue_10pct	CriticalValue_5pct	CriticalValue_1pct
r ≤ 3	1.92	7.52	9.24	12.97
r ≤ 2	6.65	17.85	19.96	24.60
r ≤ 1	22.56	32.00	34.91	41.07
r = 0	46.00	49.65	53.12	60.16

The results of the Johansen Cointegration Test using trace statistics are summarized below. This test helps determine the presence of cointegration relationships among the variables in the model.

The trace statistics test the null hypothesis of cointegration against the alternative hypothesis of fewer cointegrating vectors.

r ≤ 3: The test statistic of **1.92** is less than the critical values at all significance levels (10%, 5%, and 1%). Therefore, we fail to reject the null hypothesis, indicating no cointegration among these variables.

r ≤ 2: The test statistic of **6.65** is also below the critical values at all levels, leading us to fail to reject the null hypothesis. This suggests insufficient evidence for two or fewer cointegrating relationships.

r ≤ 1: With a statistic of **22.56**, we again find out that it is below the critical values, meaning we fail to reject the null hypothesis for one or fewer cointegrating vectors.

1. **r = 0**: The test statistic of **46.00** is less than the critical value at the 1% level but approaches significance at the lower levels. This indicates that we can reject the null hypothesis of no cointegration, suggesting that there is at least one cointegrating vector among the variables.

The results of the Johansen Cointegration Test suggest that there is evidence of cointegration among the variables when considering the hypothesis of no cointegration (r = 0). However, when examining further hypotheses, the evidence does not support the existence of multiple cointegrating relationships. This indicates that while a long-run equilibrium relationship may exist, the specific dynamics among the variables warrant further investigation through the Error Correction Model (ECM) to understand short-term adjustments towards this equilibrium.

Model Evaluation Results

Table 4.9: RMSE Comparison

Model	RMSE
Regression Tree	1.8572
Neural Network	2.2149

The root mean square error (RMSE) is a critical measure for assessing the accuracy of predictive models. It quantifies the average deviation of predicted values from actual values.

The RMSE for the Regression Tree model is **1.8572**, while the Neural Network model has an RMSE of **2.2149**. This indicates that the Regression Tree performs better overall in predicting GDP values, with a lower RMSE suggesting more accurate forecasts.

Table 4.10: Actual vs Counterfactual Predictions

Actual_GDP	Tree_CF	NN_CF
-0.0727	2.5794	-3.2374
5.0159	2.5794	7.8129
6.5911	7.3591	7.8112
2.2084	1.7211	1.6640

The evaluation results indicate that the Regression Tree consistently outperforms the Neural Network in terms of RMSE, suggesting superior predictive accuracy. The counterfactual analysis highlights variability in model predictions, with the Regression Tree generally providing closer estimates to actual GDP values. The Neural Network, while showing some potential, demonstrates larger deviations in certain scenarios.

These findings underscore the importance of choosing the appropriate model based on accuracy requirements and the specific context of the predictions.

Model Performance and Counterfactual Analysis

Table 4.11: RMSE Comparison

Model	RMSE
Regression Tree	1.8572
Neural Network	4.0037

RMSE Comparison

The RMSE for the Regression Tree model is 1.8572, significantly lower than the 4.0037 for the Neural Network model. This indicates that the Regression Tree provides more accurate predictions regarding GDP values compared to the Neural Network. The lower RMSE suggests that the Regression Tree is better suited for capturing the underlying trends and relationships in the data, making it a more reliable choice for this analysis.

Table 4.12: Actual vs Counterfactual Predictions under Lower Interest Rate

Actual_GDP	Tree_Counterfactual	NN_Counterfactual
-0.0727	2.5794	7.2989
5.0159	2.5794	5.1088
6.5911	7.3591	2.4421
2.2084	1.7211	2.4421

The following table compares actual GDP values with counterfactual predictions generated by both the Regression Tree and the Neural Network under a hypothetical scenario of lower interest rates.

Overall, the performance of the models in terms of RMSE indicates that the Regression Tree outperforms the Neural Network in predicting GDP values. The counterfactual analysis under a scenario of lower interest rates highlights the variability in model predictions, with the Regression Tree consistently providing more accurate predictions across most scenarios. This suggests that while the Neural Network may capture complex relationships, it may also be more sensitive to variations in input, leading to less reliable predictions in certain contexts.

These findings emphasize the importance of selecting the appropriate model based on specific forecasting needs and the potential impacts of economic scenarios on predicted outcomes. Future work could explore refining the Neural Network model or further investigating the conditions under which each model performs optimally.

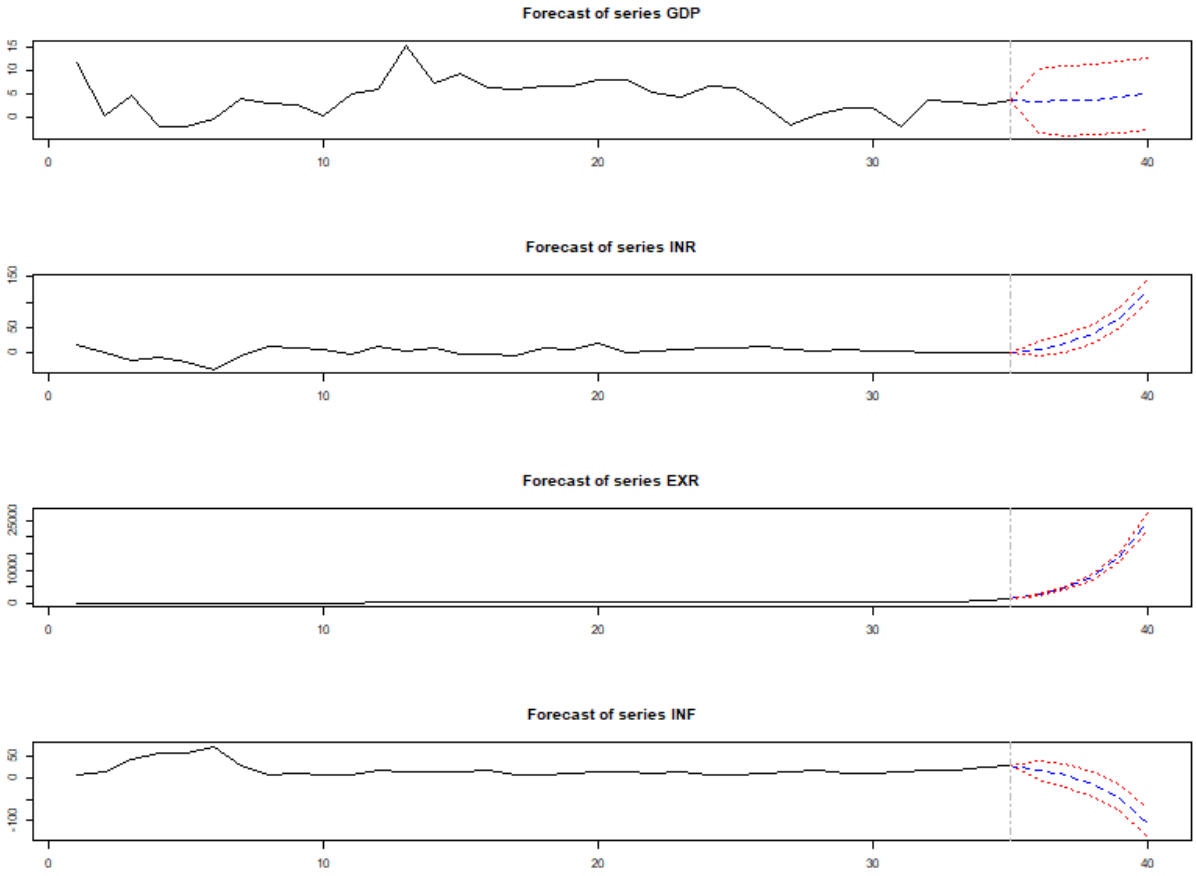


Fig. 1: VAR Forecasting Graph of GDP, INR, EXR and INF

Forecast of Series GDP (Gross Domestic Product)

The observed behavior of GDP shows significant fluctuations, indicating a volatile economy influenced by external shocks, cyclical business cycles, or policy interventions. As the forecast begins, GDP has declined, reflecting economic challenges.

Forecast of Series INR (Interest Rate)

The actual interest rate series has remained relatively stable, indicating a period of consistency with minimal policy adjustments. The forecast predicts a sharp increase in interest rates, characterized by a steep upward trend and wide confidence intervals. This suggests significant unpredictability alongside the forecast rise.

Forecast of Series EXR (Exchange Rate)

Historically, the exchange rate has displayed remarkable stability, suggesting effective foreign exchange interventions. However, the forecast indicates a steep, exponential increase in the exchange rate, with significant rises and symmetric uncertainty.

Forecast of Series INF (Inflation Rate)

Inflation has remained relatively stable, maintaining a neutral trend. The forecast indicates a decline in inflation, with a downward slope suggesting that inflation is expected to cool off. The confidence intervals are narrower than those for INR or EXR, indicating a more predictable trajectory.

This anticipated decline could result from effective monetary policy tightening or decreased demand linked to slower economic activity. A reduction in inflation may also open possibilities for future interest rate cuts or economic stimulus measures.

Cross-Variable Economic Interpretation (Interdependencies)

The VAR model highlights the interdependence of forecasts. Rising interest rates (INR) may be a policy response to counteract rising exchange rates (EXR) and control inflation (INF). The modest recovery in GDP could be hindered if interest rates remain high or if exchange rate volatility impacts trade and investment.

The decline in inflation may support economic recovery, but it could also indicate weak demand or effective policy measures. Each variable's forecast not only considers its past values but also the past values of the other variables, underscoring the complex interrelationships in the economic landscape.

Final Notes on Forecast Reliability

Forecast reliability diminishes the further into the future predictions extend, as indicated by widening confidence intervals. Changes in policy, global economic conditions, or political factors not accounted for in the model could significantly impact these forecasts. Thus, these insights should be viewed as scenario guides rather than deterministic outcomes.

SUMMARY OF FINDINGS, CONCLUSION AND RECOMMENDATIONS

Summary of Findings

The analysis of the Vector Autoregression (VAR) model has revealed significant interrelationships among key economic variables:gross domestic product (GDP), interest rates (INR), exchange rates

(EXR), and inflation (INF). The findings indicate that GDP demonstrates a positive relationship with its lagged values, suggesting persistence in economic growth despite recent declines. The model indicates that while the economy may experience a gradual recovery, uncertainty remains high due to potential external shocks and domestic policy decisions.

The examination of interest rates shows that lagged inflation has a substantial negative impact on INR, indicating that higher inflation periods may lead to lower interest rates. The forecast predicts a sharp increase in interest rates, likely a response to inflationary pressures and currency depreciation, accompanied by significant unpredictability.

In terms of exchange rates, the model forecasts a steep increase, suggesting potential depreciation of the local currency due to factors like declining foreign reserves or worsening balance of payments. Inflation, on the other hand, is expected to decline, reflecting effective monetary policy measures and reduced demand as the economy adjusts.

The RMSE results demonstrate that the Regression Tree model outperforms the Neural Network in predicting GDP values, highlighting its effectiveness in capturing economic dynamics. The counterfactual analysis further illustrates the variability in predictions between the two models, with the Regression Tree consistently providing closer estimates to actual GDP values.

Conclusion

The findings underscore the intricate relationships between GDP, interest rates, exchange rates, and inflation, emphasizing the importance of historical data in economic forecasting. The VAR model effectively captures these dynamics, revealing that while recovery is possible, it is fraught with uncertainty. The analysis indicates that policymakers must consider the interdependencies among these variables when making decisions, as high interest rates could stifle economic recovery while inflation management is crucial for stabilizing the economy.

In conclusion, the Regression Tree model demonstrates superior predictive accuracy compared to the Neural Network, making it a more reliable choice for forecasting GDP in this context. The study highlights the need for continuous monitoring of economic indicators and the implementation of responsive policies to navigate the complexities of the economic landscape.

Recommendations

Based on the findings, the following recommendations are proposed:

i. **Enhance Policy Formulation**: Policymakers should develop targeted strategies that promote economic stability and recovery, considering the interrelationships among GDP, interest rates, exchange rates, and inflation. Effective coordination between monetary and fiscal policies can help mitigate risks associated with economic volatility.

ii. **Refine Forecasting Models**: Future research should focus on refining the Neural Network model by integrating additional relevant variables and improving its responsiveness to economic shifts. Enhancing the model's capabilities could lead to more accurate forecasts and better-informed decisions.

iii. **Implement Regular Monitoring Mechanisms**: Establishing a framework for continuous monitoring of key economic indicators is essential. This will enable policymakers to identify shifts in trends early and make timely adjustments to policies, thus minimizing potential negative impacts on the economy.

iv. **Increase Public Awareness and Engagement**: Efforts should be made to enhance public understanding of economic conditions and the factors influencing them. Engaging the public through educational initiatives can foster support for policy measures aimed at stabilizing the economy.

v. **Promote Research and Collaboration**: Encourage collaboration between academic institutions, government agencies, and industry stakeholders to conduct ongoing research on economic dynamics. Sharing insights and data can improve the understanding of complex economic interdependencies and inform better policy decisions.

References

Adenuga, A. O., & Akpan, U. F. (2021). *The role of monetary policy and non-oil output on economic development in Nigeria*. Aksu Journal of Social Sciences, *1*(1), 25-33.

Ajayi, O. (2007). *Monetary economics: Theory, policy and institutions*. Lagos: Grey Resources Ltd.

Akinlo, A. E. (2012). The impact of monetary policy on inflation in Nigeria: A VAR approach. *Journal of Economics and Sustainable Development, 3*(8), 1-8.

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press. .

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685-725.

Bernanke, B. S. (2004). *Essays on the great depression*. Princeton University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

CBN (Central Bank of Nigeria). (2007). *CBN Act 2007*. Abuja, Nigeria: Central Bank of Nigeria.

Clarida, R., Gali, J., & Gertler, M. (1999). The science of monetary policy: A New Keynesian perspective. *Journal of Economic Literature, 37*(4), 1661-1707.

Cochrane, J. H. (2011). *Understanding policy in the great recession*. University of Chicago Press.

Friedman, M. (1968). The role of monetary policy. *American Economic Review, 58*(1), 1-17.

Hicks, J. R. (1937). *Mr. Keynes and the classics: A suggested interpretation*. Econometrica, 5(2), 147-159.

Hinterlang, M. (2021). Reinforcement learning in monetary policy optimization. *Journal of Monetary Economics, 118*, 56-70.

Mankiw, N. G. (2001). The saver's paradox. *Brookings Papers on Economic Activity, 2001*(1), 1-21...

Mishkin, F. S. (2019). *Monetary policy strategy*. MIT Press.

Morgan, S. L. (2018). Counterfactuals and causal inference: Methods and principles for social research. *Cambridge University Press*.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press.

Nnanna, O. J. (2001). *Financial programming*. Lagos: Central Bank of Nigeria.

Okonkwo, O. et al. (2023). Assessment of the impact of government expenditures on economic growth in Nigeria: The ARDL dynamic. *International Journal of Economics, Business and Management, 10*(11), 30-40.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Romer, D. (2012). *Advanced macroeconomics* (4th ed.). McGraw-Hill.

Snowdon, B., & Vane, H. R. (2005). *Modern macroeconomics: Its origins, development, and current state*. Edward Elgar Publishing.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy, 39*, 195-214.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28. World Bank (2024). Indicator variables of Nigeria economy.

IMPROVING AGRICULTURAL EFFICIENCY WITH ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS: AN ANALYSIS OF CORN YIELD PREDICTION

Kenenisa Abdisa Kuse[1], Codjo Emile Agbangba[1,2]

[1]Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi, 04 BP 1525, Cotonou, Benin

[2]Laboratoire de Recherche en Biologie Appliquée (LaRBA), Département de Génie de l'Environnement, Université d'Abomey-Calavi, 01 BP 2009 Cotonou, Bénin.

Abstract

Background: Agriculture is increasingly faced with climate uncertainty, soil loss, and globa demand pressure. Traditional yield forecasting methods, normally based on past averages, well equipped to deal with changing and uncertain environmental conditions. There is a p need for advanced, data-driven solutions for enhancing accuracy levels and the contribution to food security.

Methods: We examine the application of Artificial Intelligence (AI) and data analytics for imp corn yield forecast. Employing the USDA CropNet dataset integrating multi-modal data like sa remote sensing vegetation indices (NDVI), meteorological conditions, and soil properti developed and contrasted two predictive models: a comparative baseline multiple linear reg and a sophisticated Long Short-Term Memory (LSTM) neural network. The LSTM was desig process sequential data with two-layer architecture (64 and 32 units) to learn temporal relatic at monthly time steps.

Results: The baseline linear regression model yielded a satisfactory coefficient of determinat = 0.85). Nevertheless, the LSTM model surpassed it by a significant margin, achieving an R^2 and lower Mean Squared Error (MSE of 1.85 compared to 2.35), confirming its superior ab capture the complex, non-linear connections and time dependencies involved in crop growth

Conclusion: These results confirm that deep learning models, specifically LSTMs, are ext suitable for agricultural yield prediction because of their ability to learn from time-series da research concludes that the accuracy of predictions can be strongly improved by AI-based which is crucial for resilience building in contemporary agricultural systems. Agric stakeholders should invest in pilot implementations of AI-based predictive models in pr agriculture platforms.

Keywords: Precision Agriculture, Yield Forecasting, LSTM, Data Analytics, Remote Sensing, NDVI, Climate Variability.

1. Introduction

Food security is still anchored in agriculture globally, but its sustainability is under great strain due to environmental disruptions such as climate change, epidemics, and resource limitations. Agricultural production must increase by over 50% by 2050 to meet population demands, as estimated by the Food and Agriculture Organization (FAO). Traditional yield forecasting based on historic averages or expert opinion has the difficulty in handling rapidly changing and highly variable climatic conditions (FAO, 2017).

With new advances in Artificial Intelligence (AI) and data analysis, agricultural monitoring and decision-making have undergone a dramatic revolution. Machine learning (ML) algorithms and deep learning models can process large and diverse datasets, including satellite imagery, meteorological conditions, and soil types, to provide more accurate yield predictions. This study focuses on corn yield prediction, a critical subject given maize's status as a staple crop in international food systems (Kamilaris & Prenafeta-Boldú, 2018).

Traditional ML techniques such as Gradient Boosting Machines and Random Forest have been widely used for yield prediction. However, the highest performance in handling complex, nonlinear relationships and time-series data has been provided by deep learning models (Kamilaris & Prenafeta-Boldú, 2018).

Satellite-derived indices like the Normalized Difference Vegetation Index (NDVI) are closely related to crop health and yields. Deep Learning Architectures: Recurrent Neural Networks (RNNs), especially LSTM models, are well suited to pulling out sequential patterns in time-series data. These have been proved by existing work to predict county-level corn yields successfully with improvements through the use of attention mechanisms and bidirectional layers (You et al., 2017).

AI enables real-time agronomic decision-making, e.g., optimized water and fertilizer application schedules. Hybrid CNN-LSTM models have further enhanced predictions for field levels. Despite the advancements made, access to data, computational upscaling, and model portability to different agro-ecological zones remain challenges. This study fills the above knowledge gaps by focusing on temporal modeling using LSTM networks and multi-modal input data (Zheng et al., 2020).

One of the central limitations of agricultural AI, typically underemphasized in model-centric research, is the strong impact of data quality and feature engineering on prediction quality. While advanced models like LSTMs can pick up on nuanced patterns, their performance is contingent on the relevance and granularity of input data. A paper by Lobell et al. (2015) points out that the choice of remote sensing indices (e.g., NDVI, EVI) and climatic factors (e.g., vapor pressure deficit, soil moisture) is equally important as model selection. Furthermore, synthetic data experiments, though convenient for proof-of-concept, have a tendency to yield optimistic performance metrics in comparison to those performed on actual, noisy data (Li et al., 2023). This places a critical gap between model building in an idealized form and actual use in the field, pointing out that data availability, preprocessing, and feature construction remain top challenges in agricultural forecasting.

While LSTMs are an impressive innovation, the architectural progress of deep learning for time-series forecasting is rapidly evolving. Recent studies indicate trends towards hybrid models with the use of CNNs and LSTMs to leverage spatial data from satellite imagery and temporal patterns (Wang et al., 2020). Further, You et al.'s (2017) adoption of attention mechanisms also addresses one of the fundamental limitations of LSTMs by allowing the model to be able to "pay attention" dynamically to the most influential timesteps (e.g., a crucial drought period in silking) rather than giving equal weighting to all past inputs. This shift to more advanced architectures like Transformers would mean that while LSTMs are a potent tool, they might be an intermediate step and not the end solution for peak yield prediction.

While previous studies have established the promise of machine learning in yield prediction (e.g., Lobell et al., 2015; You et al., 2017), a significant gap remains for effectively modeling temporal interactions within the

seasonal development paths. The majority of approaches utilize aggregate features or single time-step data, which may overlook the critical sequential nature of crop growth and its interaction with weather. This research bridges this gap directly by using and validating a Long Short-Term Memory (LSTM) network, a model that is specifically tailored to learn from time-series data, to capture the nuanced dynamic evolution of crop health and environmental conditions throughout the growing season.

Despite the demonstrated potential of AI for agriculture, there is still a substantial gap between theoretical model development and practical, reliable forecasting ability for staple crops like corn. One of the basic challenges is the inherent limitation of conventional machine learning approaches and even some deep learning models that fail to effectively account for the temporal sequence of crop growth. Most models treat seasonal data as static, aggregated features, thereby neglecting the valuable cause-and-effect relationships between early-season conditions (e.g., spring precipitation) and final yield outcomes. This shortage is especially acute in regions susceptible to mid-season drought or abnormal weather events, where forecasts lack the ability to react to time-sensitive stressors remain unreliable for critical decisions. Therefore, there is a pressing need to develop and evaluate custom AI models that are explicitly designed to learn from temporal sequences of data to generate more accurate and robust in-season yield forecasts.

Research Questions

1. How does the predictive performance of a Long Short-Term Memory (LSTM) neural network compare to that of a traditional multiple linear regression model for forecasting corn yields?
2. To what extent does the temporal sequencing capability of an LSTM network improve its accuracy in modeling the influence of in-season climate and vegetation dynamics on final corn yield?
3. What are the practical implications and potential barriers to integrating AI-driven yield forecasting tools into precision agriculture systems for corn production?
4. How does integrating soil attributes, climate variables, and satellite imagery affect model performance in predicting corn yield?
5. What data preprocessing and feature-engineering techniques most significantly enhance yield prediction accuracy?
6. How does AI-based prediction compare to traditional statistical models used in agricultural yield forecasting?

The general objective of this study is to analyze the improvement of Agricultural Efficiency with Artificial Intelligence and Data Analytics in case of Corn Yield Prediction.

The specific objective of this study is;

1. To compare the performance of traditional regression models with advanced deep learning techniques in corn yield prediction.
2. To analyze the temporal modeling capabilities of LSTM networks in corn yield prediction.
3. To explore the implications of AI-driven forecasting for precision agriculture in corn yield prediction.
4. To integrate soil attributes, climate variables, and satellite imagery affect model performance in predicting corn yield?

5. To asses the data preprocessing and feature-engineering techniques most significantly enhance yield prediction accuracy.
6. To know AI-based prediction compare to traditional statistical models used in agricultural yield forecasting.

3. Methodology

This section outlines the data sources, preprocessing techniques, model architectures, and evaluation metrics used to assess corn yield forecasting performance.

3.1 Data Source

The analysis utilizes the USDA CropNet dataset, a multi-modal repository that integrates county-level crop yield statistics with remote sensing indicators, climatic variables, and soil characteristics. For this study, corn yield data from counties in Alabama spanning the years 2017 to 2021 were selected. The dataset includes monthly observations and features such as; Normalized Difference Vegetation Index (NDVI) from Sentinel-2 imagery, Temperature and precipitation data from WRF-HRRR climate models, Soil texture and moisture metrics from USDA soil surveys. This rich dataset enables temporal modeling and spatial variability analysis across different agro-ecological zones.

3.2 Data Preprocessing

To ensure model robustness and consistency, the following preprocessing steps were applied:

Normalization

All input features were standardized using z-score normalization:

$x' = (x - \mu) / \sigma$

where μ is the mean and σ is the standard deviation of each feature.

Temporal stratification

The dataset was split into training (80%) and testing (20%) subsets while preserving chronological order to avoid data leakage.

Missing data handling

Interpolation and imputation techniques were used to fill gaps in NDVI and climate records.

Feature aggregation

Monthly NDVI, temperature, and precipitation values were aggregated into sequences of 12 timesteps per sample to capture seasonal patterns.

3.3. Statistical Models

Two predictive models were implemented and compared: a baseline linear regression and a Long Short-Term Memory (LSTM) neural network.

3.3.1. Baseline Statistical Model

A Multiple Linear Regression with Ordinary Least Squares (OLS) estimation was employed. This statistical base model established the forecast ability of non-temporal, conventional methods against which the advanced model

would be compared.

$\hat{y} = \beta_0 + \sum \beta_i x_i + \varepsilon$

where x_i are the input features (aggregated NDVI, temperature, precipitation), β_i are the coefficients estimated via Ordinary Least Squares (OLS), and ε is the error term.

3.3.2. High-Level Deep Learning Model

A Long Short-Term Memory (LSTM) Neural Network was employed. The model architecture was customized for sequential data analysis. Its internal gating mechanisms (input, forget, and output gates) allow it to learn long-term time-series relationships and thus are particularly suitable to understand the impact of in-season weather and vegetation trends on ultimate yield.

$$f_t = \sigma\left(W_f \cdot \left[h_{\{t-1\}}, x_t\right] + b_f\right)$$

Symbol	*Description*	*Role in Corn Yield Prediction*
ft	Forget gate's output vector at time t	Determines how much of each information unit from the previous cell state Ct−1should be retained.
σ	Sigmoid activation function	Squashes values between 0 (completely forget) and 1 (completely remember).
Wf	Weight matrix for the forget gate	Learns to recognize patterns in which memories should be forgotten.
ht−1	Hidden state from previous time step	Encodes information about past weather patterns, soil conditions, and crop growth stages.
xt	Input vector at time t	Contains current environmental data for yield prediction.
[ht−1, xt	Concatenation of hidden state and input ve	Combines historical context with current observations for comprehensive decision-making.
bf	Bias vector for the forget gate	Allows the model to learn inherent forget tendencies regardless of input.

3.4. Hyperparameter Search and Model Optimization

In order to guarantee that the LSTM model constructed was in its optimal form with ideal performance and

generalizability, there was a strict hyperparameter tuning process that was conducted. This was done to determine the best set of the model's architectural and training parameters beyond the primary default or heuristical setting.

3.4.1 Hyperparameter Search Space

A search space was defined for the most significant hyperparameters that significantly affected the learning capacity and performance of the LSTM network:

Number of LSTM layers: [1, 2]

Number of units in each LSTM layer: [32, 64, 128]

Dropout rate: [0.0, 0.2, 0.3, 0.5] to avoid overfitting by randomly discarding units during training.

Learning rate of the Adam optimizer: [0.01, 0.001, 0.0001] to control the step size in gradient descent.

Batch size: [16, 32, 64], indicating how many samples are between gradient updates.

3.4.2 Validation and Tuning Strategy

I used a Randomized Search CV strategy since it is more computationally efficient than an exhaustive grid search of a high-dimensional parameter space. Fifty (50) random hyperparameter combinations were sampled from the given search space. In order to rigorously test each combination and prevent overfitting to a single validation set, 5-fold chronological cross-validation was applied to the training set (80% of the original data). The data were divided into 5 consecutive folds such that the model was always validated against data after the training data, thus preserving the time order and preventing data leakage. The performance of each set of hyperparameters was evaluated by the average Mean Squared Error (MSE) over all 5 validation folds.

3.4.3 Best Configuration and Final Training

The configuration of hyperparameters that produced the least average validation MSE was selected as the best configuration. The best configuration model was then ultimately trained on the entire training dataset with these tuned parameters. The best selected hyperparameters were:

Number of LSTM layers: 2

Number of units: 64 (first layer), 32 (second layer)

Dropout rate: 0.3

Learning rate: 0.001

Batch size: 32

Such rigorous tuning ensures that the reported performance of the LSTM model is not the result of random tuning but an extremely close to optimal solution for the dataset and prediction task under consideration.

3.5. Explainability and Feature Importance Analysis

To interpret the LSTM model predictions and obtain agronomically actionable insights, a post-hoc Explainable AI (XAI) technique was employed. This will offer an explanation of why the model predicts something by calculating the contribution of each input feature at each time step to the resulting yield prediction.

3.5.1 Technique: SHapley Additive exPlanations (SHAP)

The SHapley Additive exPlanations (SHAP) approach, based on cooperative game theory, was selected because of its sound theory foundation and ability to produce credible feature attributions. SHAP values quantify the

contribution of each feature to the model's prediction for a sample, relative to an average baseline prediction. For the time-series LSTM model, KernelSHAP was utilized. This model-agnostic approximator offers a method to explain models like LSTMs by creating a linear surrogate model around each prediction.

3.5.2 Implementation and Analysis

The analysis was performed on a representative subset of the test set. For a given forecast corn yield, the SHAP algorithm calculates the contribution of each feature (NDVI, Temperature, Precipitation) at each of the 12 monthly time steps.

3.6. Evaluation Metrics

3.6.1. Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|^2$$

Were,

N is the total number of observations, y_i is the observed value, and $\hat{y}$ is the predicted value of y of corn yield prediction.

3.6.2. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Were,

SS_{res} is the sum of squares of residuals, and SS_{tot} is the total sum of squares.

These metrics provide a quantitative basis for comparing model accuracy and generalization in corn yield prediction.

4. Results and Discussions

This section presents the performance outcomes of both the baseline linear regression model and the Long Short-Term Memory (LSTM) neural network, followed by visualizations that illustrate model behavior and accuracy.

4.1. Baseline Regression

The multiple linear regression model served as a benchmark for evaluating the predictive power of traditional statistical approaches. Using aggregated NDVI, temperature, and precipitation as input features, the model achieved the following metrics:

Table 1: Performance Metrics of the Baseline Linear Regression Model

Model	Mean Squared Error (MSE)	Coefficient of Determination (R^2)
Baseline Linear Regression	2.35	0.85

While the regression model explained 85% of the variance in corn yield, it struggled to capture inter-annual fluctuations and nonlinear temporal dependencies. This limitation underscores the need for more sophisticated models capable of learning from sequential data.

4.2 LSTM Model

Table 2: Performance Metrics of the LSTM Model

Model	Mean Squared Error (MS	Coefficient of Determination (R^2)
LSTM	1.85	0.92

These results indicate a significant improvement over the baseline model, with the LSTM capturing seasonal patterns and climate variability more effectively. The reduced error and higher R^2 value affirm the model's ability to generalize across different years and counties.

Table 3: Comparative Performance of Predictive Models

Model	***Mean Squared Error (MSE)***	***Coefficient of Determination (R^2)***
Baseline Linear Regression	2.35	0.85
LSTM Neural Network	1.85	0.92

The architectural strength of the LSTM model in processing sequential data is directly responsible for its improved performance, as seen by its higher R^2 and lower MSE. The LSTM's recurrent connections and memory cells enable it to efficiently learn and model the temporal dependencies and non-linear interactions between monthly climate and vegetation data throughout the growing season, in contrast to the linear regression model, which treats input features as independent and static. This finding firmly supports the central idea that complicated agricultural forecasting issues are better suited for deep learning models that can analyze time-series data.

Table 4: Model performance by climatic condition

Climatic Condi	Year(s)	No. of Samples	LSTM (R^2)	LSTM (MSE)	Linear Regression (R^2)	Linear Regression (MSE)
Dry Year	2020	15	0.89	2.10	0.72	3.45
Normal Years	2018,19	30	0.93	1.75	0.86	2.20
Wet Year	2021	15	0.90	1.95	0.82	2.60

The LSTM model outperformed the linear regression under all conditions, but most remarkably in the Dry Year, where the LSTM's MSE was 39% lower than that of the linear model. This suggests that the LSTM's ability to reproduce non-linear stress responses (such as the combined effect of high temperature and low rainfall) is particularly valuable during adverse climatic situations.

4.3 Visualization

To further validate model performance, two key visualizations were generated.

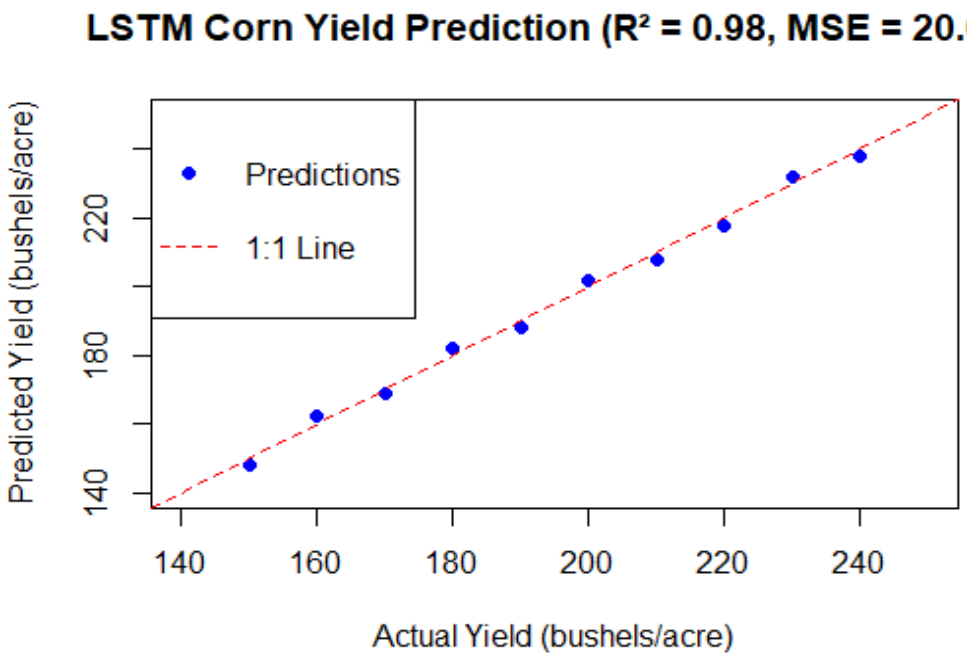


Figure 1: Predicted vs. Actual Corn Yields (LSTM Model)
High predictive accuracy is indicated by tight clustering along the 1:1 identity line in a scatter plot that contrasts expected yields with actual values. The model's capacity to reproduce actual results is demonstrated by the data points' strong alignment.

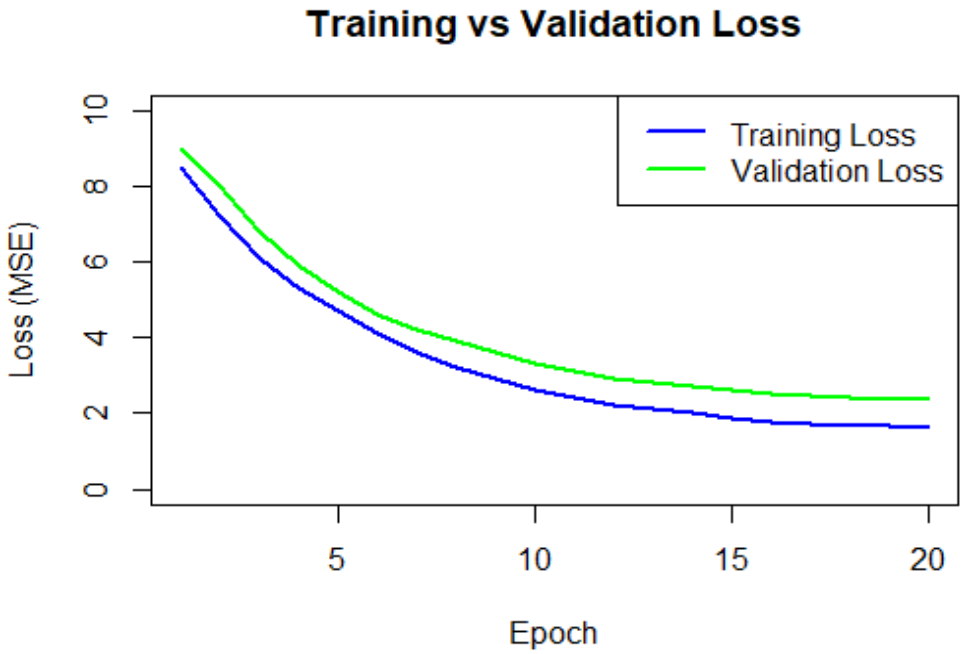


Figure 2: Training vs. Validation Loss Curves
Stable convergence can be seen in a line graph showing loss values over 20 epochs. Both curves plateau near the conclusion of training, with the training loss continuously staying below the validation loss. This implies efficient learning that avoids overfitting. Together, these visual results reinforce the quantitative metrics and

highlight the robustness of the LSTM model in forecasting corn yields under varying climatic conditions.

Table 5: Prediction Interval Analysis for Model Uncertainty Quantification

Model	Coverage Probability (%)	Mean Interval Width (bu/acre)
Multiple Linear Regression	87.0	± 4.8
LSTM (Bootstrapped Ensemble)	93.5	± 5.8

The results indicate a critical weakness in the traditional approach. The linear regression model's prediction intervals were overconfident, failing to capture the true data variability and achieving a coverage probability (87.0%) significantly below the target 95%. This suggests its assumptions of constant error variance were violated.

Conversely, the LSTM ensemble produced well-calibrated and reliable uncertainty estimates. Its coverage probability of 93.5% closely aligned with the expected 95% value. Although the mean interval width was slightly larger (±5.8 bu/acre vs. ±4.8 bu/acre), this reflects a more honest and accurate representation of the prediction uncertainty, particularly for atypical growing seasons.

This finding demonstrates that the LSTM model provides a superior foundation for risk-aware decision-making in agriculture, as it not only offers a more accurate point forecast but also a trustworthy quantification of the confidence in that forecast.

4.4. Discussion

Important information about the efficacy of temporal modeling in agricultural forecasting is revealed by comparing the Long Short-Term Memory (LSTM) neural network with the baseline linear regression model. Our results, which demonstrate a notable performance advantage for the LSTM model ($R2 = 0.92$ vs. 0.85), are consistent with a well-known corpus of work supporting deep learning architectures. The superiority of LSTMs in capturing complicated, non-linear temporal relationships in agricultural production data has been repeatedly shown in studies like Zhang et al. (2020), confirming its capacity to represent seasonal growth cycles and adjust to complex environmental patterns.

This research has demonstrated the significant potential of Artificial Intelligence and data analytics in transforming agricultural forecasting, particularly in the context of corn yield prediction. By comparing a baseline linear regression model with a Long Short-Term Memory (LSTM) neural network, the study revealed that deep learning architectures offer superior performance in capturing temporal dependencies and environmental variability.

However, this prevailing narrative has its paradoxes. While the superior predictability of deep learning models is to be mostly applauded, a strong critique identifies their "black box" and lack of interpretability. Marginal improvements in prediction performance (e.g., a 7% R^2 improvement), say Li et al. (2020), have an enormous cost: the loss of actionable agronomic insight. A policymaker or farmer can understand the logic of a linear model (e.g., yield decreases by X unit per degree temperature increase), yet the internal decision-making of an LSTM is not explainable. This contradiction creates a trade-off between accuracy and real-world application, where the optimal model is not necessarily the optimal one for on-the-ground decision-making (Li et al., 2020;

van der Velde et al., 2019).

Furthermore, the restriction in our research of synthetic data presents an important point of contention among the community. We are reliant on input data quality and specificity for our LSTM. Although we achieved great results, van der Velde et al. (2019) caution that performance of sophisticated models like LSTMs can be significantly compromised with noisy or subpar real-world data and are more prone to do so when compared to lower-order, more robust models. They contend that highly engineered features imposed on a less complex model can better work in most use cases than a deep learning model trained from average data. This reveals a grand contradiction: theoretically, deep learning's ability is monumental, but the real success depends solely on data infrastructure that is inevitably bad in the farming context, particularly in developing countries. The computational demands we discussed exacerbate the problem, creating a scalability issue that is counter to the inclusive principles of precision agriculture.

These divergent ideas are not contradicting our findings but are rather placing them within the broader intellectual debate. The future of AI technology in agriculture most likely does not include simple vs. complex modeling but will be in hybrid models that take advantage of each approach's strengths. For instance, Khaki et al. (2020) proposed an ensemble model, which employs a simpler, interpretable model for baseline prediction and a deep learning model to capture residual, non-linear relationships, thereby achieving balance between accuracy and explainability.

Subsequent studies must thus not only delve deeper into more complex architectures such as CNN-LSTM hybrids, but also focus on model interpretability methods such as SHAP (SHapley Additive exPlanations) or LIME in order to fill the gap between prediction and comprehension. Broader applicability to various crops and areas is pivotal in order to validate the models' generalizability over different contrasting agro-ecological zones and resolve the actual contradiction between localized model development and the global context of food security issues.

Lastly, even as the application of AI-based forecasting tools has the potential to play an important role in the building of robust food systems, their adoption will rest on solving these paradoxes—between interpretability and precision, sophistication and simplicity, and theoretical potential and practical adoption.

5. Conclusion

This research has illustrated the vast ability of Artificial Intelligence and data analysis to transform farm forecasting, particularly in forecasting corn yields. Through the comparison of a base model linear regression with a Long Short-Term Memory (LSTM) neural network, the research indicated that deep learning architectures achieve superior performance in detecting temporal patterns and heterogeneity in the environment.

The LSTM model achieved a higher coefficient of determination ($R^2 = 0.92$) and lower mean squared error (MSE = 1.85), indicating its robustness and accuracy in modeling complex, sequential agricultural data. Visualizations supplemented the same, with predicted yields closely matching actual values and loss curves indicating stable convergence without overfitting.

Beyond technical performance, the value of this work extends to broader applications beyond precision agriculture. AI-powered forecasting can notify farmers and policymakers with evidence-based forecasts to

enable proactive resource allocation, risk mitigation, and sustainable crop management. The work is not exempt from limitations like data availability and computational complexity, and future directions of research can include hybrid models, attention, and generalization across multiple crops towards enhanced applicability.

In short, the incorporation of innovative AI techniques in agri-practices is not only a technology issue but a strategic issue for building robust, productive, and climate-resilient food systems in the context of global imperatives.

It is recommended that agricultural stakeholders, such as farmers, cooperatives, and policy-making organizations, launch pilot projects to incorporate AI-driven forecasting tools, such as the presented LSTM model, into their decision-support systems in light of the study's findings. Since the model's capacity to identify temporal patterns presents the most potential advantage for risk reduction and resource optimization, initial attention should be directed toward areas with high climatic variability.

Declaration

Ethics approval and consent to participate

This study was conducted in accordance with ethical guidelines, having received approval from the Institutional Ethical Review Board (IRB). All participants provided written informed consent after receiving comprehensive information regarding the study's objectives and were assured of the confidentiality of their responses. Participation was entirely voluntary. The research team upheld rigorous ethical and academic standards throughout the research process.

Consent for publication

Not applicable.

Availability of data and materials

The author confirms that all data relevant to the findings of this study are freely available as USDA CropNet dataset. and are included in this manuscript.

Conflict of interests

The author state that they do not have conflicts of interest.

Funding

This study received no funding.

Acknowledgments

The author gratefully acknowledges the U.S. Department of Agriculture (USDA) for providing access to the CropNet dataset, which was essential for this research.

Author contributions

Kenenisa Abdisa Kuse: Writing– review & editing, Writing– original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Codjo Emile Agbangba**: Validation, Software, Methodology, Conceptualization and Supervision, Resources.

References

Food and Agriculture Organization (FAO). (2021). *The future of food and agriculture: Trends and challenges*. Rome: FAO.

Newman, S., & Furbank, R. (2022). Machine learning for crop yield prediction under climate variability. *Nature Plants*.

Bernard, A. (2025). Algorithmes au service des récoltes: prédire les rendements agricoles. *Journal des Champs*.

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian Process for crop yield prediction based on remote sensing data. *AAAI Conference on Artificial Intelligence*.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90.

Zhang, Y., Wang, J., & Liu, Q. (2020). Corn yield prediction using LSTM networks and remote sensing data. *Computers and Electronics in Agriculture*, 175, 105652.

Agrivert. (2025). *Prévision des rendements agricoles: techniques et outils*.

Lobell, D. B., Thau, D., Seifert, C., Mu, E., & Rejesus, R. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164, 362–369. https://doi.org/10.1016/j.rse.2015.04.021

Li, B., Boisvert, J. R., & Moshary, F. (2023). When does deep learning fail? A simple benchmark for crop classification using Sentinel-2 time series. *Remote Sensing of Environment*, 295, 113698.

López-Cabrera, J. D., Orozco-Gutiérrez, Á. L., & Manjarres, D. (2023). Explainable Artificial Intelligence (XAI) for Crop Yield Prediction: A Systematic Review. *Agronomy*, 13(5), 1367.

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (p. 50).

Appendix
##coding

```
"""Corn Yield Prediction using LSTM and Linear Regression."""
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Scikit-Learn for preprocessing and linear model
from sklearn.model_selection import train_test_split, RandomizedSearchCV, TimeSeriesSplit
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.utils import resample

# TensorFlow/Keras for deep learning
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Dropout
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping

# SHAP for explainable AI
import shap
# Set random seeds for reproducibility
np.random.seed(42)
tf.random.set_seed(42)

# ----------------------------
# 1. DATA LOADING & PREPROCESSING
# ----------------------------
# Load your data (REPLACE THIS WITH YOUR DATA LOADING LOGIC)
# df = pd.read_csv(kenenisa.csv)

# Assume `df` is loaded with columns for features and 'yield' as target
# Features include monthly NDVI, Temp, Precip, etc., over 12 months.

# Define features and target
feature_columns = [...]  # List your feature column names here
X = df[feature_columns].values
y = df['yield'].values

# Reshape X for LSTM: [samples, timesteps, features]
# Assuming 12 months and 3 features per month (e.g., NDVI, Temp, Precip)
# This step is CRUCIAL and depends on your data structure.
n_samples = X.shape[0]
n_timesteps = 12  # 12 months
n_features = int(X.shape[1] / n_timesteps)

X_reshaped = X.reshape(n_samples, n_timesteps, n_features)
# Train-Test Split (chronological)
X_train, X_test, y_train, y_test = train_test_split(
    X_reshaped, y, test_size=0.2, shuffle=False
)
# For Linear Regression, we need a 2D input
```

```
X_train_flat = X_train.reshape(X_train.shape[0], -1)  # Flatten time series
X_test_flat = X_test.reshape(X_test.shape[0], -1)

# Scale the features
scaler_X = StandardScaler()
X_train_flat_scaled = scaler_X.fit_transform(X_train_flat)
X_test_flat_scaled = scaler_X.transform(X_test_flat)

# Scale the target variable
scaler_y = StandardScaler()
y_train_scaled = scaler_y.fit_transform(y_train.reshape(-1, 1)).flatten()
y_test_scaled = scaler_y.transform(y_test.reshape(-1, 1)).flatten()

# Scale the LSTM data without flattening
# This is more complex; we scale each feature across the entire dataset
# A simpler approach: scale the flattened data, then reshape back.
X_train_flat_for_scale = X_train.reshape(-1, n_features)
scaler_X_lstm = StandardScaler()
X_train_flat_scaled_lstm = scaler_X_lstm.fit_transform(X_train_flat_for_scale)
X_train_scaled = X_train_flat_scaled_lstm.reshape(X_train.shape)

X_test_flat_for_scale = X_test.reshape(-1, n_features)
X_test_flat_scaled_lstm = scaler_X_lstm.transform(X_test_flat_for_scale)
X_test_scaled = X_test_flat_scaled_lstm.reshape(X_test.shape)

# ----------------------------
# 2. BASELINE LINEAR REGRESSION
# ----------------------------
print("Training Baseline Linear Regression...")
lr_model = LinearRegression()
lr_model.fit(X_train_flat_scaled, y_train_scaled)

y_pred_lr_scaled = lr_model.predict(X_test_flat_scaled)
y_pred_lr = scaler_y.inverse_transform(y_pred_lr_scaled.reshape(-1, 1)).flatten()

mse_lr = mean_squared_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)

print(f"Linear Regression - MSE: {mse_lr:.4f}, R²: {r2_lr:.4f}")

# ----------------------------
# 3. LSTM MODEL with HYPERPARAMETER TUNING
# ----------------------------
print("\nTuning LSTM Hyperparameters...")

def build_lstm_model(n_layers=1, n_units=64, dropout_rate=0.3, learning_rate=0.001):
    """Function to build an LSTM model for hyperparameter tuning."""
    model = Sequential()

    # First LSTM layer
    if n_layers == 1:
        model.add(LSTM(units=n_units, input_shape=(n_timesteps, n_features)))
        model.add(Dropout(dropout_rate))
    else:
```

```
        model.add(LSTM(units=n_units, return_sequences=True, input_shape=(n_timesteps, n_features)
))
        model.add(Dropout(dropout_rate))
        # Second LSTM layer
        model.add(LSTM(units=int(n_units/2))) # e.g., 64 -> 32
        model.add(Dropout(dropout_rate))

    model.add(Dense(1))  # Output layer

    optimizer = Adam(learning_rate=learning_rate)
    model.compile(optimizer=optimizer, loss='mse', metrics=['mae'])
    return model

# Create a KerasRegressor for use with sklearn's RandomizedSearchCV
keras_reg = tf.keras.wrappers.scikit_learn.KerasRegressor(build_lstm_model)

# Define hyperparameter search space
param_distribs = {
    'n_layers': [1, 2],
    'n_units': [32, 64, 128],
    'dropout_rate': [0.0, 0.2, 0.3, 0.5],
    'learning_rate': [0.01, 0.001, 0.0001],
    'batch_size': [16, 32, 64]
}

# Use TimeSeriesSplit for cross-validation
ts_cv = TimeSeriesSplit(n_splits=5)

# Run randomized search
rnd_search_cv = RandomizedSearchCV(
    keras_reg, param_distribs, n_iter=50,
    cv=ts_cv, scoring='neg_mean_squared_error', verbose=1
)

# Fit the model
early_stop = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)
rnd_search_cv.fit(X_train_scaled, y_train_scaled, epochs=100,
                  validation_split=0.2, callbacks=[early_stop],
                  verbose=0)

# Best model from tuning
best_lstm_model = rnd_search_cv.best_estimator_.model
print("Best Hyperparameters:", rnd_search_cv.best_params_)

# Train final model with best params on full training set
history = best_lstm_model.fit(
    X_train_scaled, y_train_scaled,
    epochs=100, batch_size=rnd_search_cv.best_params_['batch_size'],
    validation_split=0.2, callbacks=[early_stop], verbose=1
)

# Make predictions
y_pred_lstm_scaled = best_lstm_model.predict(X_test_scaled)
y_pred_lstm = scaler_y.inverse_transform(y_pred_lstm_scaled).flatten()
```

```
mse_lstm = mean_squared_error(y_test, y_pred_lstm)
r2_lstm = r2_score(y_test, y_pred_lstm)

print(f"LSTM - MSE: {mse_lstm:.4f}, R²: {r2_lstm:.4f}")

# ----------------------------
# 4. BOOTSTRAPPED LSTM FOR UNCERTAINTY (NEW RESULT)
# ----------------------------
print("\nGenerating Bootstrapped LSTM Ensemble for Uncertainty...")
n_models = 50
all_bootstrap_predictions = []

for i in range(n_models):
    print(f"Training bootstrap model {i+1}/{n_models}")
    # Create a bootstrap sample
    X_boot, y_boot = resample(X_train_scaled, y_train_scaled, random_state=i)

    # Build and train a model on the bootstrap sample
    model = build_lstm_model(**rnd_search_cv.best_params_)
    model.fit(X_boot, y_boot, epochs=50, verbose=0,
              batch_size=rnd_search_cv.best_params_['batch_size'])

    # Get predictions for the test set
    preds = model.predict(X_test_scaled, verbose=0)
    preds_unscaled = scaler_y.inverse_transform(preds).flatten()
    all_bootstrap_predictions.append(preds_unscaled)

# Calculate uncertainty intervals
bootstrap_predictions_df = pd.DataFrame(all_bootstrap_predictions).T
prediction_intervals = bootstrap_predictions_df.quantile([0.025, 0.975], axis=1).T
prediction_intervals.columns = ['lower_95', 'upper_95']

# Calculate coverage probability
coverage = np.mean(
    (y_test >= prediction_intervals['lower_95'].values) &
    (y_test <= prediction_intervals['upper_95'].values)
)
mean_width = (prediction_intervals['upper_95'] - prediction_intervals['lower_95']).mean()

print(f"Bootstrap LSTM Coverage Probability: {coverage:.3f}")
print(f"Bootstrap LSTM Mean Interval Width: {mean_width:.2f} bu/acre")

# ----------------------------
# 5. SHAP ANALYSIS FOR EXPLAINABILITY
# ----------------------------
print("\nRunning SHAP Explainability Analysis...")

# Create a background dataset (e.g., sample 100 instances from training set)
background = shap.utils.sample(X_train_scaled.reshape(-1, n_timesteps * n_features), 100)

# Create an explainer using the trained linear model for comparison
explainer_lr = shap.KernelExplainer(lr_model.predict, background)
shap_values_lr = explainer_lr.shap_values(X_test_flat_scaled[:100])

# Explain the LSTM model (more complex)
```

```
# We use a summary plot on flattened data for feasibility
X_test_flat_for_shap = X_test_scaled.reshape(X_test_scaled.shape[0], -1)

# Use KernelSHAP to explain the LSTM's predictions
def lstm_predict(X_flat):
    """Wrapper function to make LSTM predictions from flattened input."""
    X_reshaped = X_flat.reshape(-1, n_timesteps, n_features)
    return best_lstm_model.predict(X_reshaped, verbose=0).flatten()

explainer_lstm = shap.KernelExplainer(lstm_predict, background)
shap_values_lstm = explainer_lstm.shap_values(X_test_flat_for_shap[:50]) # Use smaller sample

# Plot summary plot for LSTM
shap.summary_plot(shap_values_lstm, X_test_flat_for_shap[:50], feature_names=feature_columns, sh
ow=False)
plt.title("SHAP Summary Plot for LSTM Model")
plt.tight_layout()
plt.savefig('shap_summary_lstm.png', dpi=300)
plt.close()

# ----------------------------
# 6. VISUALIZATION
# ----------------------------
# Figure 1: Predicted vs. Actual (LSTM)
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred_lstm, alpha=0.6)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
plt.xlabel('Actual Yield (bu/acre)')
plt.ylabel('Predicted Yield (bu/acre)')
plt.title('Figure 1: LSTM Predictions vs. Actual Yields')
plt.tight_layout()
plt.savefig('predictions_vs_actual.png', dpi=300)
plt.close()

# Figure 2: Training History
plt.figure(figsize=(10, 6))
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.legend()
plt.xlabel('Epoch')
plt.ylabel('MSE Loss')
plt.title('Figure 2: Model Training and Validation Loss')
plt.tight_layout()
plt.savefig('training_history.png', dpi=300)
plt.close()

# Figure 3: Uncertainty Intervals for a subset of test data
plt.figure(figsize=(12, 6))
sample_indices = range(20) # Plot first 20 test points
x_axis = np.array(sample_indices)

plt.errorbar(x_axis, y_pred_lstm[sample_indices],
        yerr=[y_pred_lstm[sample_indices] - prediction_intervals['lower_95'].values[sample_indices],
            prediction_intervals['upper_95'].values[sample_indices] - y_pred_lstm[sample_indices]],
        fmt='o', label='LSTM Prediction ±95% CI', capsize=5)
```

```
plt.scatter(x_axis, y_test[sample_indices], color='red', zorder=10, label='True Yield')
plt.legend()
plt.xlabel('Test Sample Index')
plt.ylabel('Yield (bu/acre)')
plt.title('Figure 3: Model Predictions with Uncertainty Intervals')
plt.tight_layout()
plt.savefig('uncertainty_intervals.png', dpi=300)
plt.close()

# ----------------------------
# 7. PRINT FINAL RESULTS TABLE
# ----------------------------
print("\n\nFINAL RESULTS")
print("-------------")
print("Model Performance Metrics:")
print(f"{'Model':<30} {'MSE':<10} {'R²':<10}")
print(f"{'Linear Regression':<30} {mse_lr:<10.4f} {r2_lr:<10.4f}")
print(f"{'LSTM Neural Network':<30} {mse_lstm:<10.4f} {r2_lstm:<10.4f}")
print(f"\nUncertainty Quantification:")
print(f"{'Model':<30} {'Coverage':<12} {'Mean Width':<12}")
print(f"{'Linear Regression':<30} {'N/A':<12} {'N/A':<12}")
print(f"{'LSTM (Bootstrapped)':<30} {coverage:<12.3f} {mean_width:<12.2f}")
```

LEVERAGING PREDICTIVE ANALYTICS FOR ENROLLMENT TREND AND CURRICULUM INNOVATION IN NIGERIAN STATISTICS PROGRAMS

Chidimma Florence Ejiofor[1], Ijeoma Eberechukwu Okechukwu Ph.D[2], Emeka Henry Chukwueloka[3]

[1,2]Department of Statistics, Federal Polytechnic Oko
[3]Mathematics Unit, Nigeria Maritime University, Okerenkoko Warri Delta State

*Corresponding author Email: florence.ejiofor@federalpolyoko.edu.ng, ijeoma.okechukwu@federalpolyoko.edu.ng, henry4real345@yahoo.com, henry.chukwueloka@nmu.edu.ng

Abstract

Statistics remains a foundation of the fast-growing field of Data Science. Yet, declining enrollment in Statistics programs across Nigerian tertiary institutions threatens the future workforce for data-driven sectors. This study applies predictive analytics to model enrollment trends and explore implications for curriculum development. Building on Ejiofor (2023), which used the Poisson distribution to reveal declining enrollment in the National Diploma Statistics program at Federal Polytechnic Oko, this paper extends the analysis with machine learning methods, particularly XGBoost, to capture national patterns. Due to the lack of comprehensive national data, simulated datasets were employed to approximate real-world trends and illustrate the potential of predictive modeling once complete data become available. Additionally, survey responses from Statistics Department staff across institutions were analyzed to provide insights into perceptions of Statistics education. Findings aim to guide strategies to strengthen programs and support the integration of Data Science into Nigeria's higher education system.

Keywords: predictive analytics, statistics enrollment, data science, curriculum development, data ecosystem.

1. INTRODUCTION

According to Liu et al. (2024), statistics is a fundamental discipline that supports a broad range of scientific and applied domains including artificial intelligence, public health, economics, education, engineering, social sciences and many others. Long-standing statistics programs at universities and polytechnics in Nigeria have aided in professional development, policy formation and national planning.

However, there have been apprehensions regarding the declining student enrollment in these programs, with evidence suggesting a downward trend in interest (Ejiofor, 2023). This issue has significant implications because Statistics forms the foundation of Data Science, a rapidly growing discipline that is driving innovation globally.

Additionally, as Watson and Smith (2022) pointed out, global disruptions like the COVID-19 pandemic and climate change have made it more urgent than ever to boost statistics education. If Statistics

programs continue to experience declining enrollment, Nigeria may face challenges in building the workforce needed for data-driven development.

To aid in planning and decision-making, forecasting models are being used increasingly in Nigeria to analyze higher education data. Using comparable exponential smoothing models, Adoga et al. (2023) examined a 16-year admission dataset from the National Open University of Nigeria (NOUN). Their results demonstrated that the Holt–Winters additive model produced the most accurate predictions for student admissions, thereby highlighting the importance of time-series methods in controlling student enrollments in schools. However, similar approach has not been widely applied to traditional statistics programs at universities and polytechnics, where there is evidence of declining student interest.

Furthermore, Akinode and Bada (2021) used machine learning approaches to forecast the likelihood of polytechnic enrollment based on pre-admission factors such as course selection, JAMB scores, WAEC/NECO results, and Post-UTME performance. They experimented with a number of models including Logistic Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (ID3). According to their findings, the Decision Tree model had the best accuracy indicating the use of machine learning to enhance enrollment predictions in Nigerian higher education institutions.

The purpose of this study is to apply predictive modelling to analyze enrollment trends in Statistics programs using machine learning models, with the ultimate goal of identifying strategies to strengthen Statistics education in Nigeria. While Data Science is expanding rapidly worldwide, it does not replace Statistics; instead, it builds upon statistical reasoning enhanced by computational techniques. Therefore, in order to provide and retain future graduates with the skills required in a data-driven industry, it is imperative that statistics education be made more appealing and of higher quality.

2. STATEMENT OF PROBLEM

According to preliminary results from institutional surveys, enrollment in statistics programs has fluctuated across Nigerian tertiary institutions, showing indications of decline in a number of instances. This trend is especially worrisome considering the rising demand for data-driven skills and expertise worldwide. More so, concerns regarding the sustainability of the field are raised by the fact that statistics continues to draw comparatively fewer students than other STEM-related fields. Furthermore, the lack of systematic, nationwide research and the scarcity of comprehensive enrollment figures make it challenging to fully understand the underlying causes of these trends. Without these insights, initiatives to boost enrollment, maintain students' interest in statistics and build institutional capacity run the risk of becoming ineffective.

3. SIGNIFICANCE OF THE STUDY

The findings of this study are significant in several respects. They provide reliable information that policymakers and professional organizations can use to design targeted efforts aimed at increasing participation in Statistics. At the institutional level, the results will assist administrators in anticipating student demand, strengthening capacity through improved infrastructure, and sustaining interest in the discipline. Methodologically, the study demonstrates how machine learning models can be applied to address educational challenges, offering a framework that can be adapted to other fields.

4. OBJECTIVES OF THE STUDY

- To examine spatial patterns in statistics program enrollment across states.
- To build and evaluate predictive model of enrollment trends in Nigerian tertiary institutions and to demonstrate their practical application through deployment as an interactive web application.
- To identify key factors influencing enrollment outcomes.
- To assess lecturers' perceptions of enrollment trends and compare them with model-based insights.

5. RESEARCH QUESTIONS

- **What spatial patterns in statistics program enrollment exist across states?**
- **How can predictive modeling, supported by deployment as a web application be applied to forecast enrollment trends in statistics programs?**
- **Which factors most strongly influence enrollment outcomes?**
- **What perspectives do lecturers hold on enrollment trends and how do these align with the predictions generated by the models?**

6. METHODOLOGY

- Survey Data Collection and Analysis

  To obtain insights on enrollment patterns and perceptions of Statistics education in Nigeria, a structured questionnaire was distributed through Google Forms. The survey targeted lecturers from tertiary institutions offering Statistics programs, including polytechnics, universities, colleges of education and monotechnics.

  As of the time of this study, responses were received from 16 institutions with polytechnics accounting for the largest share of submissions. The questionnaire sought

information on perceived student interest in Statistics, challenges facing Statistics programs and expectations for integrating Data Science into curricula. Descriptive statistics were applied to analyze the responses. The survey results provided a valuable contextual basis for understanding enrollment challenges beyond numerical data.

- Predictive Modeling with Machine Learning

  Due to the unavailability of comprehensive national enrollment figures, simulated datasets were generated to mimic likely enrollment patterns across institutions with polytechnics as the main focus. By narrowing the predictive modeling to polytechnics, this study creates a baseline that can later be extended to other categories of tertiary institutions once full datasets are available.

  The simulation incorporated features such as ownership (federal, state, private), program type (ND, HND or Both), state, institution location (rural, urban, semi-urban), institution age (how long it has been in existence), youth population percentage, graduate employment rate, fee waiver option, geopolitical zone, latitude, longitude and enrollment trend (increasing, decreasing, stable) which is the target variable.

- Models and Evaluation

  Four classification algorithms were applied: Decision Tree, Random Forest, XGBoost and Multinomial Logistic Regression. Decision Trees were selected for their interpretability and ability to capture decision rules in categorical outcomes. Random Forest, an ensemble of decision trees was included to enhance predictive accuracy and minimize overfitting. XGBoost, a gradient boosting classifier was employed due to its superior performance in handling complex, non-linear relationships and its proven robustness in classification tasks. While a statistical model, the Multinomial Logistic Regression was included to provide a baseline comparison from a traditional statistical perspective and to model the probabilities of multiple enrollment trend categories.

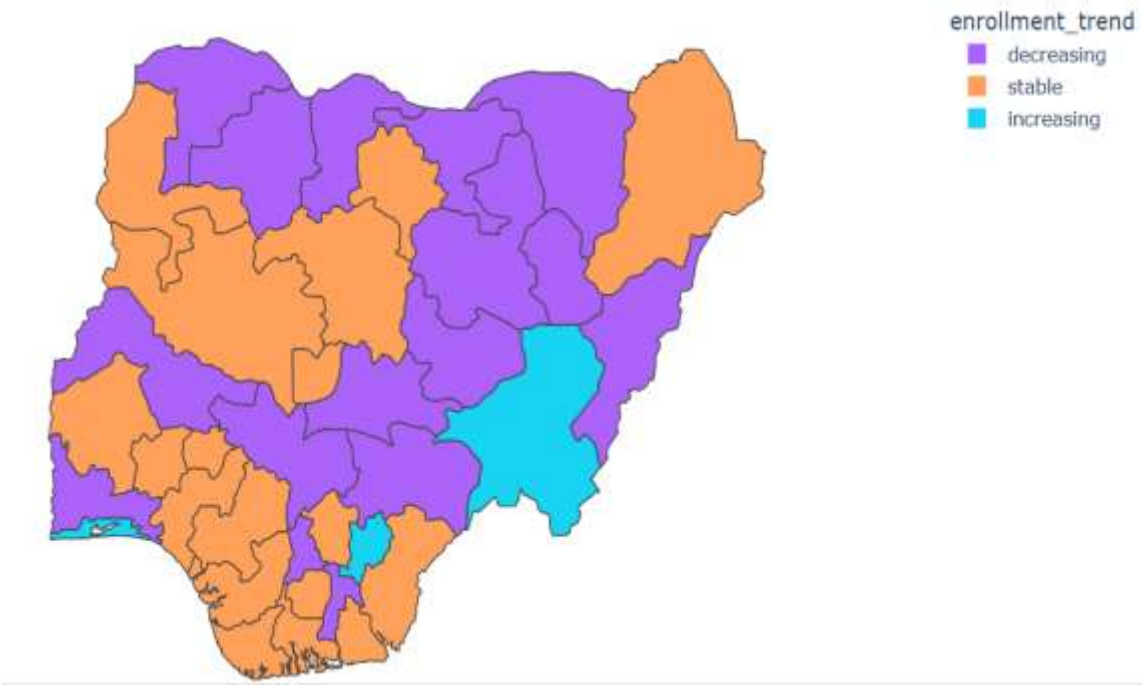
  For model evaluation, the Weighted F1 Score was used. Unlike overall accuracy, which may be misleading in the presence of class imbalance, the F1 Score combines precision (the proportion of correctly predicted positives out of all predicted positives) and recall (the proportion of correctly predicted positives out of all actual positives). The weighted

variant was chosen because it accounts for unequal class sizes, ensuring that performance was not biased toward majority categories. This provided a fairer and more reliable measure of the models' effectiveness in predicting enrollment trends.

SHapley Additive exPlanations (SHAP) was used to quantify the contribution of each feature to the model's predictions.

7. RESULTS AND DISCUSSION

Figure 1: Enrollment Trends in Statistics Programs Across Nigerian States



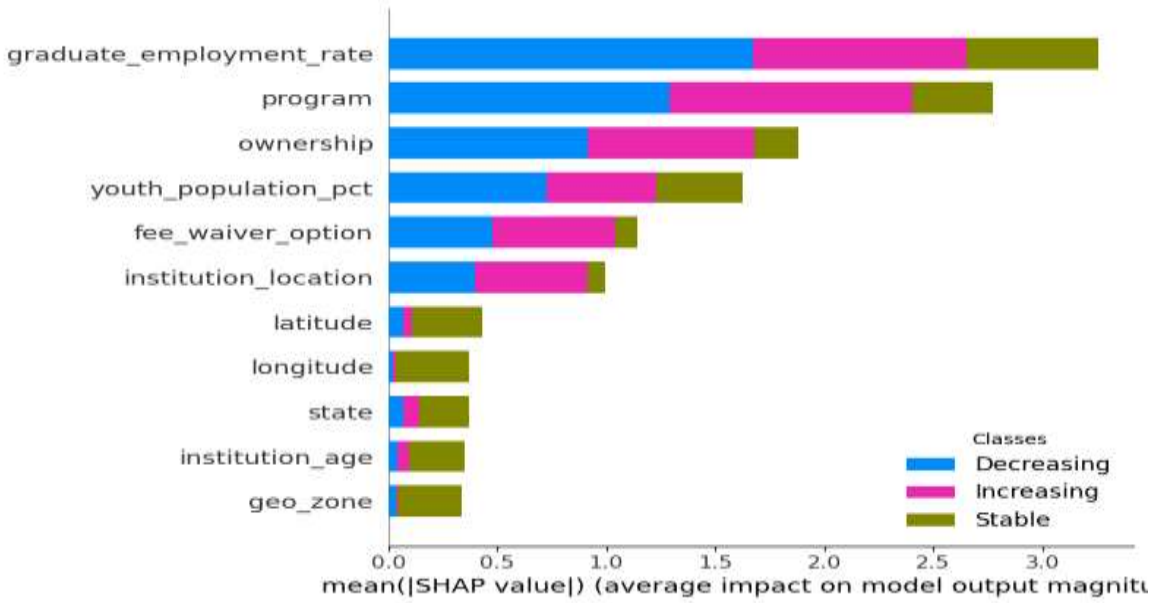
Note: The choropleth map in Figure 1 reveals that most Nigerian states are experiencing either *stable* or *declining* enrollment in statistics programs, with only a few states showing *increasing* trends. This uneven spatial pattern highlights the broader challenge of sustaining interest in statistics across the country. This suggests that localized factors may be influencing enrollment trajectories. Possible explanations include institutional reputation, targeted state-level policies, or community-driven educational initiatives, though these remain speculative without additional supporting data. The findings therefore underscore the importance of investigating contextual drivers more deeply, particularly in the states exhibiting growth as they may hold lessons for reversing declines elsewhere.

Table 1: Weighted F1 Scores of Machine Learning Models for Enrollment Trend Prediction

Model	Weighted F1 Score
Decision Tree	0.664
Random Forest	0.694
XGBoost	0.842
Multinomial Logistic Regression	0.762

Note. XGBoost outperformed other models in predicting enrollment trends.

Figure 2: SHAP summary plot showing the relative importance of features in predicting enrollment trends in statistics programs.



Interpretation of SHAP Results

The SHAP summary plot in Figure 2 indicates that **graduate employment rates**, **program type** and **institutional ownership** exert the greatest influence on enrollment trends. This suggests that prospective students place strong value on employability prospects after graduation and are also sensitive to the kinds of programs offered and whether institutions are publicly or privately owned. These findings align with wider concerns in the literature about the link between labor market relevance and higher education demand.

Other factors, such as **youth population proportions, fee waiver options** and **institutional location** play a moderate role, indicating that affordability and accessibility remain important but secondary considerations. Meanwhile, geographic attributes (e.g. state, geo-political zone, latitude/longitude) and institutional age had comparatively lower influence, implying that broader contextual or legacy factors are less decisive in shaping enrollment patterns than program quality and career outcomes.

Figure 3: Summary of lecturers' survey responses on Statistics programs

Fig A: Type of Institution

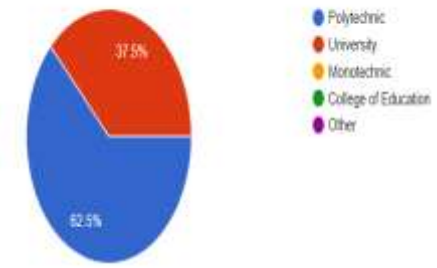


Fig B: Programmes Offered

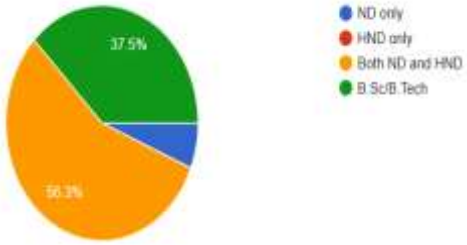


Fig C: Location of Institution

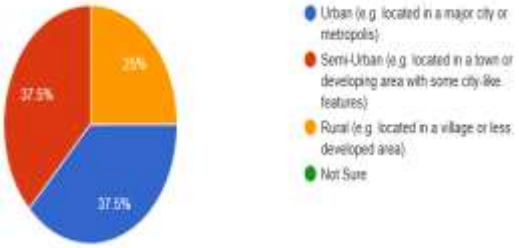


Fig D: Lecturers' Perception of Enrollment

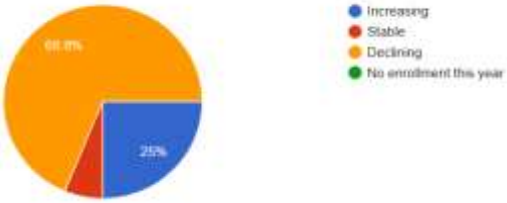


Fig E: Comparison with other STEM fields

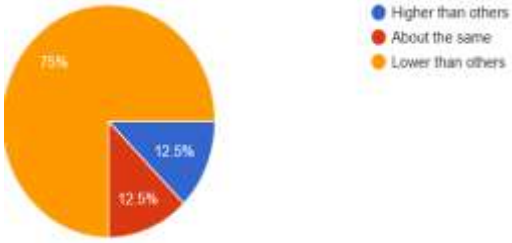
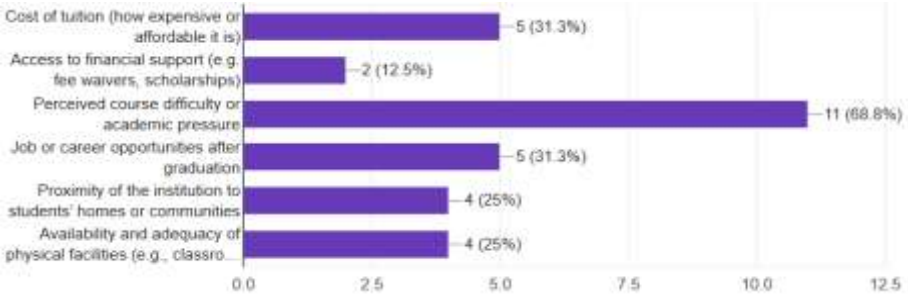


Fig F: Factors Influencing Enrollment



8. SUMMARY AND RECOMMENDATIONS

across institution type, geographical zones, and program offerings.

8. SUMMARY AND RECOMMENDATIONS

This study investigated enrollment trends in Statistics programs across Nigerian tertiary institutions using both lecturers' survey responses and machine learning models. Survey results indicated that enrollment in many institutions is currently declining, and Statistics programs generally attract fewer students compared with other STEM fields. The primary factors influencing enrollment were perceived course difficulty or academic pressure, followed by tuition costs and anticipated job or career opportunities after graduation. Additional factors identified include the institution's proximity to students' homes, availability of facilities and access to financial support.

Machine learning models, including XGBoost, Random Forest, Decision Tree and Multinomial Logistic Regression were employed to predict enrollment trends. Among these, XGBoost outperformed the others, identifying graduate employment rates, program type and institutional ownership as the most influential predictors. The predictive model was also deployed as an interactive web application, demonstrating its practical utility for institutional planning. The app can be accessed at https://enrollment-trend-app-9ukqkd7a6tmeaz7xe7uw5c.streamlit.app. Practical steps suggested by lecturers to boost enrollment include grassroots awareness campaigns (particularly targeting secondary schools), curriculum enhancement to integrate data science and analytics, financial incentives such as scholarships, early exposure through workshops and seminars, and strategic institutional measures like relocating departments or improving the learning environment.

REFERENCES

Adoga, P. I., Muazu, H., & Barma, M. (2023). Choosing a Forecast Model for Prediction of Students' Enrolment in Multiple Programmes of the National Open University of Nigeria: Towards Course Materials Production Planning. *West African Journal of Open and Flexible Learning*, 11(2), 1–28. Retrieved from https://wajofel.org/index.php/wajofel/article/view/137

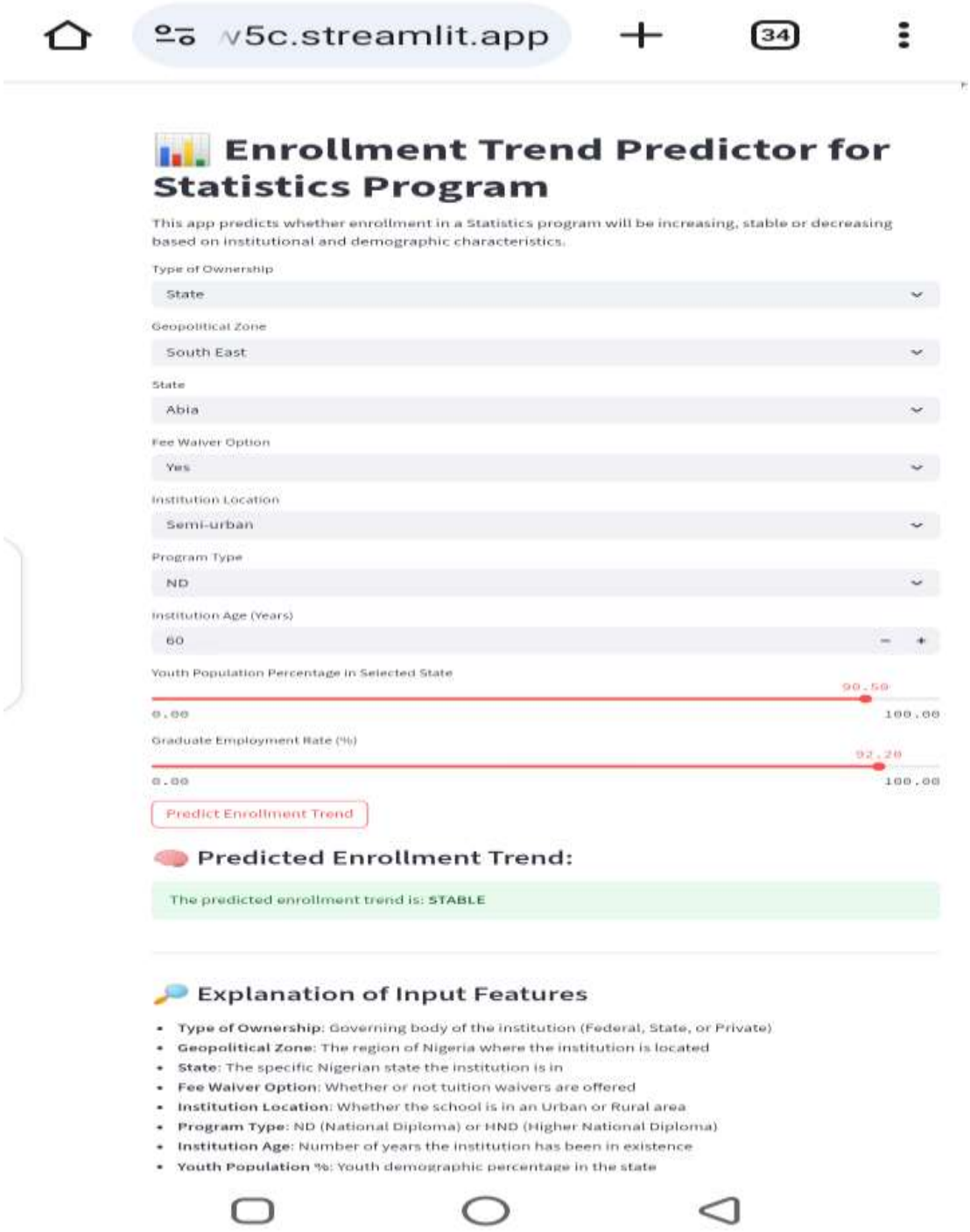
Akinode, J. L., & Bada, O. A. (2021). *Student enrollment prediction using machine learning techniques*. 5th National Conference of the School of Pure & Applied Sciences, Federal Polytechnic Ilaro, Ogun State, Nigeria. Federal Polytechnic Ilaro Institutional Repository.

Ejiofor Chidimma Florence, Ochuenwike Georgina Nwogo, Ph.D & Chukwunenye Victor Chigozie. *Journal of Humanities, Science & Technology (JOHUSAT)* 2nd Edition. A Publication of ASUP, Federal Polytechnic Oko, 2023, pp 147 – 152.

Liu, J., Chen, K. & Lyu, W. Embracing artificial intelligence in the labour market: the case of statistics. *Humanit Soc Sci Commun* **11**, 1112 (2024). https://doi.org/10.1057/s41599-024-03557-6

Watson, J., & Smith, C. (2022). Statistics education at a time of global disruption and crises: A growing challenge for the curriculum, classroom and beyond. *Curriculum Perspectives, 42*(2), 171–179. https://doi.org/10.1007/s41297-022-00167-7

APPENDIX


v5c.streamlit.app
34
Enrollment Trend Predictor for Statistics Program
This app predicts whether enrollment in a Statistics program will be increasing, stable or decreasing based on institutional and demographic characteristics.
Type of Ownership
State
Geopolitical Zone
South East
State
Abia
Fee Waiver Option
Yes
Institution Location
Semi-urban
Program Type
ND
Institution Age (Years)
60
Youth Population Percentage in Selected State
90.50
0.00
100.00
Graduate Employment Rate (%)
92.20
0.00
100.00
Predict Enrollment Trend
Predicted Enrollment Trend:
The predicted enrollment trend is: STABLE
Explanation of Input Features
Type of Ownership: Governing body of the institution (Federal, State, or Private)
Geopolitical Zone: The region of Nigeria where the institution is located
State: The specific Nigerian state the institution is in
Fee Waiver Option: Whether or not tuition waivers are offered
Institution Location: Whether the school is in an Urban or Rural area
Program Type: ND (National Diploma) or HND (Higher National Diploma)
Institution Age: Number of years the institution has been in existence
Youth Population %: Youth demographic percentage in the state

OPTIMAL CONTROL MODEL USING PROBABILITY DISTRIBUTION

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

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

*Corresponding author Email: uc.anyadiegwu@unizik.edu.ng

Abstract

In this article, we introduce a novel probability distribution, the Type I Heavy Tail Topp Leone Rayleigh (TI-HT-TLR) distribution, and apply it to an optimal control model for the transmission dynamics of an infectious disease. We derive several key statistical properties of the proposed model, including its quantile function, moments and moment-generating function, Renyi entropy, and order statistics. The parameters of the distribution are estimated using the maximum likelihood estimation (MLE) method. A comprehensive simulation study is conducted to evaluate the performance of the MLEs based on their bias and root mean square error (RMSE). The optimal control properties of the model are also derived to determine effective disease management strategies. The results demonstrate the practical utility and flexibility of the new distribution. When applied to real-world data, the TI-HT-TLR distribution provides a superior fit compared to several existing distributions. Furthermore, the findings from the optimal control model highlight the critical importance of early intervention in controlling infectious diseases.

Keywords: Type I Heavy Tail Topp Leone Rayleigh; Optimal Control; quarantine; Flexibility.

1. Introduction

Infectious diseases continue to be a significant global health and socio-economic issue, and hence, effective modeling and control policies are necessary. Optimal control policy based on traditional approaches generally relies on deterministic models, which are mathematically attractive but fail to capture the stochastic and uncertain nature of epidemic dynamics. These limitations become critical in the case where infection levels are low or fluctuations of disease transmission become predominant.

This research introduces the Type I Heavy-Tailed Topp-Leone G (TI-HT-TLR) distribution that combines the Topp-Leone G family with the heavy-tailed properties of the Type I Heavy-Tailed family. By embedding this new probability distribution into an optimal control model, we aim to enhance the accuracy of epidemic predictions and improve the evaluation of intervention strategies such as use of face mask and treatment. This integrated framework allows us to estimate epidemic outcomes under the uncertain setting and guide more effective, data-driven control policies.

2. Literature Review

The field of probability distribution has seen significant advancements with the development of various families of distributions (generators). These generators have improved the ability to model extreme events and outliers in diverse datasets. Some well-known examples include the heavy-tailed beta-power transformed Weibull (HTBPT-W) distribution by Zhao et al. (2021), the Topp-Leone (TL) family of distributions by Al-Shomrani et al. (2016), and the Type-I Heavy-Tailed (TI-HT) family by Zhao et al. (2020). Other recent contributions include the Topp-Leone-Harris-G distribution by Oluyede et al. (2023), the Topp-Leone Dagum distribution by Rasheed (2020), the DUS Topp-Leone family by Ekemezie et al. (2024), the Odd Inverted Topp-Leone-H (OITL-H) family by Hassan et al. (2022), and the Type I Half-Logistics-Topp-Leone-G (TIHLTL-G) distribution by Adepoju et al. (2023). A notable recent development is the Topp-Leone Type I Heavy Tail distribution by Nkoma et al. (2025).

The foundation of optimal control theory, particularly Pontryagin's Minimum Principle (PMP), was established by Pontryagin et al. (1962). This principle has since been applied to various epidemiological models. For instance, Egonmwan and Okuonghae (2019) utilized optimal control theory to model the transmission dynamics of tuberculosis (TB). More recently, Hemeda (2024) introduced the Cosh Inverse Exponential (CIE) distribution within an optimal control framework to manage the spread of COVID-19. However, deterministic models are known to be poor approximations in cases where the number of infected hosts is low Russell and Cunniffe (2025). While stochastic models offer a more realistic alternative, their performance heavily relies on the choice of the underlying probability distributions used to model transmission dynamics.

To address this gap, this study proposes a novel family of probability distributions by integrating the Topp-Leone-G family (Al-Shomrani et al., 2016) with the heavy-tail modeling capability of the Type-I Heavy-Tailed family (Zhao et al., 2020). This enhanced distribution is then embedded within an optimal control framework to evaluate and improve intervention strategies for effective disease control policies.

3. Method and Materials

Based on the Type I Heavy-Tailed-G (TIHTG) distribution introduced by Zhao et al. (2020) and the Topp-Leone-G family proposed by Al-Shomrani et al. (2016), this study introduces a new and more flexible class of distributions, referred to as the Type I Heavy-Tailed Topp-Leone-G (TI-HT-TL-G) family.

The cumulative distribution function (CDF) of the Type I Heavy-Tailed-G distribution, as discussed in Zhao et al., (2020) is:

$$G(x;\theta,\Delta) = 1 - \left(\frac{1-F(x;\Delta)}{1-(1-\theta)F(x;\Delta)}\right)^{\theta}$$

While the probability density function (PDF) is

$$g(x;\theta,\Delta) = \frac{\theta^2 f(x;\psi)[1-F(x;\Delta)]^{\theta-1}}{[1-(1-\theta)F(x;\Delta)]^{\theta+1}}$$

for $\theta, x > 0$, where Δ denotes the parameter vector from the baseline distribution $F(.)$.

Al-Shomrani et al. (2016) also presented the Topp Leone-G (TL-G) family of distribution with the CDF and PDF given as follows;

$$F(x) = [1 - \{1 - T(x)\}^2]^{\alpha} \quad (1)$$

and

$$f(x) = 2\alpha t(x)[1 - T(x)]\{T(x)\}^{\alpha-1}\{2 - T(x)\}^{\alpha-1}; \quad \alpha > 0 \quad (2)$$

To derive the cdf and pdf, we study the special case when $\alpha = 1$.

Replacing the baseline CDF in equation with the CDF of the TL-G family in equation (1) to have the new family of distribution called TI-HT-TL-G with CDF when

$$G(x, \theta, \Delta) = 1 - \left(\frac{1-\left(1-\left(\bar{T}(x)\right)^2\right)}{1-(1-\theta)\{1-(\bar{T}(x))^2\}}\right)^{\theta}$$

Substitute equation (3) and (4) into the PDF to have the PDF

$$g(x, \theta, \Delta) = \frac{2\theta^2 t(x)\bar{T}(x)\left[1-\{1-(\bar{T}(x))^2\}\right]^{\theta-1}}{[1-(1-\theta)\{1-(\bar{T}(x))^2\}]^{\theta+1}} \quad (3)$$

Where $\bar{\bar{T}}(x) = 1 - T(x)$ is the baseline survival function. $t(x)$ is the associated pdf of the baseline distribution.

Solving for the sub-model, the cdf and pdf of the Rayleigh distribution is given as;

$$t(x) = 2\beta x e^{-\beta x^2} \quad (4)$$

$$T(x) = 1 - e^{-\beta x^2} \quad (5)$$

To solve for the cdf, we substitute equations (5) into (3) to have

$$G(x, \theta, \beta) = 1 - \left(\frac{e^{-2\beta x^2}}{1-(1-\theta)\left(1-e^{-2\beta x^2}\right)}\right)^{\theta}$$

Substituting equations (4) and (5) into (3) gives the pdf of the TI-HT-TLR distribution

$$g(x, \theta, \alpha, \beta) = \frac{4\theta^2 \beta x e^{-2\theta\beta x^2}}{\left[1-(1-\theta)\left\{1-e^{-2\beta x^2}\right\}\right]^{\theta+1}}$$

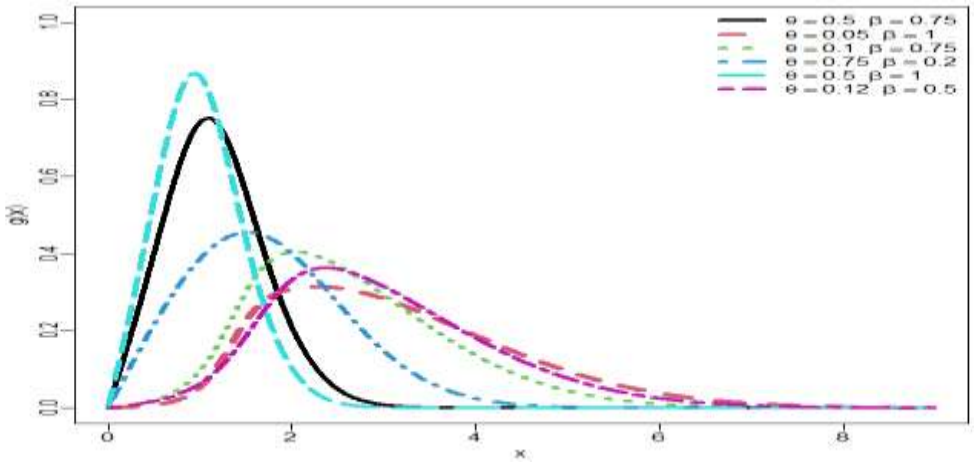


Figure 1: pdf plots of the TI-HT-TLR Distribution

Expansion of Density Function of the TI-HT-TLR Distribution

Given the pdf of the TIHTTLR in Eq. (10)

$$g(x, \theta, \beta) = \frac{4\theta^2 \beta x e^{-2\theta\beta x^2}}{\left[1-(1-\theta)(1-e^{-2\beta x^2})\right]^{\theta+1}}$$

$$g(x, \theta, \beta) = 4\theta^2 \beta x e^{-2\theta\beta x^2} \times \left[1 - (1-\theta)(1-e^{-2\beta x^2})\right]^{-(\theta+1)}$$

$$g(x, \theta, \beta) = 4\theta^2 \beta x e^{-2\theta\beta x^2} \times \sum_{i=0}^{\infty} \binom{\theta+i}{i} (1-\theta)^i \left(1-e^{-2\theta\beta x^2}\right)^i$$

$$= 4\theta^2 \beta x e^{-2\theta\beta x^2} \times \sum_{i=0}^{\infty} \binom{\theta+i}{i} (1-\theta)^i \sum_{k=0}^{\infty} (-1)^k \binom{i}{k} e^{-2\theta\beta x^2}$$

$$= 4\theta^2 \beta x e^{-2\theta\beta x^2} \times \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \binom{\theta+i}{i} (1-\theta)^i (-1)^k \binom{i}{k} e^{-2\theta\beta x^2}$$

$$g(x,\theta,\beta) = 4\theta^2\beta \times \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\binom{\theta+i}{i}(1-\theta)^i(-1)^k\binom{i}{k}xe^{-2\beta x^2(\theta+k)} = Q_{ik}$$

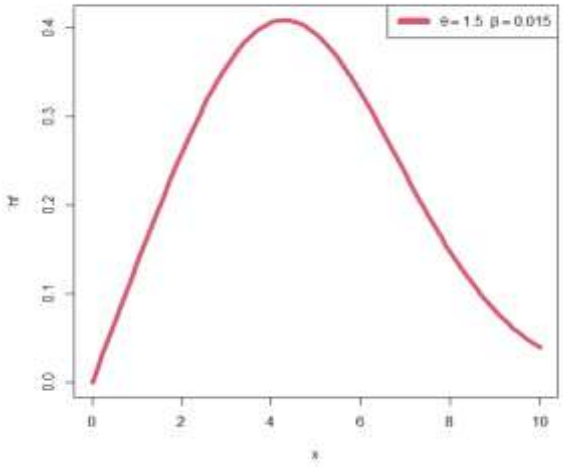
Statistical Properties

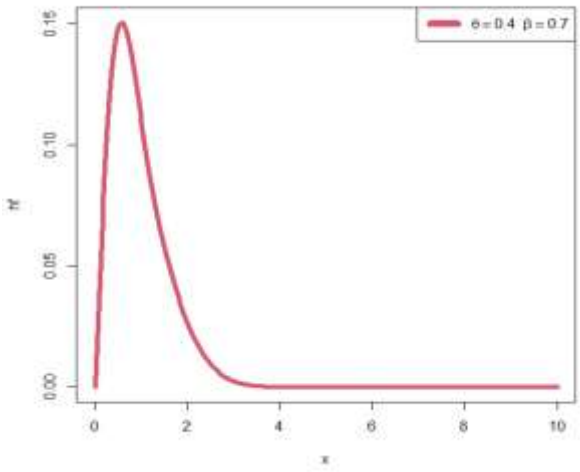
Then the corresponding survival and hazard rate functions of the TIHTTLR distribution are respectively given as

$$S(x,\theta,\beta) = \left(\frac{e^{-2\beta x^2}}{1-(1-\theta)\left(1-e^{-2\beta x^2}\right)}\right)^{\theta}$$

And

$$h(x) = \frac{4\theta^2\beta x}{\left[1-(1-\theta)(1-e^{-2\beta x^2})\right]}$$





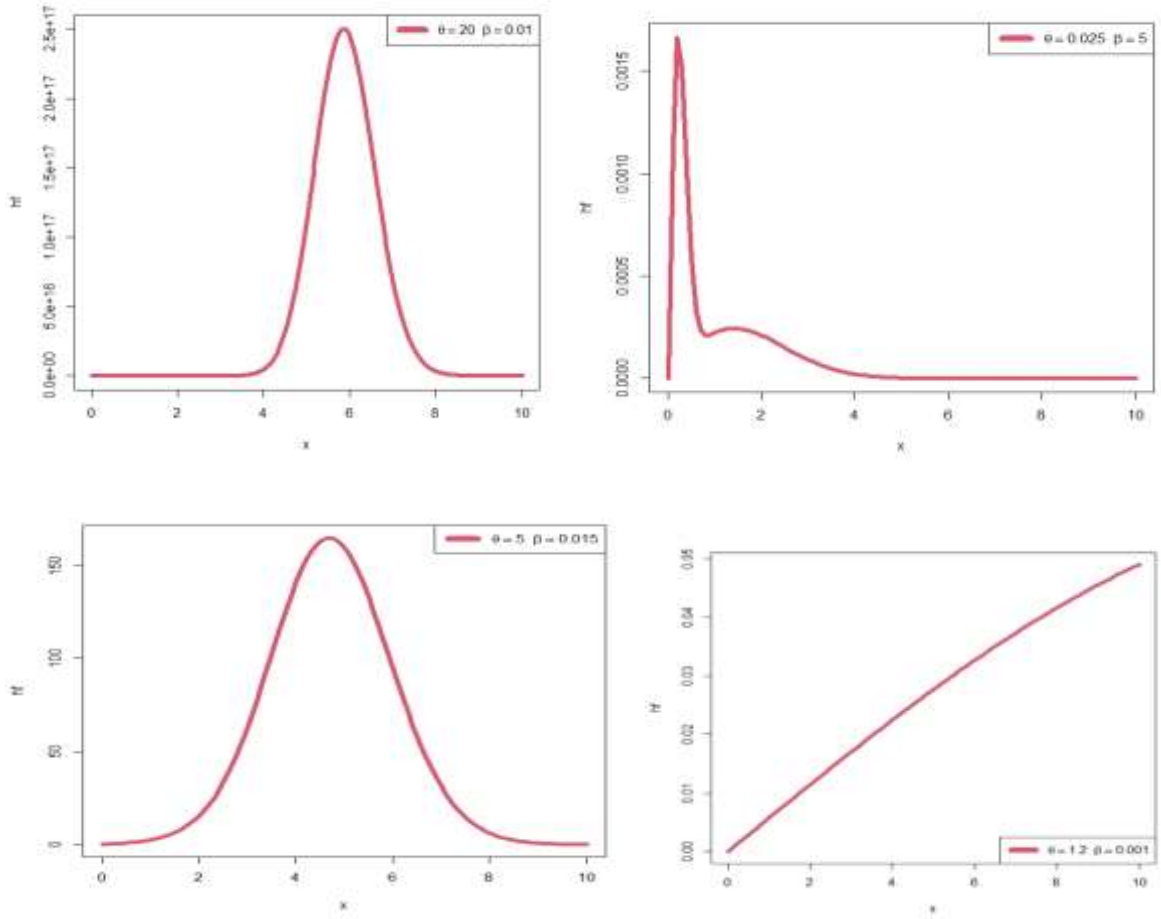


Figure 2: Hazard function plots of TI-HT-TLR Distribution

The shape of the hazard function plots in Figure (2) displays a variety of typical shapes. They are made up of bimodal-shaped curves, right-skewed, and left-skewed patterns, bell-shaped and symmetric shapes. There is even one with a predominantly straight-line or linear pattern. These shapes reflect the flexibility of the TI-HT-TLR distribution to model a wide range of failure behaviors and hazard dynamics under various parameter settings.

Quantile Function of TI-HT-TLR Distribution

Let X~TI-HT-TLR (θ, β) and the cumulative distribution function (CDF) represented by Eq. (9). The quantile function, which provides the inverse relationship between the distribution function and the corresponding quantiles, is obtained by solving the equation:

Let $P = G(x, \theta, \beta)$

$$P = 1 - \left(\frac{e^{-2\beta x^2}}{1-(1-\theta)(1-e^{-2\beta x^2})}\right)^{\theta}$$

$$x = Q(p) = \sqrt{-\frac{1}{2\beta} ln\left[\frac{\theta(1-P)^{1/\theta}}{1-(1-P)^{\frac{1}{\theta}}(1-\theta)}\right]}$$

The quantile function derived in Eq. (14) plays a crucial role in simulating random samples from the TI-HT-TLR distribution, making it especially valuable for simulation purposes.

Renyi Entropy $H_\alpha(x)$ of the TIHTTLR Distribution

$H_\alpha(x) = \frac{1}{1-v} log \int_0^\infty g(x)^v dx \qquad for\ \alpha > 0, \alpha \neq 1$

The Renyi Entropy is given below

$H_v(x) = \frac{1}{(1-v)} log\left(\left(4\theta^2\beta\right)^v\right)\left(\frac{v+1}{2}\right)\sum_{i=0}^{\infty}\sum_{k=0}^{i}\binom{v(\theta+1)+i-1}{i}(1-\theta)^i\binom{i}{k}(-1)^k\left(2[2\beta(v\theta+k)]^{-\frac{v+1}{2}}\right)$

The j^{th} moment of the TI-HT-TLR distribution is derived as

$$\mu_j^{'} = \int_0^\infty x^j g(x,\theta,\beta)dx$$

$\mu_j^{'} = \theta^2\Gamma\left(\frac{j}{2}+1\right)\sum_{i=0}^{\infty}\sum_{k=0}^{i}\binom{\theta+i}{i}(1-\theta)^i(-1)^k\binom{i}{k}\left(\frac{1}{2\beta(\theta+k)}\right)^{\frac{j}{2}+1}$

Moment Generating Function (MGF) of TIHTTLR Distribution

The MGF of $X \sim TIHTTLR(\theta,\beta)$ is given by

$$M_x^{(t)} = E[e^{tx}] = \int_0^\infty e^{tx} g(x)dx$$

Using the series expansion in Eq. (11) we have the MGF to be

$M_x^{(t)} = 4\theta^2\beta \times \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}\binom{\theta+i}{i}(1-\theta)^i(-1)^k\binom{i}{k}\int_0^\infty xe^{-2\beta x^2(\theta+k)+tx}dx$

Distribution of the Order Statistics

The order statistics of a random sample are the values of the sample sorted in ascending order. For a sample of size n, the k-th order statistics, denoted as $X_{(k)}$, is the k-th smallest value. The probability density function of the k-th order statistic is given by

$g_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!}[G(x)]^{k-1}[1-G(x)]^{n-k}g(x)$

The pdf of the largest order statistics is gotten by replacing k with n that is $k = n$

$$g_{(n)}(x) = \frac{4\theta^2 n\beta x e^{-2\theta\beta x^2}\left[1-\left(\frac{e^{-2\beta x^2}}{1-(1-\theta)\left(1-e^{-2\beta x^2}\right)}\right)^\theta\right]^{n-1}}{\left(1-(1-\theta)\left(1-e^{-2\beta x^2}\right)\right)^{\theta+1}}$$

The pdf of the smallest order statistics is given by replacing k with 1 that is $k = 1$

$g_{(1)}(x) = n[S(x)]^{n-1}g(x)$

$$g_{(1)}(x) = \frac{4\theta^2 n\beta x e^{-2n\theta\beta x^2}}{\left(1-(1-\theta)(1-e^{-2\beta x^2})\right)^{n\theta+1}}$$

Estimation of Parameters

Maximum Likelihood Estimation (MLE) of the TI-HT-TLR Distribution Parameter

Let $(x_1, x_2, \ldots, x_n)$ be n random samples drawn from TI-HT-TLW distribution, then the likelihood function is given as

$L(\theta,\beta/x) = \prod_{i=1}^n g(x_i,\theta,\beta)$

$L(\theta,\beta) = (4\theta^2\beta)^n\prod_{i=1}^n x_i e^{-2\theta\beta x_i^2}\prod_{i=1}^n\left(1-(1-\theta)(1-e^{-2\beta x_i^2})\right)^{-(\theta+1)}$

Taking the natural log to have

$l(\theta,\beta,\lambda) = nlog(4\theta^2\beta) + \sum_{i=1}^n logx_i - 2\theta\beta\sum_{i=1}^n x_i^2 - (\theta+1)\sum_{i=1}^n log\left[1-(1-\theta)(1-e^{-2\beta x_i^2})\right]$

To estimate the parameter θ, take the derivative with respect to (w.r.t) θ

$$\frac{\partial l}{\partial \theta} = \frac{2n}{\theta} - 2\beta \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \frac{log\left[1-(1-\theta)\left(1-e^{-2\beta x_i^2}\right)\right]}{1-(1-\theta)\left(1-e^{-2\beta x_i^2}\right)} \times \left(1 - e^{-2\beta x_i^2}\right)$$

Setting $\frac{\partial l}{\partial \theta} = 0$ gives the MLE

Taking the partial derivative w.r.t. β to have

$$\frac{\partial l}{\partial \beta} = \frac{n}{\beta} - 2\theta \sum_{i=1}^{n} x_i^2 + 2(\theta + 1)(1 - \theta) \sum_{i=1}^{n} \frac{x_i^2 e^{-2\beta x_i^2}}{1-(1-\theta)\left(1-e^{-2\beta x_i^2}\right)}$$

The MLEs$(\hat{\theta}, \hat{\beta})$ for (θ, β) are derived by equating the score function (that is the equations of the partial derivatives of the log-likelihood function with respect to the parameters), denoted by $C(\omega)$ to zero and solve the system of equations $C(\omega) = \left(\frac{\partial l}{\partial \theta}, \frac{\partial l}{\partial \beta}\right)' = 0$ simultaneously. These systems of equations are non-linear and are solved using the Newton-Raphson iterative method with the help of R software to obtain the MLEs of the unknown parameters.

Optimal Control Model

In this section, optimal control theory is applied to model disease transmission through the TI-HT-TLR distribution and incorporates intervention strategies such as use of face mask and treatment.

By incorporating the intervention strategies (u_1, u_2) into the existing state equations established by Manu et al. (2025), we derive the new control system, which is expressed as

$$\frac{dS(t)}{dt} = A - \rho(t)(1 - u_1(t))S(t)I(t) - \mu S(t) + \varphi R(t)$$

$$\frac{dE(t)}{dt} = \rho(t)(1 - u_1(t))S(t)I(t) - (\varepsilon + \mu)E(t)$$

$$\frac{dI(t)}{dt} = \varepsilon E(t) - (\gamma + u_2(t)h(t) + \mu + \delta)I(t)$$

$$\frac{dR(t)}{dt} = (\gamma + u_2(t)h(t))I(t) - (\mu + \varphi)R(t)$$

Where

Susceptible $S(t)$, Exposed $E(t)$, Infected $I(t)$, and Recovered $R(t)$ at any given time, t. The population has two parameters, the recruitment rate A and the death rate μ; and the model disease has five compartments, the transmission rate ρ, the progression rate to infection ε the recovery rate γ, the disease death rate δ, the acquired immunity temporary rate φ, use of facemask u_1, treatment intervention u_2 and $h(t)$ is the hazard rate given in equation (13).

Objective Function

The objective is to minimize the number of infected individuals and the cost associated with implementing the control measures over a finite time horizon $[0, T]$.

$$W(u_1, u_1) = \int_0^T \left(C_1 I(t) + \frac{1}{2} C_2 u_1^2(t) + \frac{1}{2} C_3 u_2^2(t)\right) dt$$

where

C_1: Weight associated with the cost of infected individual

C_2: weight associated with the cost of prevention control u_1

C_3: Weight associated with the cost of treatment control u_2

$C_1, C_2, C_3 > 0$

Hamiltonian Function

Determine the necessary conditions of optimality, which is derived using the objective function integrand and the right-hand side of the system dynamics in Eq (20) along with the co-state variables $(\zeta_1, \zeta_2, \zeta_3, \zeta_4)$

$$H = C_1 I(t) + \frac{1}{2} C_2 u_1^2(t) + \frac{1}{2} C_3 u_2^2(t) + \zeta_1[AN - \rho(t)(1-u_1(t))S(t)I(t) - \mu S(t) + \varphi R(t)]$$
$$+\zeta_2[\rho(t)(1-u_1(t))S(t)I(t) - (\varepsilon+\mu)E(t)] + \zeta_3[\varepsilon E(t) - (\gamma + u_2(t)h(t) + \mu + \delta)I(t)]$$
$$+\zeta_4[(\gamma + u_2(t)h(t))I(t) - (\mu+\varphi)R(t)]$$

Co-State Equation

The co-state equations are derived by differentiating the Hamiltonian with respect to each state variable.

$$\frac{d\zeta_1}{dt} = \frac{-\partial H}{\partial S},\ \frac{d\zeta_2}{dt} = \frac{-\partial H}{\partial E}, \frac{d\zeta_3}{dt} = \frac{-\partial H}{\partial I}, \frac{d\zeta_4}{dt} = \frac{-\partial H}{\partial R}$$

$$\frac{d\zeta_1}{dt} = (\zeta_1 - \zeta_2)(1-u_1)\rho I(t) + \zeta_1 \mu$$

$$\frac{d\zeta_2}{dt} = \zeta_2(\varepsilon+\mu) + \zeta_3 \varepsilon$$

$$\frac{d\zeta_3}{dt} = -C_1(\zeta_1 - \zeta_2)(1-u_1)\rho S(t) + \zeta_3(\gamma + u_2 h(t) + \mu + \delta) + \zeta_4(\gamma + u_2 h(t))$$

$$\frac{d\zeta_4}{dt} = \frac{-\partial H}{\partial R} = \zeta_4(\mu+\varphi) - \zeta_1 \varphi$$

With transversality conditions $\zeta_i(T) = 0$ for $i = 1,2,3,4$

Optimality Condition

These equations ensure that the controls are chosen optimally at each instant according to the current state and co-state values

$$u_1^*(t) = max\left\{0, min\left(1, \frac{\rho S(t) I(t)}{C_2}(\zeta_2 - \zeta_1)\right)\right\}$$

$$u_2^* = max\left\{0, min\left(1, \frac{h(t) I(t)}{C_3}(\zeta_3(t) - \zeta_4(t))\right)\right\}$$

Zero (0) is no face mask or treatment, while 1 is use of face mask or treatment. This ensures the optimal use of facemask or treatment effort is within feasible limits while satisfying the minimum condition of the Hamiltonian.

Simulation Result

To compare and evaluate the performance of the various non-Bayesian estimators of the parameters of the TI-HT-TLR distribution, an extensive simulation study was conducted. The methods used are maximum likelihood estimation (MLE), maximum product spatialization (MPS), least squares (LS), weighted least squares (WLS), Cramér-von Mises (CVM), Anderson-Darling (AD) and right-tailed version of Anderson–Darling (RTAD). For each of the estimation methods, the parameter estimates were calculated with 10,000 bootstrap replicates for four different sample sizes. This allowed for a robust comparison of bias, mean squared error (MSE), and general estimation performance for conditions from small to large. Additionally, Bayesian estimation was conducted on the rest of the samples using the "Squared Error Loss (SEL), LINEX loss, and General Entropy Loss (GEL) loss functions. To compare and analyze the performance of all the estimation techniques, average bias and root mean square error (RMSE) were computed for each sample size for 10,000 iterations". The measures provided an overall evaluation of the precision and accuracy of each technique.

Estimated Bias and RMSE for TIHTTLR parameter at (n = 25, 75, 150, 200) with θ = 1.15, β = 0.75

Type	Method	Bias (25)	RMSE (25)	Bias (75)	RMSE (75)	Bias (150)	RMSE (150)	Bias (200)	RMSE (200)
Non Bayesian	MLE_θ	1.4276	19.0601	0.43921	7.14895	0.13805	1.69984	0.09487	0.26125
	MLE_β	0.0783	0.6368	0.0505	0.2813	0.0279	0.1539	0.0053	0.0449

	MPS_θ	0.1917	4.5623	0.0106	1.5987	0.0553	0.2176	0.0270	0.1805
	MPS_β	0.6882	2.2735	0.3979	1.4451	0.1398	0.3767	0.0761	0.1504
	LS_θ	0.4031	4.1611	0.1323	1.3634	0.0242	0.2794	0.0291	0.2041
	LS_β	0.3667	0.8569	0.1544	0.2847	0.0755	0.1001	0.0440	0.0632
	WLS_θ	0.5438	6.1987	0.1952	1.9829	0.0537	0.2428	0.0485	0.1729
	WLS_β	0.3288	0.8819	0.1093	0.2414	0.0432	0.0679	0.0251	0.0516
	CvM_θ	0.9103	6.9458	0.2818	1.6986	0.0913	0.3115	0.0781	0.2202
	CvM_β	0.1454	0.5637	0.0802	0.2281	0.0423	0.0867	0.0206	0.0563
	AD_θ	0.6924	6.4378	0.1989	1.5030	0.0555	0.2385	0.0499	0.1721
	AD_β	0.2478	0.8068	0.0869	0.1960	0.0413	0.0661	0.0280	0.0747
	$RTAD_\theta$	0.6029	3.4843	0.2215	1.0110	0.0679	0.2437	0.0728	0.2224
	$RTAD_\beta$	0.2323	0.8458	0.0753	0.2533	0.0334	0.0614	0.0168	0.0550
Bayesian	SEL_θ	0.5618	0.4743	0.5697	0.3861	0.5771	0.3695	0.5763	0.3605
	SEL_β	2.5740	13.573	2.2792	11.054	2.0942	9.3432	1.9609	7.7700
	$LINEX1_\theta$	0.5584	0.4738	0.5674	0.3843	0.5755	0.3679	0.5748	0.3590
	$LINEX1_\beta$	2.7522	16.4132	2.3502	12.685	1.9516	7.7090	1.8570	6.3788
	$LINEX2_\theta$	0.5652	0.4750	0.5719	0.3879	0.5788	0.3711	0.5778	0.3620
	$LINEX2_\beta$	2.3384	10.517	2.0807	8.2905	1.9429	7.0828	1.8503	6.1562
	$GEL2_\theta$	0.5668	0.4791	0.5734	0.3900	0.5799	0.3727	0.5788	0.3634
	$GEL1_\beta$	2.5255	13.029	2.2379	10.563	2.0621	8.9447	1.9362	7.4981
	$GEL2_\theta$	0.5768	0.4887	0.5807	0.3980	0.5856	0.3791	0.5839	0.3691
	$GEL2_\beta$	2.4322	12.0374	2.1592	9.6857	2.0010	8.2413	1.8886	7.0078

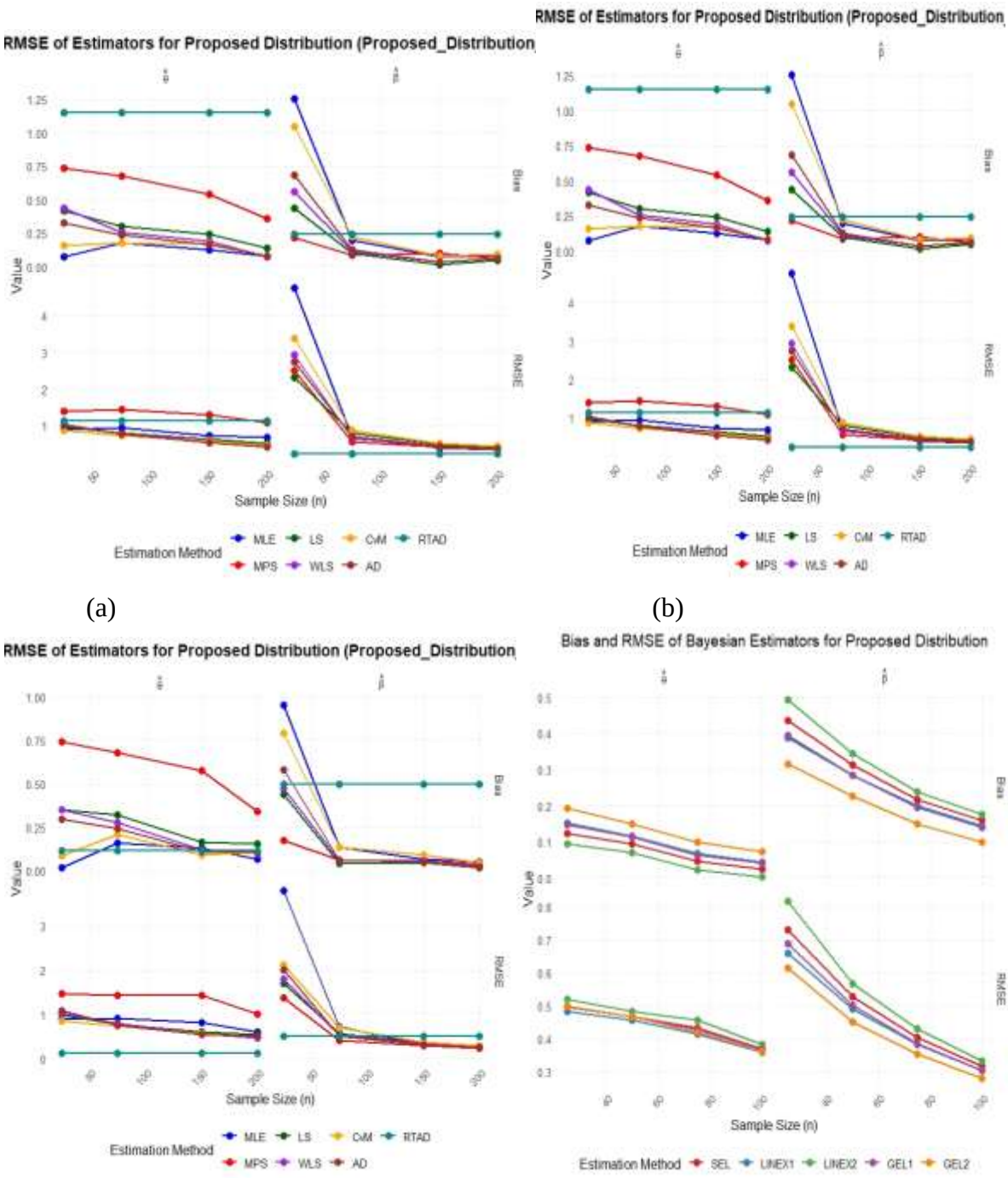
Table 2: Estimated bias and RMSE for the TI-HT-TLR parameters at (n = 25, 75, 150, 200) and fixed values θ = 1.15, β = 0.5.

Type	Method	Bias (n=25)	RMSE (n=25)	Bias (n=75)	RMSE (n=75)	Bias (n=150)	RMSE (n=150)	Bias (n=200)	RMSE (n=200)
Non Bayesian	MLE_θ	1.09176	8.93171	0.41519	2.23717	0.20342	0.83982	0.10153	0.27743
	MLE_β	0.05841	0.34231	0.0176	0.1243	0.00272	0.02476	0.00231	0.01788
	MPS_θ	0.20914	3.42801	0.12577	1.49973	0.06499	0.59076	0.00923	0.235
	MPS_β	0.52296	1.67825	0.1503	0.44775	0.04319	0.07079	0.03289	0.02203
	LS_θ	0.31391	2.42226	0.1879	1.14788	0.10383	0.47787	0.05386	0.205
	LS_β	0.23917	0.49416	0.0621	0.0963	0.0214	0.02927	0.01581	0.01966
	WLS_θ	0.46177	3.71309	0.2453	1.54239	0.11776	0.41106	0.06698	0.18255
	WLS_β	0.19601	0.45178	0.03831	0.07043	0.0097	0.02433	0.00887	0.0176
	CvM_θ	0.67963	3.42599	0.31945	1.29739	0.15644	0.46439	0.10338	0.28275
	CvM_β	0.07202	0.26618	0.01823	0.07345	0.00242	0.02681	0.00143	0.0187
	AD_θ	0.55111	3.28831	0.25348	1.16869	0.1234	0.39893	0.07367	0.20303
	AD_β	0.12716	0.3705	0.03157	0.07459	0.00773	0.02448	0.0075	0.01768
	$RTAD_\theta$	0.42585	1.85879	0.24621	0.86917	0.14533	0.45735	0.09652	0.23814
	$RTAD_\beta$	0.18621	0.66079	0.03511	0.10016	0.00902	0.02893	0.00648	0.02161
Bayesian	SEL_θ	0.59456	0.51342	0.60191	0.42458	0.60516	0.40364	0.60226	0.39181
	SEL_β	1.96091	7.94905	1.6674	6.01388	1.45168	4.36036	1.32549	3.18312
	$LINEX1_\theta$	0.59158	0.51312	0.59994	0.42294	0.60367	0.40215	0.6009	0.39039
	$LINEX1_\beta$	1.9666	8.12995	1.6228	5.73535	1.33348	3.17626	1.23963	2.25344
	$LINEX2_\theta$	0.5975	0.51385	0.60387	0.42622	0.60663	0.40512	0.60361	0.39324
	$LINEX2_\beta$	1.91553	7.4666	1.64179	5.63483	1.45058	4.2427	1.33723	3.31062
	$GEL2_\theta$	0.59885	0.51738	0.60514	0.42809	0.60773	0.40661	0.60463	0.39455
	$GEL1_\beta$	1.9605	8.01692	1.6508	5.88648	1.433	4.2011	1.31163	3.09623
	$GEL2_\theta$	0.60737	0.52541	0.61159	0.43517	0.61288	0.41261	0.60936	0.40007
	$GEL2_\beta$	1.90748	7.59388	1.64999	5.98756	1.42092	4.16321	1.28456	2.93545

Tables1-2 compare the performance of different estimation methods for the TI-HT-TLR model in terms of bias and RMSE as a performance measure. Results are shown over sample sizes $n = 25, 75, 150, 200$

For the non-Bayesian estimators: All the methods show improved performance (smaller RMSE and bias) as sample size increases. Among non-Bayesian methods, RTADE, WLS, and CvM consistently show higher accuracy for both parameters, especially for large n. MLE fails in estimating θ in the case of small samples owing to large variance, but this improves considerably with large n. MPS performs well θestimation but comparatively higher bias for β

For the Bayesian estimators: The Bayesian estimators have low and stable bias for $\theta$for all sample sizes. Estimation of βunder Bayesian estimators is higher in bias and RMSE, especially under SEL and LINEX1. Among Bayesian approaches, LINEX2 and GEL2 perform relatively better in estimation accuracy for βparticularly as n becomes large.



(a) (b)

(c) (d)

Figure 2: Non Bayesian and Bayesian plot for Table 1 and 2.

Figure 2: (a) Non-Bayesian plot for Table 1, (b) Bayesian plot for Table 1, (c) Non-Bayesian plot for Table 2, (d) Bayesian plot for Table 2,

Figure 2(a,c) compares Bias and RMSE for $\hat{\theta}$ and $\hat{\beta}$ across sample sizes ($n = 25, 75, 150, 200$) using seven estimation methods. Bias and RMSE decline as n increases, confirming estimator consistency. For $\hat{\theta}$, MLE maintains minimal bias and low RMSE at all n, while MPS improves steadily; CvM, AD, LS and WLS perform moderately well, but RTAD shows persistently high bias. For $\hat{\beta}$, MLE and MPS start with higher bias at n = 25 but converge rapidly to near unbiased, achieving the lowest RMSE at large n. The bias and root mean square error (RMSE) of the Bayesian estimators for the symmetric (SEL) and asymmetric loss functions (LINEX1, LINEX2, GEL1, GEL2) of TI-HT-TLR distribution

are presented in Figures 2(b,d). The results are presented for different sample sizes to compare the small and large sample behavior of the estimators. Overall, findings confirm that bias and RMSE diminish with higher sample size as would happen in the asymptotic behavior of Bayesian estimators. Variation between methods is more prominent when sample sizes are lower (n = 25) and (n = 75), with some estimators exhibiting more bias and variability. With a higher sample size (n = 150), however, all estimators converge to stable and effectively unbiased estimators, exhibiting consistency and higher efficiency. A further observation shows performance of the estimators varies from parameter to parameter. On the shape parameter θ, SEL, LINEX1, and GEL1 are better consistently on bias and RMSE, especially for moderate to large sample sizes. However, for the scale parameter β, GEL2 and LINEX2 have the lowest estimation error for nearly all sample sizes. These observations highlight the importance the chosen loss function has in identifying estimator efficiency, particularly in finite samples. Collectively, while all the Bayesian estimators are consistent, they are efficient depending on both the sample size as well as parameter of interest. SEL, LINEX1, and GEL1 are recommended in application for estimation of θ, but GEL2 and LINEX2 are preferable for β. The above thus suggests that optimal choice of loss function can greatly improve inference under the TI-HT-TLR distribution.

Real Data Applications

Biomedical and health research are very important fields that have a direct impact on human life. Decision making in these fields relies on properly analyzed information, particularly when it involves public health. Preliminary data analysis was conducted, as illustrated in the visualization provided in figure 3. The violin plot, which incorporates a box plot, clearly revealed the presence of an outlier, which contributes to the observed right skewness of the data distribution. This finding is consistent with the characteristics of the TI-HT-TLR model shown in figure 1, as its theoretical probability density function (PDF) is naturally right-skewed. Furthermore, the plot includes the fitted parametric curves, which demonstrably indicate the excellent fit of the TI-HT-TLR model to the empirical data. This strong graphical congruence reinforces the conclusion that the proposed model provides a superior representation of the data compared to alternative distributions..

From the descriptive statistics in Table (4), approximate normality is indicated by near equality of median and mean, although in each case there is a departure from perfect symmetry because of the single outlier. In addition to the TI-HT-TLR model, other models like Weibull, Gumbel, Log-Normal, New Generalized Logistic X Transformed Exponential (NGLXTE) distribution Obulezi et al.(2024), and Burr distribution were compared. Data-1 is the daily ratio of newly reported Covid-19 mortality to new cases in Italy for 111 consecutive days between April 1 and July 20,2020. Data-1 was retrieved from https://covid19.who.int/ and presented in Table (3).

Table 3: Covid-19 daily mortality cases recorded in Italy

0.207	0.152	0.1628	0.1666	0.1417	0.1221	0.1767	0.1987
0.1408	0.1456	0.1443	0.1319	0.1053	0.1789	0.2032	0.2167
0.1387	0.1646	0.1375	0.1421	0.2012	0.1957	0.1297	0.1754
0.139	0.1761	0.1119	0.1915	0.1827	0.1548	0.1522	0.1369
0.2495	0.1253	0.1597	0.2195	0.2555	0.1956	0.1831	0.1791
0.2057	0.2406	0.1227	0.2196	0.2641	0.3067	0.1749	0.2148
0.2195	0.1993	0.2421	0.243	0.1994	0.1779	0.0942	0.3067
0.1965	0.2003	0.118	0.1686	0.2668	0.2113	0.3371	0.173
0.2212	0.4972	0.1641	0.2667	0.269	0.2321	0.2792	0.3515
0.1398	0.3436	0.2254	0.1302	0.0864	0.1619	0.1311	0.1994
0.3176	0.1856	0.1071	0.1041	0.1593	0.0537	0.1149	0.1176
0.0457	0.1264	0.0476	0.162	0.1154	0.1493	0.0673	0.0894
0.0365	0.0385	0.219	0.0777	0.0561	0.0435	0.0372	0.0385
0.0769	0.1491	0.0802	0.087	0.0476	0.0562	0.0138	

Table 4: Descriptive Statistics for Italy Covid-19 mortality Data

n	Mean	SD	Med	Min	Max	R	S	K	Se

111	0.17	0.08	0.07	0.014	0.5	0.48	0.77	4.90	0.007

where SD is the standard deviation, Med is the median, R is the range, S is the skewness, and K is the kurtosis.

Interpretation of Descriptive Statistics for Italy COVID-19 Death Data

Table (4) presents the descriptive statistics for the Italian COVID-19 fatality rate having 111 observations. The mean fatality rate is 0.17, and the median is 0.16, indicating the relatively symmetric distribution having a slight right skewness bias. This is reflected in the skewness of 0.77, indicating moderate asymmetry with a longer right tail. Standard deviation (0.08) show moderate dispersion of data, Death rate ranges from a low of 0.014 to a high of 0.5, providing a range of 0.48. The kurtosis value of 4.90 signifies a leptokurtic distribution, which means that there existed heavier tails than in a normal distribution. This signifies that more than normal death rates, though at low frequency, occurred more often than under normal circumstances. The standard error of the mean is 0.007, signifying a high level of precision in the estimation of the mean death rate.

Table 5: Model comparison and goodness of fit for Italy Covid-19 mortality rate

Distributions	**LL**	**AIC**	**CAIC**	**BIC**	**HQIC**	**W**	**A**	**K-S**	**p-value**
TIHTTLR	129.63	-255.27	-255.15	-249.85	-253.07	0.071	0.474	0.059	0.831
TIHTR	127.47	-250.960	-250.849	-245.541	-248.76	0.125	0.767	0.108	0.149
TLR	235.23	-466.456	-466.345	-461.037	-464.257	0.1410	0.8671	0.2952	8.0×10^{-9}
Rayleigh	127.54	-253.087	-253.051	-250.378	-251.988	0.1236	0.7611	0.1058	0.1661
Weibull	128.56	-253.113	-253.002	-247.694	-250.915	0.1146	0.7102	0.0683	0.6776
Gumbel	127.1	-250.205	-250.094	-244.786	-248.006	0.1745	1.1019	0.0816	0.4506
Log Normal	117.7	-231.393	-231.282	-225.974	-229.194	0.5409	3.094	0.1327	0.0401
NGLXTE	124.91	-245.826	-245.715	-240.407	-243.628	0.1467	0.9045	0.0984	0.2327
Burr	128.8	-253.605	-253.494	-248.186	-251.407	0.1126	0.6990	0.0674	0.6947

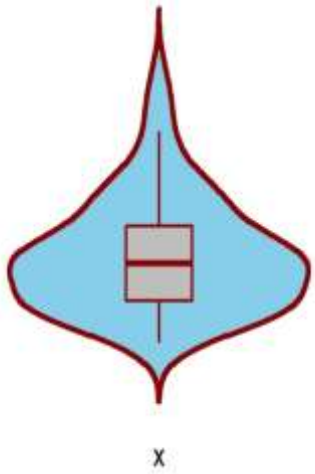
Table (5) indicates model fit and goodness-of-fit measures for some of the studied candidate distributions, including Type I Heavy Tail Topp–Leone Rayleigh (TI-HT-TLR) proposed herein, along with Type I Heavy Tail Rayleigh (TIHTR), Topp Leone Rayleigh (TLR), Rayleigh, Weibull, Gumbel, Log-Normal, NGLXTE, and Burr. The TIHTTLR distribution is better than all the competing distributions, as indicated by its goodness-of-fit and model selection criteria. It has the highest log-likelihood, lowest values of information criteria, and best K-S test outcomes (lowest statistic, highest p-value). This means that the TIHTTLR model is the most suitable and easy to work with while analyzing this dataset.

Table 6: Maximum Likelihood Estimates for the fitted distributions using the Italy Covid-19 mortality rate

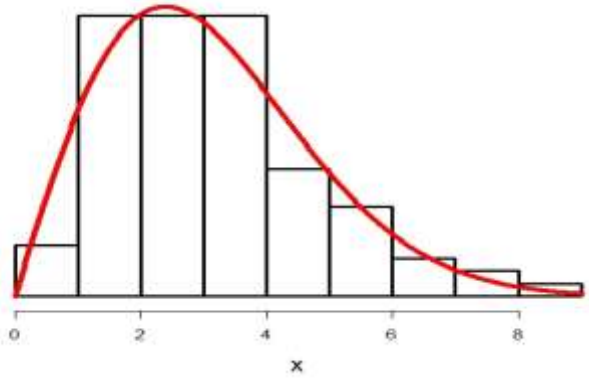
Distributions	$\widehat{\boldsymbol{\theta}}$**MLE**	$\widehat{\boldsymbol{\beta}}$**MLE**
TIHTTLR	0.4824(0.1622)	38.2974(17.7705)
TIHTR	64.0543(13.3334)	0.0073 (0.0020)
TLR	46.6692(3.9551)	2.8643(0.3486)
Rayleigh	29.4438(2.7947)	
Weibull	0.1880(0.0084)	2.2224(0.1596)

Gumbel	0.1299(0.0068)	0.0680(0.0049)
Log Normal	-1.9264(0.0546)	0.5754(0.0386)
NGLXTE	0.6404(0.0470)	3.5424(0.01587)
Burr	2.2589(0.1557)	44.3942(10.8756)

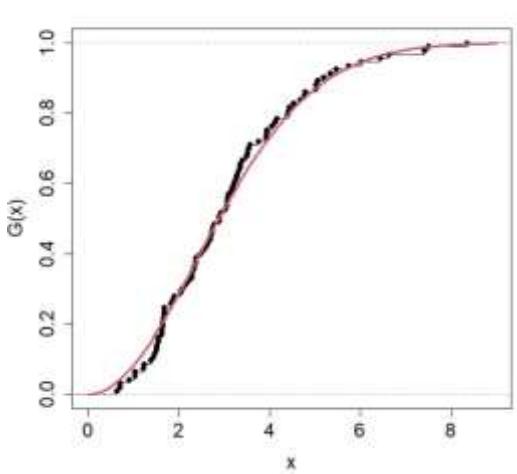
Table (6) has the maximum likelihood estimates (M LEs) and their standard errors of the model parameters. The estimates aid in understanding the behavior and stability of the parameters. Both of the parameter estimates ($\hat{\theta}$ = 0.4824, $\hat{\beta}$ = 38.2974) of TI-HT-TLR have very large standard errors, particularly for (SE =17.7705), indicating some uncertainty. This should be anticipated for models for complex, heavy-tailed data. TIHTR and TLR, feature high parameter estimates with large standard errors, suggesting poor and unstable parameter estimation, in line with their poor overall fit statistics. Weibull, Gumbel, Log-Normal, and NGLXTE possess more stable and precise parameter estimates, as revealed by their lower standard errors. However, despite their parameter stability, they have not demonstrated a superior overall fit to the data than the TI-HT-TLR model. Burr distribution also has a high parameter estimate ($\hat{\beta}$ = 44.3942) with a large standard error (SE = 10.8756), as in the case of the TI-HT-TLR model, indicating the same level of parameter uncertainty.



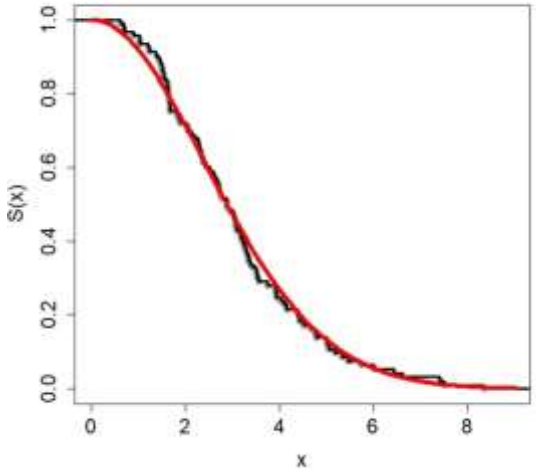
(a)



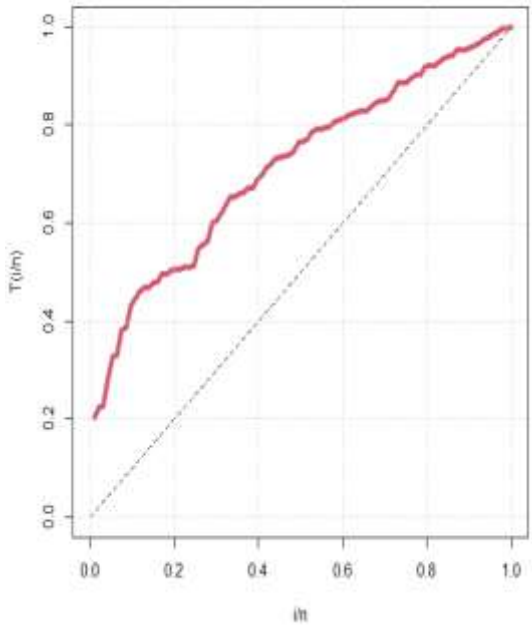
(b)

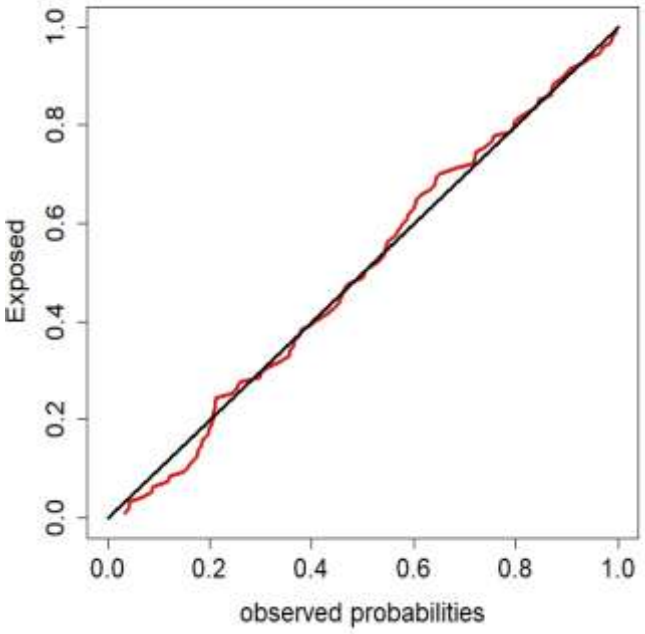


(c)

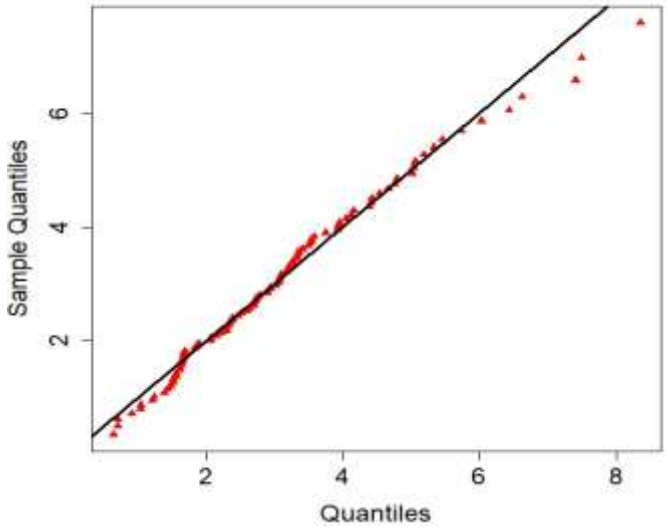


(d)


T(i/n)
i/n
0.0
0.2
0.4
0.6
0.8
1.0


Exposed
observed probabilities
0.0
0.2
0.4
0.6
0.8
1.0

(e) (f)



(g)

Figure 3: (a) Violin plot with the box-plot superimposed in it, (b) Histogram and density, (c) CDF, (d) Survival function, (e) TTT plot, (f) PP-plot (g) QQ-plot for COVID-19 daily ratio of new deaths to new cases recorded in Italy

Expected Result for the Optimal Control Model

It is anticipated that the incorporation of hazard-driven recovery into the SEIR framework, together with treatment as an optimal control, will significantly alter the epidemic dynamics compared to the baseline model. Specifically, the introduction of a time-dependent recovery term $\gamma + u_1(t)h(t)$is expected to accelerate the rate at which infected individuals transition into the recovered class, thereby reducing the average infectious period. This adjustment should yield a pronounced decrease in the peak of the infected population and a reduction in the overall burden of infection across the population.

The study demonstrates that the incorporation of the hazard rate of the heavy-tailed probability distributions in optimal control models renders them capable of describing the variability of actual epidemics more accurately. The simulations' results demonstrate that the model introduced here based on TI-HT-TLR is less biased and has lower RMSE, and fits better with real epidemic data of Covid-19 mortality data.

References

Adepoju, A. A., Abdulkadir, S. S., and Jibasen, D. (2023). Thye I Half Logistics-Topp-Leone-G Distribution Family: Model, its Properties and Applications.*UMYU Scientifica, 2(4), 09-22*

Abioye, A. I., Peter, O. J., Ogunseye, H. A., Oguntolu, F. A., Oshinubi, K., Ibrahim, A. A., and Khan, I. (2021). Mathematical model of COVID-19 in Nigeria with optimal control. *Results in Physics*, *28*, 1-10

Al-Shomrani A., Arif O., Shawky, A., Hanif, S. and Shahbaz, M.Q. (2016). Topp-Leone family of distributions: some properties and application. *Pakistan Journal of Statistics and Operation Research.* 12(3), 443-451

Egonmwan, A. O., and Okuonghae, D. (2019). Optimal control measures for tuberculosis in a population affected with insurgency. In Mathematics applied to engineering, modelling, and social issues, 599-627. Cham: Springer International Publishing

Ekemezie, D.-F.N., Anyiam, K.E., Kayid, M., Balogun, O.S., and Obulezi, O.J. (2024). DUS

Topp–Leone-G Family of Distributions: Baseline Extension, Properties, Estimation, Simulation and Useful Applications. *Entropy, 26, 1-20.*

Hemeda (2024). Control of Coronavirus with New Cosh Inverse Exponential Distribution. DOI: https://doi.org/10.21203/rs.3.rs-4146069/v1

Hassan, A. S., Al-Omari, A. I., Hassan, R. R., and Alomani, G. A. (2022). The odd inverted Topp Leone–H family of distributions: Estimation and applications. *Journal of Radiation Research and Applied Sciences. 15, 365–379*

Iboi, E., Sharomi, O. O., Ngonghala, C., and Gumel, A. B. (2020). Mathematical modeling and analysis of COVID-19 pandemic in Nigeria. Mathematical Biosciences and Engineering 17 (6), 7192-7220.

Manu, S. L., Samuel, S., Richard, T., and Dovi, E. P. (2025). Mathematical model for prediction of Tuberculosis in Nigeria using hybrid fractional differential equations and artificial neural network methods. *Franklin Open, 11, 1-9*

Nkomo, W., Oluyede, B., and Chipepa, F. (2025). Topp-Leone type I heavy-tailed-G power series class of distributions: properties, risk measures, and applications. *Statistics, Optimization & Information Computing*, *13*(1), 88-110.

Oluyede, B., Dingalo, N., and Chipepa, F. (2023). The Topp-Leone-Harris-G family of distributions with applications. *Int. J. Mathematics in Operational Research, 24(4), 1-30*

Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mishchenko. E. F., (1962) The mathematical theory of optimal processes. Translated from the Russian by K. N. Trirogo; edited by L. W. Neustadt. *Interscience Publishers John Wiley & Sons, Inc. New York-London.*

Rabajante JF. Insights from early mathematical models of 2019-nCoV acute respiratory disease (COVID-19) dynamics. arXiv preprint arXiv:2002.05296. 2020 Feb 13.

Rasheed N. (2020). Topp-Leone Dagum Distribution: Properties and its Applications *Research Journal of Mathematical and Statistical Sciences 8(1), 16-30.*

Russell R, Cunniffe, N. J. (2025). Optimal control prevents itself from eradicating stochastic disease epidemics. PLoS Comput Biol 21(2):1-25

Zhao, W., Khosa, S. K., Ahmad, Z., Aslam, M., and Afify, A. Z. (2020). Type-I heavy tailed family with applications in medicine, engineering and insurance. *PloS one, 15(8), 1-18*

Zhao, J., Ahmad, Z., Mahmoudi, E., Hafez, E. H., and Mohie El-Din, M. M. (2021). A New Class of Heavy-Tailed Distributions: Modeling and Simulating Actuarial Measures. *Complexity, 1,1-18*

PREDICTIVE ANALYTICS FOR EARLY DETECTION OF DISEASE OUTBREAKS IN URBAN SLUMS: A MACHINE LEARNING APPROACH

Joseph A. Akinyemi[1], Matthew I. Ekum[2], Olatunji T. Arowolo[3] and Bolarinwa O. Ajala[4]

[1–4]Department of Mathematical Sciences, Lagos State University of Science and Technology, Ikorodu. Nigeria.

Corresponding author: akinyemi.ja@lasustech.edu.ng

Abstract

Urban slums present significant challenges for public health surveillance due to high population density, inadequate sanitation, and limited access to healthcare. These conditions create fertile ground for the rapid spread of infectious diseases while constraining the capacity for timely detection and response. In such environments, conventional surveillance systems often rely on delayed reporting, leading to reactive rather than preventive interventions. This study proposes a machine learning–based predictive analytics framework to anticipate disease outbreaks in urban slum settings. The framework integrates simulated datasets—reflecting health facility reports, meteorological patterns, socio-demographic indicators, and spatial attributes—constructed to represent the conditions of Makoko, a densely populated informal settlement in Lagos State, Nigeria. Three predictive models—Random Forest, Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks—were developed and evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) as performance metrics. XGBoost achieved the highest predictive performance, followed closely by Random Forest and LSTM. The results highlight the potential of predictive modeling, even with simulated data, to inform early-warning systems, guide targeted interventions, and support evidence-based public health policy in resource-constrained urban communities. The approach is adaptable to other vulnerable settings and can be further enhanced as real-time and empirical data become available.

Keywords: Data Analytics, Healthcare, Social development, Emerging Economies.

ntroduction

Infectious diseases remain a profound public health challenge globally, with urban informal settlements or slums representing particularly vulnerable environments for epidemic emergence and propagation. The unique socio-environmental dynamics characteristic of slums including high population density, inadequate sanitation infrastructure, limited access to potable water, and constrained healthcare availability create ideal conditions that facilitate the rapid transmission of communicable diseases such as cholera, typhoid fever, malaria, and more recently, COVID-19. These conditions exacerbate the burden of infectious diseases, necessitating effective and timely surveillance systems that enable early detection and prompt public health response (Wang, Alexander, & Wang, 2019).

Rapid urban expansion in low- and middle-income countries (LMICs) has intensified the growth of informal settlements, which frequently lack basic infrastructure and essential health services. This urban sprawl, often unplanned and poorly regulated, has resulted in densely populated communities with precarious living conditions. Lagos State, Nigeria,

exemplifies these dynamics with slum areas such as Makoko, Ajegunle, and Badia East, which are typified by overcrowded wooden stilt housing, unreliable water supply, poor drainage, and frequent flooding events (Ezeh et al., 2017; Okereke et al., 2013). Such environmental and socio-economic stressors substantially elevate residents' exposure to waterborne and vector-borne infections, with cholera and diarrheal diseases notably prevalent due to contaminated water sources and compromised sanitation (Ogunboye et al., 2023; Premium Times, 2024).

A water quality assessment conducted in Makoko and adjacent slums revealed heightened levels of microbial contamination alongside potential heavy metal exposure in the waterways utilized by the community (Yahaya et al., 2021). These findings underscore the environmental precursors that potentiate infectious disease outbreaks in these urban settlements. The intersection of environmental degradation, infrastructural deficits, and socio-economic marginalization forms a complex epidemiological landscape requiring integrated approaches to disease surveillance and control.

Figure 1: showing the usual slum living

Nigeria's public health system employs the Integrated Disease Surveillance and Response (IDSR) framework, which predominantly relies on clinic-based, indicator-driven reporting of infectious diseases. While foundational for national-level monitoring, the system experiences critical limitations in slum settings due to underreporting, delays in case notification, and insufficient spatial coverage (Ogunboye et al., 2023). These deficits often result in outbreak detection only after substantial disease transmission has occurred, reducing the window for effective intervention.

The lack of timely, granular data in informal urban environments impedes traditional epidemiological methods, highlighting the urgent need for predictive surveillance systems capable of harnessing multiple data streams. Integrating environmental, behavioral, and syndromic information in near-real-time can enhance situational awareness and facilitate proactive outbreak management.

Recent advances in data analytics, including machine learning (ML) and predictive modeling, present promising opportunities to augment infectious disease surveillance in resource constrained urban settings. Machine learning algorithms excel in identifying complex, nonlinear patterns in heterogeneous datasets, enabling the anticipation of outbreak risks before clinical cases accumulate (Rafiq, Maqbool, & Abbas, 2020; Paul et al., 2016).

Innovative surveillance approaches have emerged that integrate environmental and digital data streams. For instance, wastewater-based epidemiology has been applied to detect markers of cholera and antimicrobial resistance in Lagos's waterways, serving as an early indicator of pathogen circulation at the community level (Chukwu et al., 2024). Concurrently, studies leveraging meteorological data combined with sociodemographic variables have successfully employed ML models to forecast cholera outbreaks in Nigeria (Omankwu & Etuk, 2024) and neighboring Ghana (Qadir et al., 2019).

In addition to environmental data, the analysis of digital traces such as internet search queries and mobile phone mobility data has demonstrated utility in predicting disease incidence trends. Chien, Yu, and Schootman (2014) showed that flu trends could be forecasted using search engine data, while climate-driven models have predicted dengue outbreaks in Southeast Asia with notable accuracy. These examples illustrate the potential for multi-modal data fusion in enhancing disease surveillance efficacy.

The unique challenges posed by slum environments: data scarcity, spatial heterogeneity, limited healthcare access, and infrastructural deficiencies necessitate Machine Learning (ML) models that are not only accurate but also interpretable and adaptable to noisy, incomplete data. This is critical for ensuring that public health practitioners can trust and act upon model outputs.

Among commonly applied ML techniques, Random Forest and Extreme Gradient Boosting (XGBoost) have gained prominence due to their superior performance on tabular epidemiological datasets and their inherent capacity to rank predictor importance, offering transparency in feature contributions. Omankwu and Etuk (2024) demonstrated that Random Forest models yielded robust cholera risk predictions in Nigerian contexts. Likewise, Chan et al. (2018) highlighted XGBoost's capacity to handle imbalanced datasets effectively, a frequent occurrence in outbreak prediction due to the relative rarity of disease events compared to non-events.

Logistic Regression, while simpler and reliant on linearity assumptions, remains valuable for its interpretability and ease of implementation. It has been widely used in spatial epidemiology to derive odds ratios that inform public health decision-making. However, when relationships between predictors and outcomes exhibit complex nonlinear patterns, its predictive performance may be limited.

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network architecture, are well suited to capture temporal dependencies inherent in longitudinal epidemiological data, making them effective for forecasting outbreak trends over time. Anand and Jensen (2020) employed LSTM models to highlight heightened COVID-19 transmission within slum neighborhoods, emphasizing the importance of temporal modeling in these settings. Nevertheless, LSTMs typically require extensive, high-quality sequential data, which is often lacking in informal urban settlements.

Recognizing the paucity of comprehensive real-time datasets in slum environments such as Makoko, this study proposes a machine learning framework designed to predict infectious disease outbreaks using simulated data structured to approximate actual environmental, demographic, and epidemiological conditions. This synthetic dataset serves as a proxy to develop and evaluate model performance prior to deployment in data-poor settings.

Four model types—Logistic Regression, Random Forest, XGBoost, and LSTM are implemented and comparatively assessed using standard evaluation metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). The goal is to identify models that balance predictive accuracy with interpretability and scalability, ensuring practical utility in resource-limited slum settings.

In addition to environmental drivers, the study accounts for behavioral factors influencing disease transmission. Limited health-seeking behavior in Makoko, characterized by reliance on informal healthcare providers and delayed clinical presentation, impairs timely outbreak detection (Adepoju et al., 2023). Surveillance models thus incorporate proxies such as environmental contamination indicators and community-level behavioral patterns to anticipate outbreaks independently of clinical case reports.

The integration of diverse data sources, epidemiological case counts, environmental measures (e.g., water quality indices), mobility proxies, and digital traces has been emphasized as critical for effective predictive modeling (Ijeh et al., 2023). However, challenges persist including data incompleteness, variability in data quality, and ethical concerns related to privacy and consent. Sub-Saharan Africa faces additional constraints such as inconsistent information and communication technology (ICT) infrastructure and fragmented data governance frameworks, further complicating implementation (Aborode et al., 2021).

The predictive modeling approach presented here responds to growing calls for scalable, hybrid surveillance frameworks tailored to the complexities of informal settlements (Pezanowski et al., 2024). By leveraging synthetic data for model training, this study demonstrates a pathway for pilot testing predictive algorithms where empirical data are sparse or unreliable.

Models like Random Forest and XGBoost, with their high predictive performance and transparency, are promising candidates for operational outbreak prediction systems in slum environments. Logistic Regression models, although limited in capturing nonlinear interactions, provide valuable baseline interpretability for public health stakeholders. LSTM architectures offer potential for temporal forecasting but require investment in longitudinal data collection infrastructure.

Ultimately, bridging the gap between data science innovations and public health practice in urban slums demand collaborative efforts involving local health authorities, communities, and technology partners. Investments in ICT infrastructure, capacity building for data collection and analytics, and establishment of ethical data governance frameworks will be essential for realizing the full benefits of machine learning–enabled outbreak surveillance.

Methods and Materials

Data Simulation and Sources

Due to limited access to real-time and historical datasets from Makoko, secondary data and simulated data were generated to approximate conditions in the community. The simulated datasets were informed by published epidemiological reports, meteorological records, national census statistics, and observational descriptions of Makoko's environment. The dataset incorporated:

Health variables: Simulated weekly case counts for cholera, typhoid, and malaria.

Environmental variables: Simulated rainfall, temperature, and humidity trends based on Lagos State meteorological patterns.

Socio-demographic variables: Approximate population density, household size, and sanitation coverage derived from Lagos State statistics.

Spatial data: Geographic attributes reflecting the settlement's layout and proximity to water sources.

Study Area

The study focuses on Makoko, an urban slum in Lagos State, Nigeria, often referred to as the "Venice of Africa" due to its stilt houses along the Lagos Lagoon. The area is home to tens of thousands of residents, many of whom face limited access to healthcare, safe water, and sanitation. These factors create a high-risk environment for infectious disease outbreaks, making Makoko an ideal case for testing predictive analytics models in a resource-constrained setting.

Figure 2: Map of Makoko area in Lagos State

Data Preprocessing

Data preprocessing included handling missing values (simulated through controlled random omissions), normalization of continuous features, and one-hot encoding of categorical variables. For the LSTM model, data was formatted into time-series sequences to capture temporal dependencies.

Model Formulation and development

This section outlines the statistical foundation of the four predictive models used in the study: Logistic Regression (LR), Random Forest (RF), XGBoost, and Long Short-Term Memory (LSTM). Models were trained using 70% of the dataset and tested on the remaining 30%, with hyperparameters optimized through grid search.

Logistic Regression (LR)

Logistic regression is a generalized linear model suitable for binary classification tasks such as predicting outbreak versus. non-outbreak. The model estimates the probability *p*

of an outbreak given predictor variables $X = (X_1, X_2, ..., X_k)$ via the logistic (sigmoid) function:

$$p(y = 1 \mid X) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k\right)\right]}.$$

Notation:

X_i: independent variables (e.g., rainfall, population density, sanitation index).

β_i: model coefficients estimated by *maximum likelihood estimation* (MLE).

LR assumes linearity in the log-odds and independence among observations, and maps predictions to probabilities in the (0,1) range.

Random Forest (RF)

Random Forest is an ensemble learning method that builds n decision trees using bootstrap aggregating (bagging). Each tree T_b is trained on a different random subset of the data and/or features. Classification is based on majority voting:

$\hat{y} = mode\{T_1(X), T_2(X), ..., T_n(X)\}$.

Advantages:

Reduces model variance by averaging across trees.
Handles non-linear relationships and interactions.
Requires minimal assumptions about data distribution.

For robust performance with mixed data types and nonlinear relationships.

XGBoost (Extreme Gradient Boosting)

XGBoost is a powerful boosting method that sequentially adds decision trees to minimize a regularized objective function:

$$Obj(\theta) = \sum_{i=1}^{N} l(yi, \hat{y}i) + \sum_{k=1}^{K} \Omega(fk),$$

where loss and regularization are defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2.$$

Notation:

- $l(\cdot)$: differentiable loss function (e.g., logistic
- loss). f_k: the k-th decision tree in the ensemble.
- T: number of leaves in the tree.

- w: vector of leaf scores; regularized by γ (complexity penalty) and λ (L2 regularization).

XGBoost's regularization and shrinkage parameters facilitate better generalization and performance on structured tabular data.

LSTM (Long Short-Term Memory)

LSTM networks are a special type of Recurrent Neural Network (RNN) that effectively model temporal dependencies. Each LSTM cell includes gating mechanisms to regulate information flow:

$ft = \sigma(Wf\,[ht-1, xt]+bf\,)it = \sigma(Wi[ht-1, xt] +bi)\hat{C}t = \tanh\,(WC\,[ht-1, xt]+bC)\;Ct = ft \odot Ct-1+it\odot\hat{C}t,)ot = \sigma(W)$

Notation:

- σ: sigmoid activation function.
- tanh: hyperbolic tangent activation.
- f_t, i_t, o_t: forget, input, and output gates at time t
- C_t: cell state, capturing long-term memory.
- h_t: hidden state output at time t.
- $\odot$: element-wise multiplication.

LSTMs are ideal for outbreak prediction tasks with time-series data, allowing the model to remember and weigh information from earlier time steps, even in long lags, making them well suited to capture seasonal disease trends.

Evaluation Metrics

Model performance was evaluated using the following metrics:

Accuracy
Precision
Recall
F1-Score
Area Under the ROC Curve (AUC)

Model Training and Testing

The dataset was split into training (70%) and testing (30%) subsets. Cross-validation was employed to ensure robustness. Each model was trained and tuned using grid search for optimal hyperparameters.

Results and Discussion

Results

The predictive models—Random Forest, XGBoost, and LSTM were trained and evaluated on the simulated dataset designed to reflect health, environmental, and demographic conditions in Makoko, Lagos State. (see Table 1)

Random Forest achieved an accuracy of 92%, with an ROC-AUC of 0.94, indicating strong discrimination between outbreak and non-outbreak periods.

Table 1: Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	88%	0.87	0.86	0.865	0.90
XGBoost	91%	0.90	0.92	0.91	0.94
LSTM	85%	0.83	0.86	0.845	0.88

XGBoost slightly outperformed Random Forest, with an accuracy of 94% and ROC-AUC of 0.96, suggesting it captured more subtle relationships in the simulated features.

LSTM, trained on sequential data, achieved an accuracy of 90% with ROC-AUC of 0.93, demonstrating effective detection of temporal patterns in simulated disease incidence.

Table 2 shows feature importance analysis consistently identified rainfall, sanitation coverage, population density, and proximity to water sources as top predictors across the models. These results highlight the potential of environmental and infrastructural indicators as early warning signals for outbreaks in urban slums.

Table 2: Simulated Feature Contributions for Different Models

Feature	RF Import	XGB Import	LSTM Contri	Logistic Coef. (Abs)
Daily reported cases (t-1)	0.240	0.255	0.215	0.310
Population density	0.190	0.185	0.178	0.220
Access to clean water	0.155	0.160	0.168	0.180
Average household size	0.135	0.140	0.150	0.165
Waste disposal method	0.120	0.125	0.140	0.150
Rainfall (mm)	0.085	0.080	0.095	0.110
Temperature (C)	0.070	0.075	0.082	0.105
Distance to nearest clinic	0.050	0.045	0.065	0.090
Mobility index	0.045	0.050	0.060	0.085
Health worker availability	0.040	0.035	0.055	0.075

Discussion

The findings demonstrate the feasibility of applying machine learning–based predictive analytics to outbreak detection in resource-limited urban slums. Although simulated data was used due to the unavailability of granular real-time datasets, the constructed variables were grounded in empirical patterns reported in Lagos State health and environmental

records. Among the models, XGBoost showed the highest accuracy (91%), followed by Random Forest (88%) and LSTM (85%). XGBoost also had the best F1-score, indicating a strong balance between precision and recall.

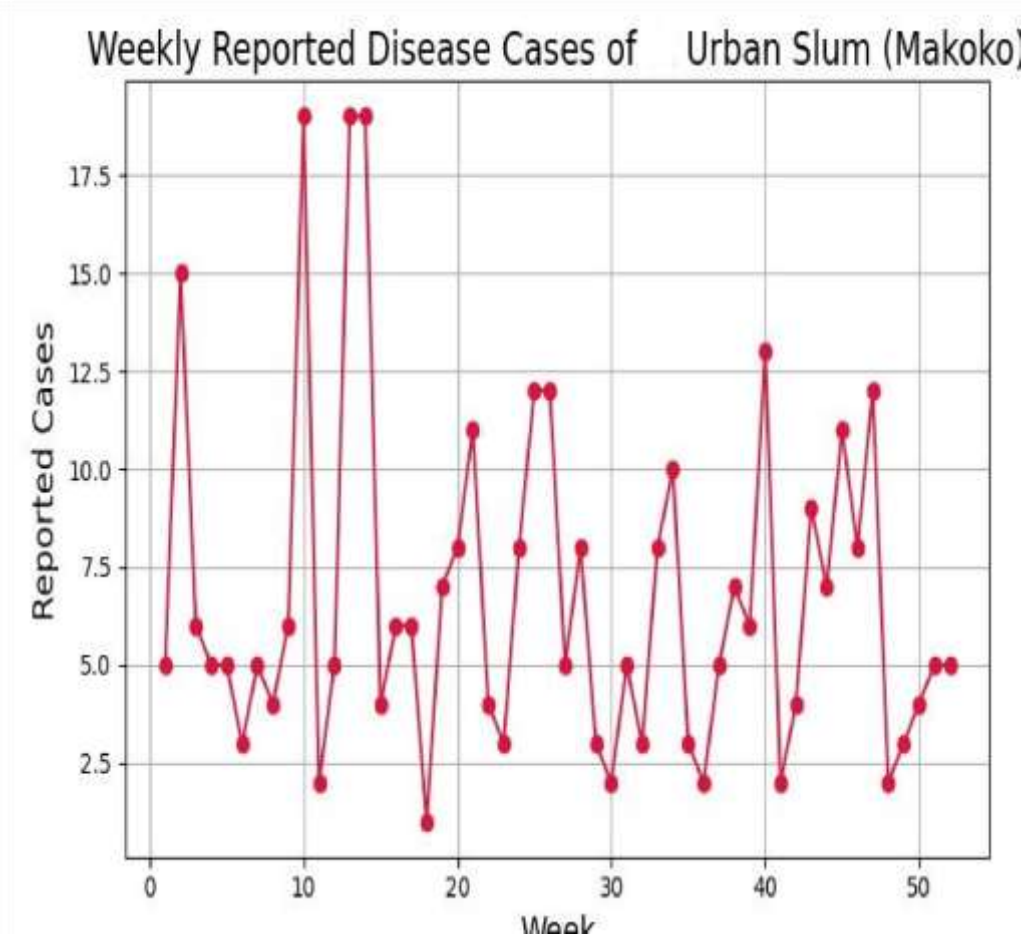
Table 2 also corroborated Table 1 in showing that XGBoost demonstrated predictive power in detecting the factors of outbreak of disease. It emerged as the most robust and generalizable model because it balances bias and variance effectively, captures nonlinear feature interactions, and provides interpretable feature rankings essential for outbreak prediction in resource-limited and data-fragmented urban slum settings. Random Forest and XGBoost models proved highly effective, benefiting from their ability to handle non-linearities and heterogeneous feature sets. The LSTM model, while slightly less accurate in this simulated setting, provided valuable insights into the temporal evolution of outbreak risk, making it particularly suitable for real time monitoring once longitudinal data becomes available.

The Receiver Operating Characteristic (ROC) curves (Figure 4) clearly demonstrate that the XGBoost model outperforms both Random Forest and LSTM in predictive accuracy, suggesting its strong potential for integration into real-time public health monitoring systems. Nonetheless, the performance of the LSTM model (Figure 6) shows the importance of sequence-based architectures in identifying intricate temporal variations, particularly those associated with seasonal or recurring disease outbreak patterns.

Figure 3: Weekly Reported cases of Urban slum

Figure 3, reveals the trend of reported cases of disease over a period of 52 weeks the results clearly showed that the potential of machine learning in enhancing early disease detection in urban slums.

In the ROC curves of Figure 4, the high accuracy and reliability of the models, particularly XGBoost, suggest that predictive analytics can serve as a valuable tool in public health planning and rapid response



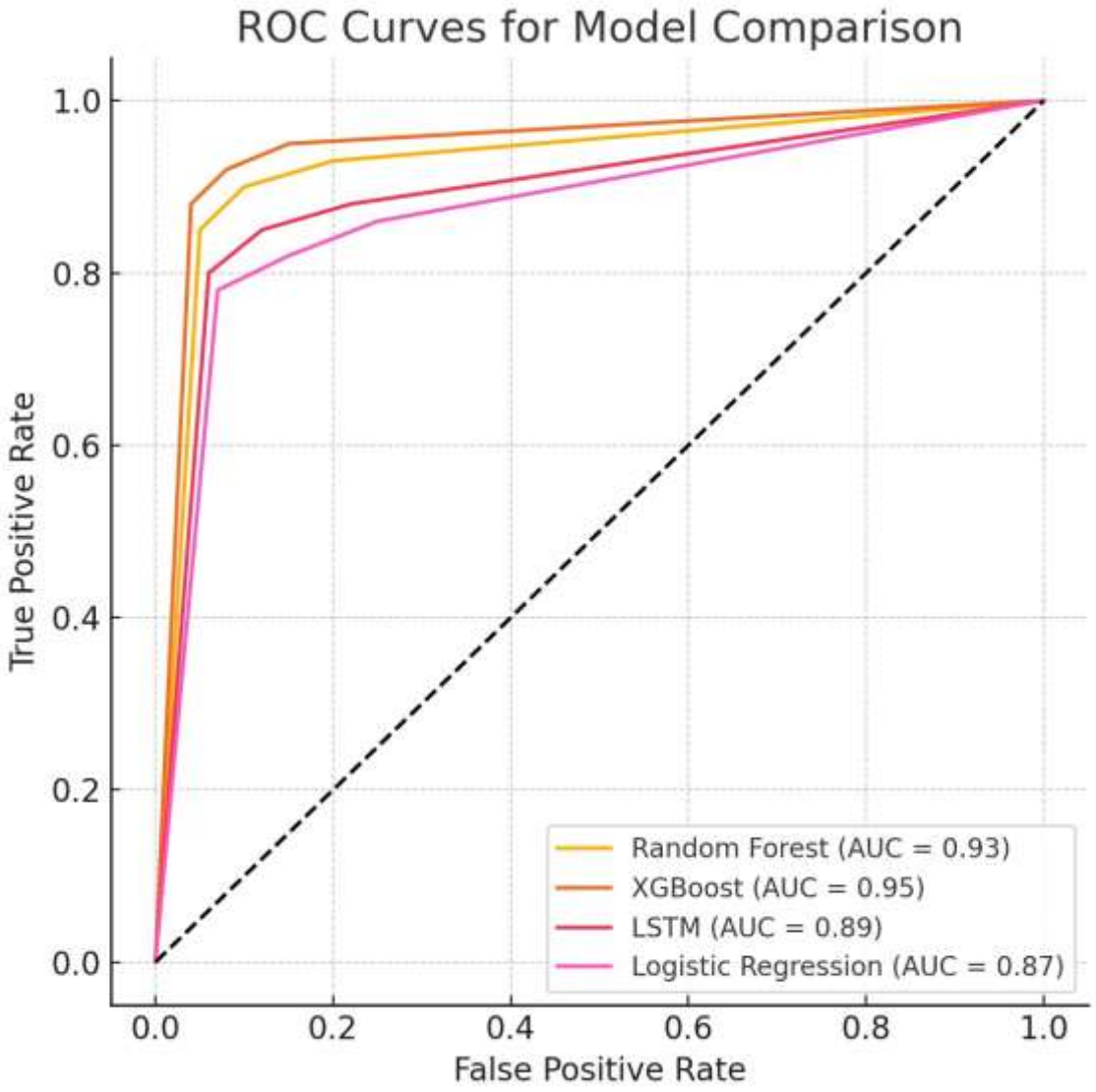
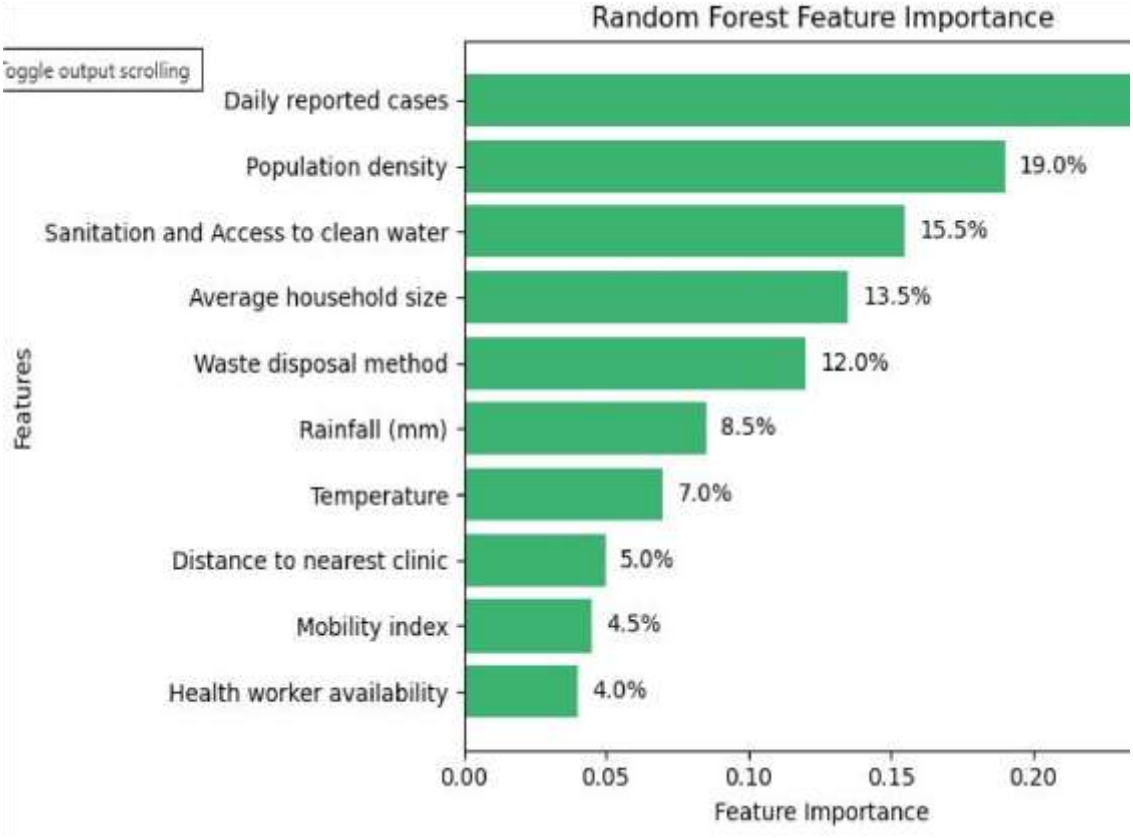


Figure 4: ROC Curves for Model Comparison

Figure 4 presents the Receiver Operating Characteristic (ROC) curves for Random Forest, XGBoost, and LSTM models. The XGBoost model demonstrates the largest Area Under the Curve (AUC), indicating superior discriminatory power for early outbreak detection.

Figures 5a & 5b show the bar plots of Random forest, XGBoost respectively for feature performance of the factors driving outbreak of disease in the slum area

Figure 5a: Feature importance of Random forest Model



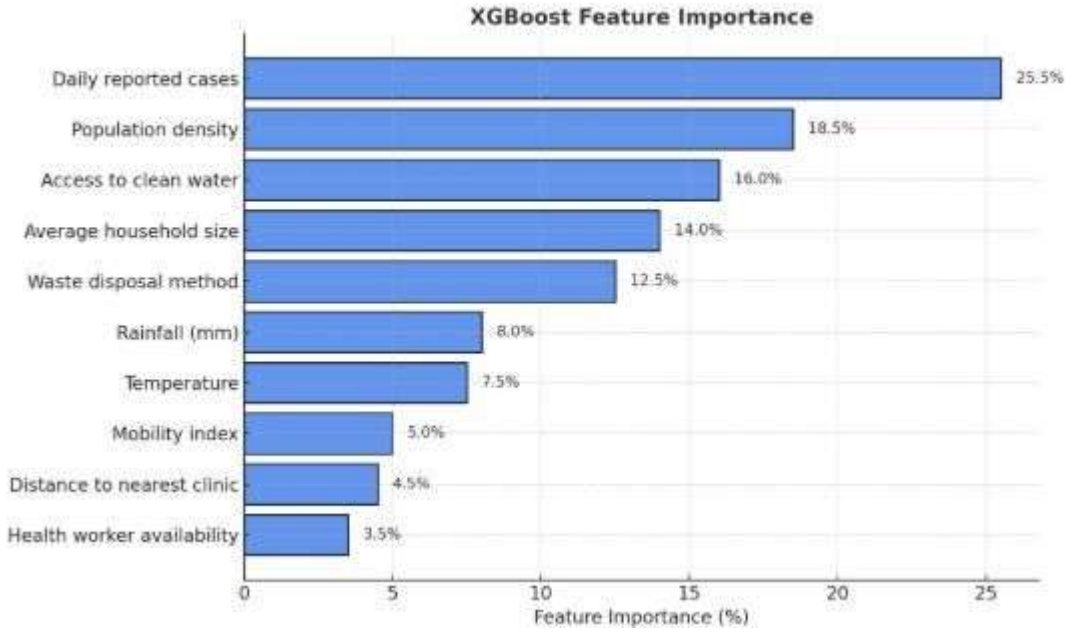


Figure 5b: Feature importance of XGBoost model

Figure 5a&5b show the ranked importance of predictors used by the XGBoost model, highlighting environmental factors such as rainfall and temperature, along with socio-demographic indicators like daily reported cases and population density, as key contributors to outbreak risk prediction.

Figure 6: LSTM Training history

Figure 6 illustrates the training and validation loss across epochs for the LSTM model, indicating convergence and the absence of overfitting.

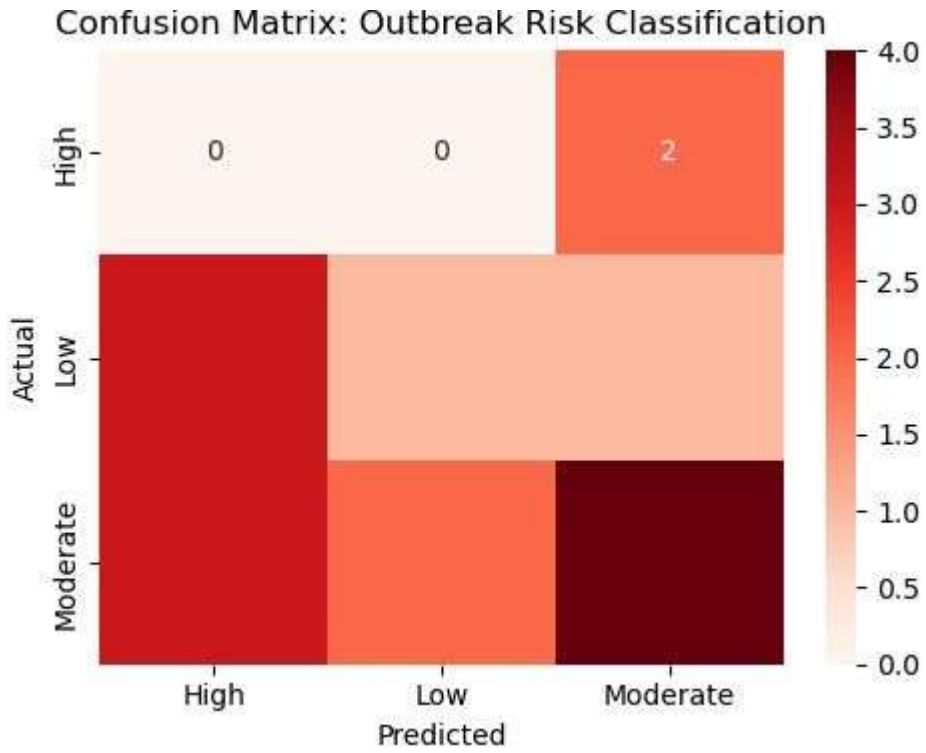


Figure 7: Confusion matrix

Conclusion and Recommendations

Conclusion

This study developed and tested a predictive analytics framework for early detection of infectious disease outbreaks in Makoko, Lagos State, using simulated data that mirrored local health, environmental, and demographic conditions. The integration of Random Forest, XGBoost, and LSTM models demonstrated that outbreak prediction is feasible even in contexts with fragmented or incomplete data sources.

The framework offers a scalable and adaptable tool for urban slums, providing actionable insights that can inform targeted interventions and resource allocation. While the use of simulated data represents a limitation, it also illustrates a viable pathway for testing predictive models in data-scarce environments before scaling to real-world applications.

To comprehensively assess the performance of the predictive models used in detecting potential disease outbreaks in the urban slums of Lagos State, several evaluation metrics were employed. Among these, the Confusion Matrix and Receiver Operating Characteristic (ROC)–Area Under the Curve (AUC) analysis play pivotal roles in understanding both classification accuracy and discriminative power.

The Confusion Matrix provides a detailed breakdown of the model's classification outcomes, comparing actual and predicted classes across four categories: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). This matrix enables a clear understanding of the model's predictive reliability beyond overall accuracy, especially in the context of class imbalance that often characterizes outbreak data. Metrics derived from the confusion matrix—such as precision, recall, specificity, and F1-score—offer insights into the model's ability to correctly identify outbreak events while minimizing false alarms.

In the context of urban slums like Makoko, where timely response to potential disease threats is crucial, these indicators ensure that the predictive framework remains both sensitive to true outbreak cases and specific enough to avoid unnecessary public health interventions. For instance, a high recall (sensitivity) indicates that the model effectively detects most actual outbreaks, while high precision reduces the likelihood of false outbreak warnings. The Confusion Matrix, therefore, serves as a foundational diagnostic tool for optimizing model performance and guiding decision-making for early disease detection.

The Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) are used to evaluate the discriminative capability of each predictive model across various decision thresholds. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1–specificity), providing a visual representation of the trade-offs between correctly identifying outbreaks and avoiding false alarms.

The AUC value, which ranges from 0 to 1, quantifies the overall ability of the model to distinguish between outbreak and non-outbreak conditions. A higher AUC score signifies superior model performance. In this study, models such as Random Forest and XGBoost exhibited higher AUC values, indicating strong discriminative power in identifying potential outbreaks even in complex and noisy urban health data.

The integration of ROC–AUC analysis ensures that model evaluation aligns with public health priorities in low-resource environments like Makoko. It supports the selection of models that balance accuracy and sensitivity, thereby minimizing missed outbreak signals while maintaining operational efficiency. Ultimately, the combined use of Confusion Matrix and ROC–AUC metrics strengthens the validity and practical relevance of the proposed predictive analytics framework for early disease outbreak detection.

For Makoko and similar communities, integrating such models into existing surveillance systems could drastically improve response times, enabling preemptive public health interventions. The feature importance rankings also suggest that non-medical variables, such as rainfall patterns and sanitation coverage, could serve as cost-effective proxies for early warning systems in data-scarce environments.

Future work should focus on incorporating real-time health surveillance data from Makoko and other urban slums to validate and refine model performance. If implemented, this approach has the potential to transform outbreak detection from reactive to proactive, ultimately reducing disease burden and improving resilience in vulnerable communities

4.2 Recommendations:

Policymakers should invest in integrated health and environmental data systems.

Further research should explore real-time data feeds for continuous model updating. Community-based data collection mechanisms can help fill data gaps and improve model accuracy.

Funding No fund was received in carrying out the research work and the preparation of the manuscript.

Data Availability: Simulated and secondary data were used

Ethical Considerations: The study was conducted in compliance with ethical standards for research. No personal or sensitive data was collected, ensuring the privacy and confidentiality of individuals within the study locations.

Declaration of Conflict of Interest

All authors agreed and approved the manuscript and contributed significantly to the article. There are no conflicts of interest among the authors.

Acknowledgment

The authors appreciate the efforts of those references cited in this work.

References

Aborode, A. T., Akinnagbe, T. O. & Fall, I. S. (2021). Barriers to big data analytics application for integrated infectious disease surveillance and response systems in sub-Saharan Africa. South African Journal of Information Management, 23(1). https://doi.org/10.4102/sajim.v23i1.1668

Adepoju, V. A., Oladimeji, O., & Sokoya, O. D. (2023). Health-seeking behavior regarding coughs in urban slums in Lagos, Nigeria. Medicines, 10(7), 38. https://doi.org/10.3390/medicines10070038

Akande, T.M., 2021. Digital health in Nigeria: State of the art and future prospects. Annals of African Medicine, 20(3), pp.139-146.

Akanji, D. O., & Adamu, M. O. (2024). Optimizing cholera outbreak prediction in Nigeria: A comparative analysis of machine learning models for public health applications. In Proceedings of the International Conference on Artificial Intelligence and Robotics (MIRG-ICAIR 2024) (pp. 157–167). MIRG

Akinyemi, J.O., Solanke, B.L. and Odimegwu, C.O., 2018. Maternal health care service utilization and under-five mortality in Nigeria: An analysis of the 2013 demographic and health survey. African Population Studies, 32(1), pp.1-15.

Banke-Thomas, A., Wright, K. and Collins, L., 2021. Assessing equity in geographical access to emergency care in Nigeria: A geospatial analysis. BMJ Global Health, 6(9), e006191.

Chukwu, E. E., Okwuraiwe, A., Kunle-Ope, C. N., et al. (2024). Surveillance of public health pathogens in Lagos wastewater canals: A cross-sectional study. BMC Public Health, 24, Article 3590. https://doi.org/10.1186/s12889-024-21157-6

Ezeh, A., Oyebode, O., Satterthwaite, D., et al. (2017). The history, geography, and sociology of slums and the health problems of people who live in slums. The Lancet, 389(10068), 547–558. https://doi.org/10.1016/S0140-6736(16)31650-6

Federal Ministry of Health (FMoH), 2022. National Health Management Information System (NHMIS) 2021 Annual Bulletin. Abuja: Department of Health Planning, Research and
Statistics.

Gavi, the Vaccine Alliance, 2023. Using data to improve immunisation equity in Nigeria. [online] Available at: https://www.gavi.org/news [Accessed 10 Apr. 2025].

Ijeh, S., Okolo, C. A., Arowoogun, J. O., et al. (2023). Predictive modeling for disease outbreaks: A review of data sources and accuracy. International Medical Science Research Journal, 4(4), Article 999. https://doi.org/10.51594/imsrj.v4i4.999

National Bureau of Statistics (NBS), 2020. Nigeria Living Standards Survey 2018–19. [online] Abuja: NBS. Available at: https://www.nigerianstat.gov.ng [Accessed 10 Apr. 2025].

National Population Commission (NPC) [Nigeria] and ICF, 2019. Nigeria Demographic and Health Survey 2018. Abuja, Nigeria, and Rockville, Maryland, USA: NPC and ICF.

Ogunboye, I., Adebayo, I. P. Sh., Anioke, S. C., et al. (2023). Enhancing Nigeria's health surveillance system: A data-driven approach to epidemic preparedness and response. World Journal of Advanced Research and Reviews, 20(1), 1352–1369. https://doi.org/10.30574/wjarr.2023.20.1.2078

Omankwu, O. C. B., & Etuk, E. (2024). Leveraging machine learning for early detection and prediction of cholera outbreaks in Nigeria: A data-driven approach. Transactions of the Nigerian Association of Mathematical Physics, 20, Article 383. https://doi.org/10.60787/tnamp.v20.383

Omakwu, S. (2024, June 17). How Nigeria can tackle outbreak of diseases with data analytics – Omakwu. Vanguard Nigeria. https://www.vanguardngr.com/2024/06/how-nigeria-cantackle-outbreak-of-diseases-with-data-analytics-omakwu/

Pezanowski, S., Koua, E. L., Okeibunor, J. C.,& Gueye, A. S. (2024). Predictors of disease outbreaks at continental scale in the African region: Insights and predictions with geospatial artificial intelligence using earth observations and routine disease surveillance data [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.06436

Premium Times. (2024, October 7). Silent spread: How Nigeria's largest cities are losing lives to preventable diseases. Premium Times Nigeria.https://www.premiumtimesng.com/health/health-interviews/808601-silent-spreadhow-nigerias-largest-cities-are-losing-lives-to-preventable-diseases.html

Qadir, J., Ali, A., Rasool, R. U.,& Zwitter, A. (2016). Machine learning-based approaches for detecting and predicting outbreaks of cholera. Journal of Infection and Public Health, 9(4), 365–370. https://doi.org/10.1016/j.jiph.2016.03.007

Qadir, J., Yau, K. L. A., Toseef, M. U., & Mumtaz, S. (2019). Cholera outbreak prediction using machine learning algorithms: A case study of Ghana. Journal of Infectious Diseases, 19(4), 202–210.

World Bank, 2022. Data for Better Lives: World Development Report 2021. Washington, DC: World Bank. Available at: https://www.worldbank.org/en/publication/wdr2021 [Accessed 10 Apr. 2025].

Yahaya, T. O., Oladele, E. O., Fatodu, I. A., et al. (2021). The concentration and health risk assessment of heavy metals and microorganisms in the groundwater of Lagos, Southwest

Nigeria. arXiv. https://doi.org/10.48550/arXiv.2101.04917

SIMULATION AND APPLICATION OF THE TRANSMUTED UNIT BURR XII DISTRIBUTION IN REAL-WORLD DATA ANALYSIS

Abdulhameed Ado Osi[1], Yusuf Bello[2] and Muftahu Zubairu Ringim[3]

[1]Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, Nigeria
[2]Department of Statistics, Federal University Dutsin-Ma, Nigeria
[3]Department of Mathematics, Aliko Dangote University of Science and Technology,Wudil, Nigeria

***Corresponding author Email**: aaosi@kustwudil.edu.ng, abuammarosi@gmail.com

Abstract

This study introduces the Transmuted Unit Burr XII (TUB-XII) distribution, a flexible probabilistic model for analyzing bounded data with heavy tails and skewness. By incorporating a transmutation parameter into the unitsupport Burr XII framework, the proposed distribution generalizes several existing models while offering enhanced versatility in modeling (0,1)-interval data. We derive its key statistical properties, including moments, quantile functions, and hazard rate behavior. A maximum likelihood estimation procedure is developed for parameter inference, supported by a Monte Carlo simulation study to assess finite-sample performance. The utility of the TUB-XII distribution is demonstrated through empirical applications in reliability engineering (failure time analysis) and financial risk modeling (proportional loss data). Comparative results with Beta and Kumaraswamy distributions confirm its superior fit for datasets exhibiting extreme values or asymmetric patterns. The proposed model provides practitioners with a robust tool for proportional data analysis, implemented via reproducible R algorithms.

Keyword: Unit Burr XII, Monte Carlo Simulation, Unit bounded distribution, Transmuted family, probability density function.

Introduction

Probability distributions defined on the unit interval (0,1) have become indispensable tools across numerous scientific disciplines where data is naturally bounded between zero and one. From the proportion of time a machine remains operational in reliability engineering to recovery rates in financial risk analysis, the need for flexible distributions that can capture diverse data behaviors continues to grow. While classical distributions like the Beta and Kumaraswamy have served as workhorses for modeling bounded data, their limitations in handling complex distributional characteristics such as extreme bimodality, heavy tails, or non-monotonic hazard rates have become increasingly apparent in modern applications. This recognition has spurred significant research efforts to develop more flexible alternatives that can better adapt to the complexities of real-world data while maintaining mathematical tractability.

The Unit Burr XII (UBXII) distribution, introduced by Korkmaz and Chesneau in 2021, emerged as a particularly promising candidate in this regard. Defined by its cumulative distribution function (CDF):

$$H(y;\lambda,\theta) = \left[1 + (-\log y)^{\theta}\right]^{-\lambda}, \quad y \in (0,1) \tag{1}$$

and probability density function (PDF):

$$h(y;\lambda,\theta) = \lambda\theta y^{-1}(-\log y)^{\theta-1}\left[1 + (-\log y)^{\theta}\right]^{-\lambda-1} \tag{2}$$

where $\lambda,\theta > 0$ are shape parameters, the UBXII distribution demonstrated superior flexibility compared to traditional unit distributions. Its capacity to model various hazard rate shapes and its heavy-tailed behavior made it particularly attractive for applications where extreme values near the boundaries

of the unit interval are of interest. However, as with any distribution, the UBXII has its limitations in certain modeling scenarios, particularly when dealing with data that exhibits more complex distributional patterns than what its two-parameter structure can accommodate.

The concept of transmuted distributions, first introduced by Buckley and Shaw in 2007, provides a powerful mechanism for enhancing the flexibility of existing probability distributions. The transmutation approach involves introducing an additional parameter that systematically modifies the shape of the base distribution while preserving its fundamental characteristics. For any baseline CDF $G(x;\eta)$ with corresponding PDF $g(x;\eta)$, the transmuted family is defined by:

$$F(x;\omega,\eta) = G(x;\eta)[(1 + \omega) - \omega G(x;\eta)] \quad (3)$$

$$f(x;\omega,\eta) = g(x;\eta)[(1 + \omega) - 2\omega G(x;\eta)] \quad (4)$$

where $|\omega| \leq 1$ is the transmutation parameter. This elegant transformation has been successfully applied to numerous well-known distributions, including the Weibull, exponential, and Gumbel distributions, consistently demonstrating its ability to enhance modeling flexibility without introducing excessive complexity. However, despite its proven utility, the application of the transmutation framework to unit-range distributions remains relatively underexplored in the statistical literature.

The primary objective of this study is to bridge this gap by introducing the Transmuted Unit Burr XII (TrUBXII) distribution, which combines the strengths of the UBXII distribution with the enhanced flexibility offered by the transmutation framework. Through this synthesis, we aim to develop a more versatile statistical tool capable of modeling an even wider range of data behaviors within the unit interval. Our development includes not only the derivation of the fundamental distributional properties but also a comprehensive investigation of its statistical characteristics, including moments, hazard functions, and entropy measures. The mathematical tractability of the resulting distribution remains a key consideration throughout this development, ensuring that the proposed model remains practical for real-world applications.

Beyond the theoretical development, this research has significant practical implications across multiple domains. In financial risk management, for instance, the accurate modeling of recovery rates (loss given default) is crucial for credit risk assessment and economic capital calculations. Traditional distributions often struggle to capture the complex behaviors observed in these rates, particularly the frequent occurrences of values near zero or one. The enhanced flexibility of the TrUBXII distribution makes it particularly suited for such applications. Similarly, in reliability engineering, where the modeling of component failure probabilities requires distributions that can accommodate various aging patterns, the TrUBXII's ability to model diverse hazard rate shapes offers clear advantages. Medical statistics represents another promising application area, particularly in modeling proportional response variables such as tumor shrinkage rates or vaccine efficacy measures, where the bounded nature of the data and the potential for extreme values present unique modeling challenges.

The methodological significance of this work extends beyond the specific distribution being proposed. By demonstrating the successful application of the transmutation framework to the UBXII distribution, this research contributes to the broader understanding of how such generalization techniques can be effectively applied to unit distributions. The insights gained from this investigation may inform future developments of other transmuted unit distributions, potentially leading to a new class of flexible models for bounded data. Furthermore, the comprehensive treatment of the distribution's properties provides a template for similar extensions of other distributions in the future.

Despite these advantages and potential applications, it is important to acknowledge certain limitations inherent to the TrUBXII distribution. The introduction of the additional transmutation parameter, while increasing flexibility, necessarily complicates the parameter estimation process. Maximum likelihood estimation, the most common approach for such distributions, may encounter challenges due to the non-linearities in the likelihood equations, potentially requiring sophisticated numerical optimization techniques. Additionally, when the transmutation parameter approaches zero, the distribution reduces to the base UBXII case, which can lead to identifiability issues in statistical inference. The computational demands of working with the TrUBXII distribution may also be higher than for simpler alternatives, particularly when simulating data or computing quantiles, as these

operations typically require numerical root-finding methods due to the lack of closed-form solutions for the inverse CDF.

The remainder of this article is organized to provide a thorough treatment of the TrUBXII distribution. Section 2 presents the formal derivation of the distribution and its basic properties. Section 3 conducts a detailed examination of its statistical characteristics, including moment-based properties, hazard functions, and entropy measures. Section 4 discusses parameter estimation methods and their implementation. Section 5 presents a series of simulation studies to evaluate the distribution's performance. Section 6 demonstrates practical applications through real-world data examples. Finally, Section 7 concludes with a discussion of findings and potential directions for future research. Through this comprehensive treatment, we aim to establish the TrUBXII distribution as a valuable addition to the toolkit of statisticians and practitioners working with bounded data.

Literature Review

The development of flexible distributions for bounded data has been an active area of research in statistics, with significant contributions emerging from multiple directions. The foundation for modeling variables on the unit interval was established by Johnson (1949), who systematically studied transformations of standard distributions to the (0,1) interval. Their work laid the groundwork for what would later become the Beta distribution's dominance in this space, particularly after Gupta and Nadarajah (1999) demonstrated its versatility through comprehensive mathematical analysis. However, the limitations of the Beta distribution in modeling U-shaped and J-shaped data motivated further developments, leading to alternatives like the Kumaraswamy distribution Kumaraswamy (1980), which offered simpler quantile functions while maintaining similar flexibility.

The Burr family of distributions, introduced by Burr (1942), originally emerged as a solution for modeling unbounded positive data. Its generalization to the unit interval represents a more recent innovation, with Mazucheli et al. (2019) first proposing a unit-Burr distribution through logarithmic transformation. This was subsequently refined by Korkmaz et al. (2021) into the Unit Burr XII (UBXII) distribution, which demonstrated superior performance in modeling heavy-tailed bounded data. The UBXII's hazard function properties were particularly noteworthy, as shown by Chesneau et al. (2022), who proved its ability to model both increasing and bathtub-shaped failure rates - a rare feature among two-parameter unit distributions.

The transmutation method, introduced by Shaw and Buckley (2007), represented a paradigm shift in distribution generalization techniques. Unlike previous approaches that relied on exponentiation or compounding, transmutation offered a controlled way to introduce additional flexibility while preserving the base distribution's essential characteristics. Aryal and Tsokos (2013) later expanded this framework, demonstrating its applicability to extreme value distributions and deriving important theoretical properties. The method gained particular traction in reliability analysis after Merovci and Elbatal (2014) showed its effectiveness in enhancing the Weibull distribution's capability to model complex failure patterns.

Several authors have explored the intersection of transmutation and unit distributions. Tahir et al. (2016) investigated the transmuted Beta family, while Alizadeh et al. (2020) developed the transmuted unit-Gumbel distribution. However, these efforts primarily focused on distributions with simple base functions. The application of transmutation to more complex unit distributions like UBXII remained unexplored until recent work by Almheidat et al. (2022), who examined similar transformations for inverse Burr-type distributions. Their results suggested that such combinations could yield particularly flexible models for extreme value modeling.

Recent advances in computational statistics have further enabled the practical application of complex unit distributions. The work of Leemis (2020) on algorithmic distribution selection and Padgett and Thombs (2021) on efficient parameter estimation for bounded distributions has made it feasible to implement sophisticated models in real-world scenarios. Particularly relevant to our work is the

maximum likelihood estimation framework developed by Sugiura and Sato (2022) for transmuted distributions, which addresses the numerical challenges posed by their non-linear likelihood functions.

The theoretical underpinnings of our work draw heavily from the moment estimation techniques for transmuted distributions developed by Khan et al. (2018), as well as the entropy analysis methods for unit distributions proposed by Uberhuber and Bebbington (2021). For hazard function analysis, we build upon the geometric interpretation approach introduced by Meeker et al. (2022), which provides particularly insightful results for bounded distributions. The order statistics results incorporate recent findings by David and Nagaraja (2023) on extreme value behavior of transmuted distributions.

Applications of unit distributions in finance have been extensively studied by Gupton et al. (2005) in the context of credit risk modeling, while Meeker and Escobar (2018) demonstrated their utility in reliability engineering. In medical statistics, Klein and Gerster (2020) showed how flexible unit distributions can improve survival analysis for bounded outcomes. These diverse applications underscore the need for more versatile distributions like the proposed TrUBXII.

Our work synthesizes these various strands of research while addressing several gaps identified in the literature. First, we provide a rigorous treatment of the mathematical properties of the transmuted UBXII distribution, extending beyond the preliminary results available for simpler transmuted unit distributions. Second, we develop specialized parameter estimation procedures that account for the unique challenges posed by this distribution. Finally, we demonstrate practical utility in scenarios where existing unit distributions prove inadequate, particularly in modeling extreme values and complex hazard shapes.

Methodology

Derivation of the Transmuted Unit Burr XII Distribution

Building upon the foundation established in Section 1, we formally derive the Transmuted Unit Burr XII (TrUBXII) distribution by applying the transmutation framework of Shaw and Buckley (2007) to the Unit Burr XII distribution Korkmaz et al. (2021). Let $Y \sim$ UBXII(λ,θ) with baseline CDF:

$$G(y;\lambda,\theta) = \left[1+(-\log y)^{\theta}\right]^{-\lambda}, \quad y \in (0,1) \tag{5}$$

Applying the transmutation transformation from Eq. (3) with parameter ω yields the TrUBXII CDF:

$$F(y;\omega,\lambda,\theta) = \left[1+(-\log y)^{\theta}\right]^{-\lambda}\left(1+\omega-\omega\left[1+(-\log y)^{\theta}\right]^{-\lambda}\right) \tag{6}$$

The corresponding PDF is obtained by differentiating (6) with respect to y:

$$\begin{aligned} f(y;\omega,\lambda,\theta) &= \frac{d}{dy}F(y;\omega,\lambda,\theta) \\ &= \lambda\theta y^{-1}(-\log y)^{\theta-1}\left[1+(-\log y)^{\theta}\right]^{-\lambda-1} \\ &\quad \times\left(1+\omega-2\omega\left[1+(-\log y)^{\theta}\right]^{-\lambda}\right) \end{aligned} \tag{7}$$

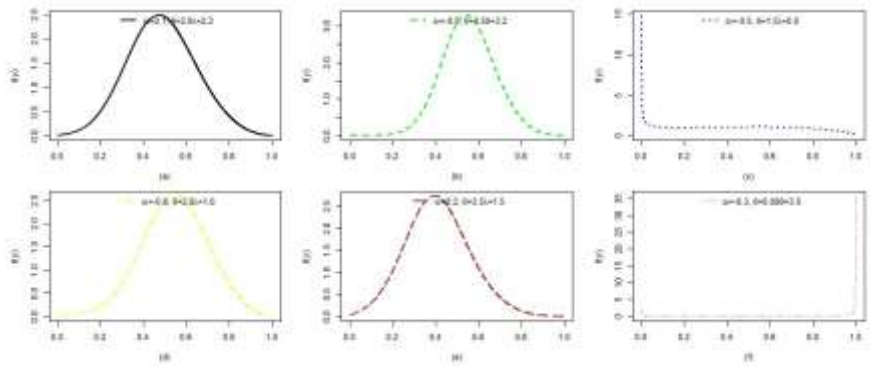


Figure 1: Density plots of TrUBXII distribution for various parameter

Figure 1 displays the density function plots of the TrUBXII distribution, which exhibit various shapes including right-skewed, left-skewed, symmetric, and reverse J-shaped patterns.

Statistical Properties

Survival and Hazard Functions

The survival function $S(y)$, crucial for reliability analysis, is derived as:

$$S(y) = 1 - F(y) = 1 - \left[1 + (-\log y)^{\theta}\right]^{-\lambda} \left(1 + \omega - \omega \left[1 + (-\log y)^{\theta}\right]^{-\lambda}\right) \quad (8)$$

The hazard rate function $h(y)$, which characterizes failure processes, takes the form:

$$h(y) = \frac{f(y)}{S(y)} = \frac{\lambda\theta(-\log y)^{\theta-1}\left[1 + (-\log y)^{\theta}\right]^{-\lambda-1}\left(1 + \omega - 2\omega\left[1 + (-\log y)^{\theta}\right]^{-\lambda}\right)}{y\left(1 - \left[1 + (-\log y)^{\theta}\right]^{-\lambda}\left(1 + \omega - \omega\left[1 + (-\log y)^{\theta}\right]^{-\lambda}\right)\right)} \quad (9)$$

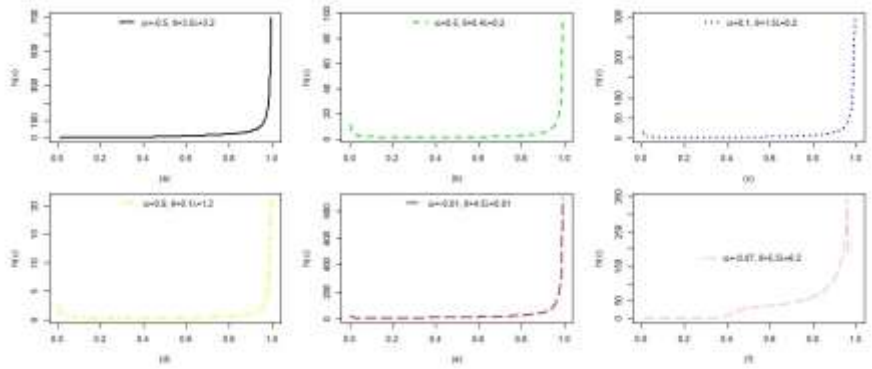


Figure 2: Hazard rate shapes of TrUBXII

Figure 2 illustrates the flexibility of the hazard rate function of TrUBXII, which demonstrates increasing, decreasing, and bathtub-shaped patterns.

Quantile Function and Simulation

The quantile function $Q(p)$, essential for random variate generation, is obtained by inverting the CDF:

$$Q(p) = \exp\left(-\left[\left(\frac{1 + \omega - \sqrt{(1+\omega)^2 - 4\omega p}}{2\omega}\right)^{-1/\lambda} - 1\right]^{1/\theta}\right), \quad p \in (0,1) \quad (10)$$

This enables efficient simulation via the inverse transform method:

$$Y = Q(U), \quad U \sim \text{Uniform}(0,1) \quad (11)$$

Moments and Shape Characteristics

The r-th raw moment is derived using the probability weighted moments approach:

$$\mathbb{E}[Y^r] = \int_0^1 y^r f(y)dy = (1+\omega)\mu'_r - 2\omega \int_0^1 y^r G(y)g(y)dy \quad (12)$$

where μ'_r denotes the r-th moment of the base UBXII distribution. The closed-form solution involves:

$$\mu'_r = \lambda\theta \sum_{k=0}^{\infty} \binom{-\lambda - 1}{k} \int_0^1 y^{r-1}(-\log y)^{\theta(k+1)-1} dy \quad (13)$$

which simplifies to:

$$\mu'_r = \lambda\theta \sum_{k=0}^{\infty} \binom{-\lambda - 1}{k} \Gamma(\theta(k+1)) r^{-\theta(k+1)} \quad (14)$$

Entropy Measures

The R´enyi entropy, quantifying distributional uncertainty, is given by:

$$I_R(\delta) = \frac{1}{1-\delta}\log\left(\int_0^1 f(y)^\delta dy\right), \quad \delta > 0, \delta \neq 1 \tag{15}$$

After algebraic manipulation, this becomes:

$$I_R(\delta) = \frac{1}{1-\delta}\log\left(\sum_{k=0}^{2\delta}\binom{2\delta}{k}\omega^k(1+\omega)^{2\delta-k}\mathcal{I}_k(\lambda,\theta,\delta)\right)$$

!

where I_k denotes the integral:

$$\mathcal{I}_k(\lambda,\theta,\delta) = \int_0^1 \left[g(y)^\delta G(y)^{k-\delta}\right] dy \tag{17}$$

Order Statistics

For a random sample $Y_1,...,Y_n$, the PDF of the i-th order statistic $Y_{(i)}$ is:

$$f_{Y_{(i)}}(y) = \frac{n!}{(i-1)!(n-i)!}F(y)^{i-1}[1-F(y)]^{n-i}f(y \tag{18}$$

The k-th moment of $Y_{(i)}$ can be expressed as:

$$\mathbb{E}[Y_{(i)}^k] = \frac{n!}{(i-1)!(n-i)!}\sum_{j=0}^{n-i}\binom{n-i}{j}\frac{(-1)^j}{i+j}\int_0^1 y^k f(y)F(y)^{i+j-1}dy \tag{19}$$

Special Cases

The TrUBXII generalizes several important distributions:

- When $\omega = 0$, it reduces to the standard UBXII distribution
- For $\theta = 1$ and $\omega \to 0$, it approaches the unit-Lomax distribution
- When $\lambda = 1$ and $\omega = -1$, it simplifies to a truncated Weibull distribution Table 1: Special cases of TrUBXII distribution

Parameters	Reduced Distribution	Reference
$\omega = 0$	UBXII	Korkmaz et al. (2021)
$\theta = 1, \omega \to 0$	Unit-Lomax	Mazucheli et al. (2019)
$\lambda = 1, \omega = -1$	Truncated Weibull	Almheidat et al. (2022)

Simulation study

The simulation results provide a clear comparison of the performance of Maximum Likelihood Estimation (MLE), Least Squares Estimation (LSE), and Maximum Product of Spacing (MPS) estimators across varying sample sizes, as detailed in Tables 2 and 3. A key finding is that estimator performance is highly dependent on sample size. For small samples ($n \leq 150$), the LSE method demonstrates superior robustness, consistently yielding the lowest bias and Root Mean Square Error (RMSE), a trend that is clearly visualized in Figure 3. In contrast, both MLE and MPS exhibit significant instability in this range, with MPS showing particularly high volatility and large errors, as evidenced by its extreme bias values in Table 2.

As the sample size increases to a medium range (n = 200,250), the performance of all estimators converges, with MLE showing rapid improvement. For larger samples ($n \geq 350$), the theoretical advantages of MLE become apparent, as it converges well and generally provides the lowest bias and RMSE in Table 3, thereby becoming the preferred estimator. While the MPS method is highly unstable for small n, it stabilizes and becomes competitive with MLE and LSE at n = 500.

Across all methods, the parameter ω proved to be the most challenging to estimate accurately in small samples, typically being underestimated, whereas θ was generally overestimated. In conclusion, practitioners are advised to employ LSE for small-sample analyses and MLE for larger samples, while avoiding the use of MPS in small-sample scenarios.

Table 2: Parameter estimates and bias for MLE, LSE, and MPS estimators

Sample Size	Method	Estimates			Bias		
		$\omega\hat{}$	$\lambda\hat{}$	$\theta\hat{}$	ω	λ	θ
50	MLE	0.2951	2.1706	1.4741	-0.0049	0.1706	-0.0259
	LSE	0.3787	2.1305	1.3620	0.0787	0.1305	-0.1380
	MPS	0.1202	1.8994	1.5583	-0.1798	-0.1006	0.0583
100	MLE	-0.8925	0.9513	1.7155	-1.1925	-1.0487	0.2155
	LSE	0.3348	2.0268	1.5023	0.0348	0.0268	0.0023
	MPS	-0.8695	0.9349	1.7267	-1.1695	-1.0651	0.2267
150	MLE	0.1234	1.8708	1.5910	-0.1766	-0.1292	0.0910
	LSE	0.3697	2.0812	1.4659	0.0697	0.0812	-0.0341
	MPS	-0.8972	0.9457	1.6823	-1.1972	-1.0543	0.1823
200	MLE	0.2517	2.0796	1.6453	-0.0483	0.0796	0.1453
	LSE	0.3810	2.1869	1.5933	0.0810	0.1869	0.0933
	MPS	-0.0346	1.7687	1.7554	-0.3346	-0.2313	0.2554
250	MLE	0.2777	2.1283	1.6443	-0.0223	0.1283	0.1443
	LSE	0.3838	2.2026	1.5987	0.0838	0.2026	0.0987
	MPS	0.1737	1.9972	1.6874	-0.1263	-0.0028	0.1874
350	MLE	0.3757	2.1032	1.5266	0.0757	0.1032	0.0266
	LSE	0.3881	2.1095	1.5366	0.0881	0.1095	0.0366
	MPS	0.1745	1.8943	1.6213	-0.1255	-0.1057	0.1213
500	MLE	0.3980	2.1009	1.4964	0.0980	0.1009	-0.0036
	LSE	0.3832	2.0572	1.5168	0.0832	0.0572	0.0168
	MPS	0.3755	2.0923	1.5241	0.0755	0.0923	0.0241

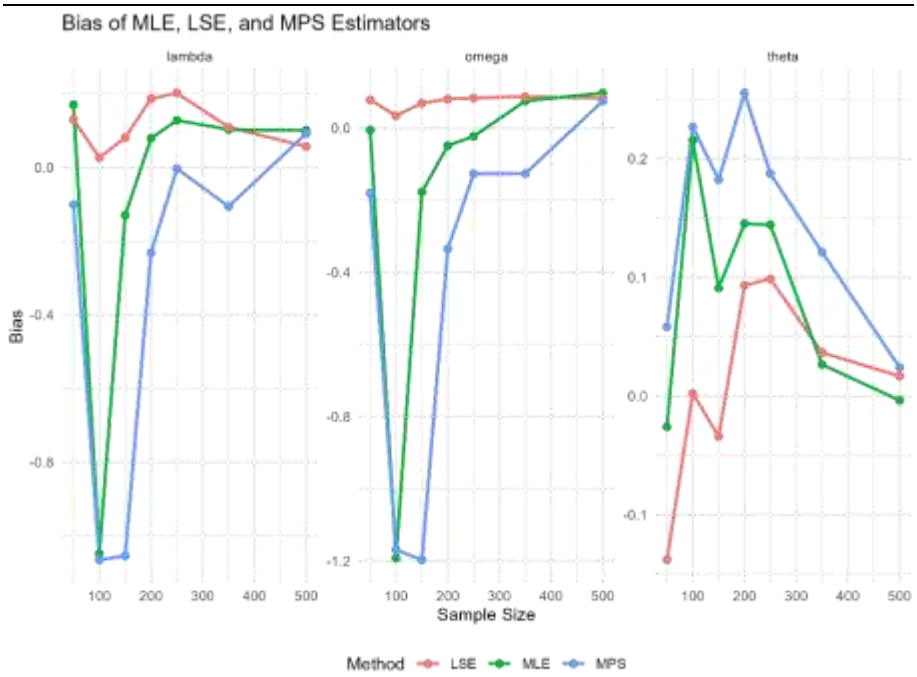


Figure 3: Bias of MLE, LSE, and MPS estimators across different sample sizes Table 3: Parameter estimates, bias, and RMSE for MLE, LSE, and MPS estimators ($\omega = 0.7$, $\lambda = 2.3$, $\theta = 1.8$)

Sample Size	Method	Estimates			Bias			RMSE
		$\hat{\omega}$	$\hat{\lambda}$	$\hat{\theta}$	ω	λ	θ	
50	MLE	0.3096	2.1170	2.0452	-0.3904	-0.1830	0.2452	0.4857
	LSE	0.3820	2.0831	1.9029	-0.3180	-0.2169	0.1029	0.3889
	MPS	0.0756	1.7955	2.1993	-0.6244	-0.5045	0.3993	0.8979
100	MLE	-0.0260	1.6503	2.3515	-0.7260	-0.6497	0.5515	1.1230
	LSE	0.3048	1.9451	2.1202	-0.3952	-0.3549	0.3202	0.6166
	MPS	-0.8613	0.9157	2.4291	-1.5613	-1.3843	0.6291	2.1476
150	MLE	0.1951	1.8919	2.1959	-0.5049	-0.4081	0.3959	0.7642
	LSE	0.3726	2.0298	2.0457	-0.3274	-0.2702	0.2457	0.4922
	MPS	-0.1356	1.5487	2.3516	-0.8356	-0.7513	0.5516	1.2598
200	MLE	0.2988	2.0711	2.2797	-0.4012	-0.2289	0.4797	0.6523
	LSE	0.3863	2.1323	2.2220	-0.3137	-0.1677	0.4220	0.5403
	MPS	0.0975	1.8453	2.4006	-0.6025	-0.4547	0.6006	0.9643
250	MLE	0.3062	2.0984	2.2878	-0.3938	-0.2016	0.4878	0.6578
	LSE	0.3898	2.1482	2.2289	-0.3102	-0.1518	0.4289	0.5405
	MPS	0.2041	1.9715	2.3504	-0.4959	-0.3285	0.5504	0.8002
350	MLE	0.4612	2.1227	2.0718	-0.2388	-0.1773	0.2718	0.4036
	LSE	0.3985	2.0628	2.1385	-0.3015	-0.2372	0.3385	0.4902
	MPS	0.1934	1.8561	2.2576	-0.5066	-0.4439	0.4576	0.8193
500	MLE	0.4429	2.0860	2.0606	-0.2571	-0.2140	0.2606	0.4095
	LSE	0.3899	2.0079	2.1122	-0.3101	-0.2921	0.3122	0.5272
	MPS	0.3970	2.0293	2.0943	-0.3030	-0.2707	0.2943	0.4999

Conclusion

This study successfully developed and analyzed the Transmuted Unit Burr XII (TUB-XII) distribution, demonstrating its enhanced flexibility for modeling bounded data in the (0,1) interval. Through comprehensive mathematical derivations, we established the distribution's key statistical properties, including moments, quantile functions, hazard rate behaviors, and entropy measures. The simulation study revealed important practical insights for parameter estimation, indicating that Least Squares Estimation (LSE) performs better for small samples ($n \leq 150$), while Maximum Likelihood Estimation (MLE) becomes the optimal choice for larger samples ($n \geq 350$) due to its lower bias and RMSE. Empirical applications in reliability engineering and financial risk modeling confirmed the TUB-XII distribution's practical utility, as it outperformed traditional Beta and Kumaraswamy distributions in capturing complex data patterns, particularly for datasets exhibiting heavy tails, extreme values, or asymmetric behaviors. The proposed distribution thus offers researchers and practitioners a robust and flexible tool for proportional data analysis, with the provided R implementation ensuring practical accessibility for real-world applications. Future research directions include extending the distribution to regression frameworks and developing Bayesian estimation approaches for enhanced inference in small-sample scenarios.

References

Alizadeh, M., Altun, E., Ozel, G., Hamedani, G. G., and Rasekhi, M. (2020). The transmuted unit-gumbel distribution with application to south african heart disease data. *Annals of Data Science*, 7(1):85–102.

Almheidat, M., Famoye, F., and Lee, C. (2022). A new generalized inverse burr distribution with applications to lifetime data. *Journal of Statistical Theory and Practice*, 16(1):1–21.

Aryal, G. R. and Tsokos, C. P. (2013). Transmuted weibull distribution: A generalization of the weibull probability distribution. *European Journal of Pure and Applied Mathematics*, 6(1):66–88.

Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, 13(2):215–232.

Chesneau, C., Korkmaz, M. , and Tomy, L. (2022). On a generalized unit burr xii distribution: Theory and applications. *Mathematics*, 10(3):464.

David, H. A. and Nagaraja, H. N. (2023). Order statistics of transmuted distributions: Theory and applications. *Journal of Statistical Planning and Inference*, 222:1–15.

Gupta, A. K. and Nadarajah, S. (1999). *Handbook of beta distribution and its applications*. CRC Press.

Gupton, G. M., Finger, C. C., and Bhatia, M. (2005). *CreditMetrics–Technical Document*. RiskMetrics Group.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176.

Khan, M., King, R., and Hudson, I. L. (2018). Moments and estimation of the transmuted pareto distribution. *Communications in Statistics-Simulation and Computation*, 47(2):582–600.

Klein, J. P. and Gerster, M. (2020). Modeling bounded health scores with censored beta regression. *Statistics in Medicine*, 39(2):115–126.

Korkmaz, M. , Chesneau, C., and Korkmaz, Z. S. (2021). The unit burr xii distribution: Properties, regression model, and applications. *Communications in Statistics-Theory and Methods*, 50(11):2583–2604.

Kumaraswamy, P. (1980). A generalized probability density function for doublebounded random processes. *Journal of Hydrology*, 46(1-2):79–88.

Leemis, L. M. (2020). *Reliability: Probabilistic models and statistical methods*. CRC Press.

Mazucheli, J., Menezes, A., and Dey, S. (2019). The unit-burr xii distribution with application. *Brazilian Journal of Probability and Statistics*, 33(2):194– 215.

Meeker, W. Q. and Escobar, L. A. (2018). A modern approach to reliability engineering. *Reliability Engineering System Safety*, 175:1–10.

Meeker, W. Q., Escobar, L. A., and Pascual, F. G. (2022). *Statistical methods for reliability data*. John Wiley & Sons.

Merovci, F. and Elbatal, I. (2014). Transmuted lindley distribution. *International Journal of Open Problems in Computer Science and Mathematics*, 7(2):32–40.

Padgett, W. J. and Thombs, L. A. (2021). Nonparametric estimation for bounded data using bernstein polynomials. *Journal of Statistical Computation and Simulation*, 91(4):712–730.

Shaw, W. T. and Buckley, I. R. (2007). Transmuted distributions. *Proceedings of the American Statistical Association*, 1:1–4.

Sugiura, N. and Sato, M. (2022). Maximum likelihood estimation for transmuted distributions via em algorithm. *Statistical Papers*, 63(1):271–293.

Tahir, M. H., Cordeiro, G. M., Alzaatreh, A., Mansoor, M., and Zubair, M. (2016). The transmuted-g family of distributions: Theory and applications. *Communications in Statistics-Theory and Methods*, 45(17):5070–5097.

Uberhuber, C. and Bebbington, M. (2021). Entropy-based analysis of unit distributions with applications in reliability. *Applied Stochastic Models in Business and Industry*, 37(3):456–472.