SCOPUA
Journal of Applied Statistical Research
Editor in Chief | Dr. Muhammad Aslam

SCOPUA Journal of Applied Statistical Research

Volume 2 (Issue 1) Published on 2026-03-01

https://doi.org/10.64060/JASR.v2i1

ISSN (e)	3104-4794
DOI	10.64060/JASR
Frequency	Quarterly
Editor in Chief	Prof. Dr. Muhammad Aslam
Contact	jasr@journals.scopua.com

Journal Homepage

https://journals.scopua.com/index.php/JASR

Publisher

Scientific Collaborative Online Publishing Universal Academy

Aims & Scope

SCOPUA Journal of Applied Statistical Research (JASR), ISSN 3104-4794, aims to serve as a comprehensive, peer-reviewed, open-access platform dedicated to advancing knowledge and fostering innovation in the field of applied statistical science. The journal publishes high-quality, original research articles, reviews, and case studies that explore the development and application of statistical methodologies to solve complex real-world problems. Targeting a global audience of statisticians, researchers, and data analysts, the journal seeks to be a leading forum for the dissemination of cutting-edge statistical knowledge and to support the growing demand for advanced statistical applications in various sectors.

Its scope covers a wide range of topics, including statistical theory and methods, biostatistics, statistical computing, data analytics, neutrosophic statistics, and statistical quality control. JASR is particularly focused on interdisciplinary research that bridges statistics with fields such as engineering, social sciences, business, environmental science, and computer science, promoting the integration of statistical tools in diverse areas. The journal welcomes submissions that address contemporary challenges in big data analytics, artificial intelligence, uncertainty modelling, and decision science, with special attention to innovative applications that have tangible societal or industrial impact.

SCOPUA Journal of Applied Statistical Research

Contents (Volume 2 & Issue 1 – 2026)

SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.3

A comprehensive review on the advancement of the Xgamma Distribution

Harsh Tripathi [1*], Mahendra Saha [2], Abhimanyu Singh Yadav [3]

[1]Symbiosis Statistical Institute, Symbiosis International (Deemed University) ,Pune, India

[2]Department of Statistics, University of Delhi, India

[3]Department of Statistics, Banaras Hindu University, Varanasi, India

* Corresponding Email: rsearchstat21@gmail.com

Received: 01 September 2025 / Revised: 06 December 2025 / Accepted: 10 December 2025 / Published online: 18 December 2025

ABSTRACT

The present article provides a comprehensive review of the development and applications of the Xgamma distribution (XGD). The Xgamma distribution, conceptually parallel to the Lindley distribution, emerges as a combination of the exponential and gamma distributions with an appropriate weighting coefficient. Owing to its flexibility and analytical tractability, the XGD has gained considerable attention in reliability and lifetime data modeling. In recent years, numerous researchers have proposed various extensions and generalizations of the Xgamma distribution by employing diverse transformation techniques and distributional families. These developments have further enhanced its modeling capability and adaptability to complex data structures. A particularly interesting aspect of this review is the discussion of practical applications, where several real data sets used in previous studies are examined to demonstrate the empirical relevance and versatility of the extended forms of the XGD across different scientific and engineering fields.

Keywords: Xgamma distribution; finite mixture distribution; extended Xgamma families; transformation-based distributions; real data applications.

1. Introduction

Statistics underpins almost every scientific discipline, including medicine, finance, economics, agriculture, engineering, and the social sciences. Its methods guide evidence-based decision-making, support policy formulation, and help in understanding complex real-world phenomena. Within both academic research and applied work, statistics provides the conceptual and computational tools needed to summarize data, quantify uncertainty, and draw reliable conclusions. The discipline of statistics is broadly structured into several major areas such as descriptive statistics, inferential statistics, design of experiments, time series analysis, statistical quality control, econometrics, regression analysis, and multivariate analysis. In each of these areas, researchers continually develop new methodologies and theoretical results to address emerging data structures and practical demands, thereby contributing both to the advancement of statistical science and to societal welfare. These developments have had substantial impact in diverse sectors, including medical research, the corporate and financial industries, manufacturing, agriculture, and public health, where rigorous statistical tools are now integral to planning, monitoring, and evaluation.

Statistical thinking also plays a crucial role in governance and public policy. Almost all modern government policies ranging from health interventions and educational reforms to economic planning and environmental regulation rely heavily on statistical evidence. Large-scale surveys, censuses, clinical trials, and administrative databases provide the empirical foundation upon which policies are designed, implemented, and assessed. Without sound statistical methodology, such large and complex data sources would be difficult to interpret, and policy decisions would risk being inefficient or misguided. Within parametric statistics, probability distributions form one of the central pillars of theory and practice. They provide mathematical models for random phenomena in virtually every field, including medical survival times, insurance claims, financial returns, rainfall amounts, crop yields, and industrial lifetimes. By specifying a probability distribution, one can characterize the behavior of a random variable and derive key statistical properties such

as moments, quantiles, reliability measures, and risk functionals, which in turn inform inference, prediction, and decision-making.

Because of their wide utility, an extensive body of work has focused on proposing, studying, and extending families of probability distributions. Over time, numerous univariate, bivariate, and multivariate distributions have been introduced, along with systematic "generating" families that allow new models to be constructed from existing ones. Classical and widely used distributions include the exponential, gamma, normal, Weibull, logistic, Burr, binomial, Poisson, geometric, and hypergeometric distributions, among many others. These models arise naturally in different applied contexts, yet none is universally adequate for all types of data, especially when real data exhibit features such as skewness, heavy tails, multimodality, or complex dependence.

In the last few decades, there has been a marked shift toward the development and expansion of distribution theory through generalizations, extensions, and compound or transformed forms of existing models. Authors have proposed new families by introducing additional shape parameters, applying transformations, or compounding baseline distributions with mixing mechanisms to increase flexibility. This line of research is driven by practitioners' need for models that can accurately capture diverse real-world behaviors, such as varying hazard rate shapes in reliability and survival analysis or extreme events in finance and environmental studies. The ongoing search for flexible, robust, and tractable probability distributions continues to motivate the development of new parametric forms that offer better fit, richer interpretation, and wider applicability across scientific and industrial domains.

XGD is a direct outcome of the ongoing search for flexible probability models capable of providing good fit and meaningful interpretation across a wide variety of real-life situations. Since its introduction, the XGD has been employed by many researchers in diverse applied fields, particularly in reliability and lifetime data analysis, where its structural properties and tractable form make it an attractive alternative to classical distributions. The original work of Sen et al. (2016) demonstrated both the superior performance and the appealing theoretical features of the XGD when compared with several competing models, thereby establishing it as a useful baseline distribution for further generalization and application.

Following this foundational contribution, a rich body of work has emerged that focuses on extending and modifying the XGD to enhance its flexibility and adapt it to different data-analytic contexts. Several notable variants have been introduced, including the weighted Xgamma, quasi Xgamma, wrapped Xgamma, inverse Xgamma, and unit Xgamma distributions. Each of these versions is constructed to address specific modeling needs, such as accommodating different support domains, capturing more complex shapes of the probability density function, or allowing for a wider range of skewness and tail behavior. In most cases, the proposed models are rigorously studied in terms of their structural properties, estimation procedures, and reliability characteristics.

A key aspect of this development is the strong empirical support provided through applications to real data sets. For each extended or generalized form of the XGD, authors typically analyze one or more real-life data sets arising from areas such as engineering reliability, biomedical survival times, environmental measurements, or industrial production. These applications not only illustrate the practical relevance of the models but also allow comparison with classical and competing distributions, often showing that XGD-based models yield better fit or more realistic hazard rate behavior. In particular, the flexibility of the hazard rate function capable of capturing increasing, decreasing, bathtub-shaped, or other complex patterns has been central to justifying the use of these extensions in reliability and survival contexts.

Interestingly, the literature on XGD and its extensions has grown continuously since 2016, reflecting sustained interest in this class of distributions. Recent contributions include, among others, works by Wani et al. (2022), Abulebda et al. (2022), Sen et al. (2024), Alomain et al. (2025), Tripathi et al. (2025), and Boudjerda (2025), each proposing new generalized or transformed versions of the XGD and demonstrating their usefulness through theoretical investigations and real-data applications. These studies collectively expand the XGD family into a broad class of models with varying shapes, supports, and dependence structures, thereby enriching the toolbox available to applied statisticians.

In this context, the main aim of the present review is to provide a systematic and coherent summary of all work related to the generalizations and extensions of the Xgamma distribution. The study brings together the various proposed forms of XGD, outlines their construction mechanisms, details their key statistical and reliability properties, and documents their areas of application. By organizing and synthesizing the scattered literature, this review is intended to serve as a comprehensive reference for researchers and practitioners interested in using XGD-based models, as well as a foundation for future developments in the theory and application of flexible lifetime distributions.

Rest of article is organized as follows: In section 2, we discussed regarding the genesis and evolution of XGD. We placed the new generalization or extension of XGD in section 3. We have reported some of data sets which are used to by different researchers in their articles of variety developed XGD in section 4. Conclusive words regarding the presented study are discussed in section 5.

2. Evolution of Xgamma distribution (XGD)

The exponential and gamma distributions are two of the most elegant and widely used models in probability and statistics, and their beauty lies both in their fundamental properties and in their simple functional forms. For an exponential distribution with rate parameter $\theta > 0$,, the probability density function (PDF) and cumulative distribution function (CDF) are

$$f_{\mathrm{Exp}}(x) = \theta \mathrm{e}^{-\theta x}, \quad x > 0, \tag{1}$$

$$F_{\mathrm{Exp}}(x) = 1 - \mathrm{e}^{-\theta x}, \quad x > 0. \tag{2}$$

The exponential distribution is celebrated for its mathematical simplicity and its unique memoryless property, which makes it a natural choice for modeling lifetimes with a constant failure rate and inter-arrival times in Poisson processes, especially in reliability and queueing contexts. The gamma distribution generalizes the exponential model by introducing a shape parameter that allows much richer behavior. For a gamma distribution with shape parameter $k = 3$ and rate parameter $\theta > 0$, the PDF and CDF can be written as

$$f_{\Gamma}(x) = \frac{\theta^3 x^2}{2} \mathrm{e}^{-\theta x}, \quad x > 0, \tag{3}$$

$$F_{\Gamma}(x) = 1 - \mathrm{e}^{-\theta x}\left(1 + \theta x + \frac{\theta^2 x^2}{2}\right), \quad x > 0. \tag{4}$$

By varying the shape parameter in general, the gamma distribution can capture decreasing, increasing, or unimodal density shapes and a wide range of hazard rate patterns, which is particularly useful in reliability, survival analysis, and Bayesian modeling, where conjugacy and tractable moments are highly valued. XGD harnesses the strengths of both exponential and gamma distributions by representing a finite mixture of an exponential distribution with rate θ and a gamma distribution with parameters $(3, \theta)$, with mixing proportions $\theta/(1+\theta)$ and $1/(1+\theta)$, respectively. The resulting PDF is

$$f_{\mathrm{XGD}}(x) = \frac{\theta^2}{1+\theta}\left(1 + \frac{\theta x^2}{2}\right)\mathrm{e}^{-\theta x}, \quad x > 0,\ \theta > 0, \tag{5}$$

and the CDF is

$$F_{\mathrm{XGD}}(x) = 1 - \frac{1+\theta+\theta x+\frac{\theta^2 x^2}{2}}{1+\theta}\mathrm{e}^{-\theta x}, \quad x > 0,\ \theta > 0. \tag{6}$$

Through this mixture construction, the Xgamma distribution retains the tractability and interpretability of its parent distributions while introducing additional flexibility in the shape of the density and hazard rate, making it an appealing model for positive continuous data in reliability, survival, and related fields. The foundational contribution of Sen et al. (2016) formally introduced the XGD, established its key statistical properties, and demonstrated its superiority over several competing models using real data applications in lifetime analysis. Subsequent research has developed a rich family of XGD-based models, including weighted Xgamma, quasi Xgamma, wrapped Xgamma, inverse Xgamma, and unit Xgamma distributions, each designed to address specific modeling needs such as different supports, more complex density shapes, or more flexible hazard rate patterns. More recent extensions and generalizations such as those proposed by Wani et al. (2022), Abulebda et al. (2022), Sen et al. (2024), Alomain et al. (2025), Tripathi et al. (2025), and Boudjerda (2025)—have further expanded the XGD family, providing additional parameters and transformation schemes, and consistently supporting their proposals through detailed theoretical investigations and applications to diverse real-life data sets that highlight the practical power and versatility of Xgamma-based modeling.

3. Different forms of XGD

Motivation is to developed new version of XGD to provide more flexible versions of XGD. In some cases is not a good fit for data or suited well to real situation to overcome this situation, there are various versions of XGD present in literature and make XGD more reliable for the use cases. These new versions allow the researcher to choose one among many good suited versions of XGD distribution. Many authors have taken XGD distribution as a baseline model and generate its different forms by using several transformations and families. All the recent developments of XGD are presented in following table [see, Table 1] along with year and author name.

***Table 1**: List of different forms of XGD*

S.No	Year	Author (Authors)	Model
1	2016	Sen et al.	Xgamma distribution
2	2017	Sen et al.	Weighted Xgamma distribution
3	2017	Sen and Chandra	Quasi Xgamma distribution
4	2018	Sen at al.	Quasi Xgamma Poisson distribution
5	2018	Altun and Hamedani	Log Xgamma distribution
6	2018	Maiti et al.	Discrete Xgamma distribution
7	2019	Hazem and Sen	Wrapped Xgamma distribution
8	2019	Yadav et al.	Inverse Xgamma distribution
9	2019	Sen et al.	Quasi Xgamma-Geometric distribution

10	2020	Cordeiro et al.	Xgamma family
11	2020	Yousof et al.	Xgamma Weibull distribution
12	2020	Bantan et al.	Half-logistic Xgamma distribution
13	2020	Hassan et al.	Lindley-Quasi Xgamma distribution
14	2020	Mazucheli et al.	Discrete Quasi Xgamma distribution
15	2020	Para et al.	Poisson Xgamma distribution
16	2021	Yadav et al.	Exponentiated Xgamma distribution
17	2021	Wani and Shafi	Weighted Lindley-Quasi Xgamma distribution
18	2021	Mohamed et al.	Three parameter Xgamma frechet distribution
19	2022	Shukla et al.	Alpha power transformed Xgamma distribution
20	2022	Tripathi et al.	Marshall-Olkin Xgamma distribution
21	2022	Tripathi and Mishra	Transmuted inverse Xgamma distribution
22	2022	Hashmi et al.	Unit Xgamma distribution
23	2022	Wani et al.	Size Baised Lindley-Quasi Xgamma distribution
24	2022	Abulebda et al.	Bivariate Xgamma Distribution
25	2022	Yadav et al.	Xgamma Exponential Distribution
26	2023	Yadav et al.	Weighted Xgamma Exponential Distribution
27	2024	Sen et al.	Truncated version of Xgamma distribution
28	2025	Alomair et al.	Improved Extension of Xgamma distribution
29	2025	Tripathi et al.	Generalized Inverse Xgamma distribution
30	2025	Boudjerda	New Power Xgamma Distribution

All mentioned models are used in survival studies, reliability theory, censoring scheme, truncation scheme, environmental data etc. It is important to discuss the CDF and hazard rate function (HRF) of each developed model or extension of XGD one by one in detail.

3.1 Weighted Xgamma distribution

Sen et al. (2017) chalked out weighted Xgamma distribution with the concept of weighted distribution. A general form of the weighted PDF is:

$$f(x) = \frac{w(x)f_0(x)}{E[w(x)]}$$

With the baseline PDF of XGD, CDF (G(x)) and HRF (H(x)) of weighted xgamma distribution are:

$$G(x) = \frac{2\eta}{\zeta![2\eta+(1+\zeta)(2+\zeta)]}\left[\gamma(\zeta+1,\eta x) + \frac{1}{2\eta}\gamma(\zeta+3,\eta x)\right]; x > 0, \eta > 0, \zeta = 1,2,3.. \quad (7)$$

and

$$H(x) = \frac{\eta^{\zeta+1}\left(x^{\zeta}+\frac{\eta}{2}x^{\zeta+2}\right)e^{-\eta x}}{\left[\Gamma(\zeta+1,\eta x)+\frac{1}{2\eta}\Gamma(\zeta+3,\eta x)\right]} \quad (8)$$

Authors focused on special case of weighted xgamma distribution, for $\zeta = 1$ and it is formulated as Length biased Xgamma distribution (LBXD). Sen et al. (2017) mentioned that HRF of LBXD is log-concave and having the increasing failure rate nature. Graphical presentation of the HRF of the same placed in Figure 1.

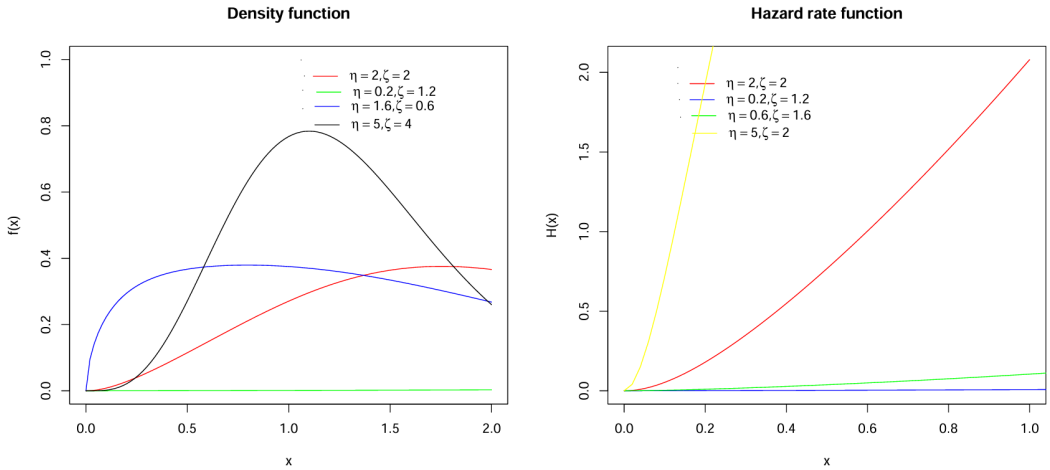

***Figure 1**: Weighted XGD*

3.2 Quasi Xgamma distribution

Sen and Chandra (2017) developed the extension of XGD and called this extension as Quasi Xgamma distribution. Authors have added one more parameter to XGD to make it more flexible. CDF and HRF of Quasi Xgamma distribution are given below:

$$G(x) = 1 - \frac{\left(1+\zeta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\zeta)} e^{-\eta x}, \quad x > 0,\ \eta > 0,\ \zeta > 0 \tag{9}$$

and

$$H(x) = \frac{\eta\left(\zeta+\frac{\eta^2}{2}x^2\right)}{1+\zeta+\eta x+\frac{\eta^2}{2}x^2}, \quad x > 0,\ \eta > 0,\ \zeta > 0 \tag{10}$$

,respectively. They explained the special cases of it along with the statistical properties. Also authors evaluated the mode of Quasi Xgamma distribution and the same displayed by using graphs. Graphical presentation of the HRF of the same placed in Figure 2.

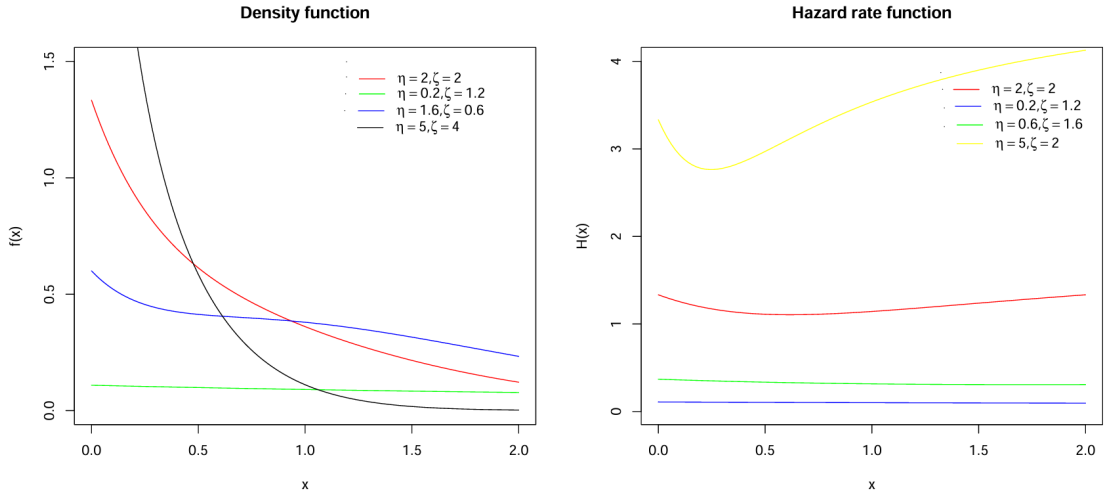

***Figure 2**: Quasi XGD*

3.3 Quasi Xgamma Poisson distribution

Modeling of lifetime data can be done by Quasi Xgamma Poisson distribution. Need of Quai Xgamma Poisson distribution is established by Sen et al. (2018) and showed through real life data. Quasi Xgamma Poisson distribution coincide with different distributions like: Xgamma Poisson, Quasi Xgamma and Xgamma distribution. CDF and HRF are:

$$G(x) = \frac{e^{\lambda} - \exp\left(\frac{\lambda e^{-\eta x}}{(1+\zeta)}\left(\frac{\eta^2}{2}x^2+\eta x+\zeta+1\right)\right)}{e^{\lambda}-1}, \quad x > 0,\ \eta > 0,\ \zeta > 0,\ \lambda > 0 \tag{11}$$

and

$$H(x) = \frac{\frac{\lambda\eta}{(1+\zeta)}\left(\zeta+\frac{\eta^2}{2}x^2\right)\exp\left(\frac{\lambda e^{-\eta x}}{(1+\zeta)}\left(\frac{\eta^2}{2}x^2+\eta x+\zeta+1\right)-\eta x\right)}{\exp\left(\frac{\lambda e^{-\eta x}}{(1+\zeta)}\left(\frac{\eta^2}{2}x^2+\eta x+\zeta+1\right)-1\right)}, \quad x > 0,\ \eta > 0,\ \zeta > 0,\ \lambda > 0 \tag{12}$$

HRF of Quasi Xgamma Poisson distribution can take different shapes which makes it applicable for real life situation and the same illustrated through the graph for various parameters. Graphical presentation of the HRF of the same placed in Figure 3.

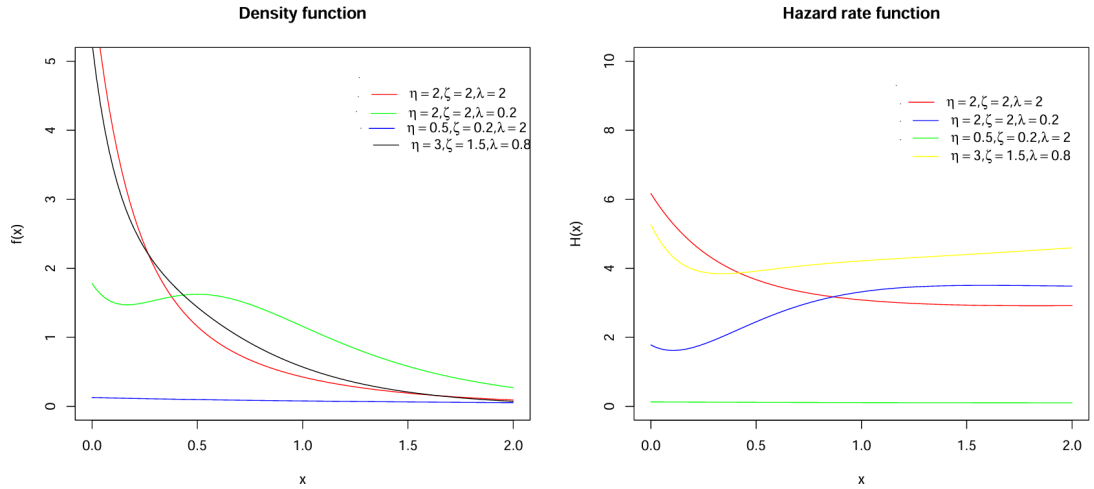

***Figure 3**: Quasi XGD Poisson*

3.4 Log Xgamma distribution

Altun & Hamedani outline the Log Xgamma in 2018. Researchers are very much inclined to develop the probability distributions of unbounded support. Log Xgamma distribution is used in the situation where values are in bounded range such as percentage and proportions. CDF and HRF of the Log Xgamma are given below:

$$G(x) = x^{\eta}(\eta+1)^{-1}\left[1+\eta-\eta log(x)+\frac{\eta^2 log(x)^2}{2}\right]; 0<x<1, \eta>0 \tag{13}$$

and

$$H(x) = \frac{\frac{\eta^2}{1+\eta}\left(1+\frac{\eta}{2}log(x)^2\right)x^{\eta-1}}{1-x^{\eta}(\eta+1)^{-1}\left[1+\eta-\eta log(x)+\frac{\eta^2 log(x)^2}{2}\right]} \tag{14}$$

HRF of Log Xgamma distribution having increasing and bathtub shape property and to validate the same Altun & Hamedani (2018) used a data set those having the range of observation between 0 to 1. Graphical presentation of the HRF of the same placed in Figure 4.

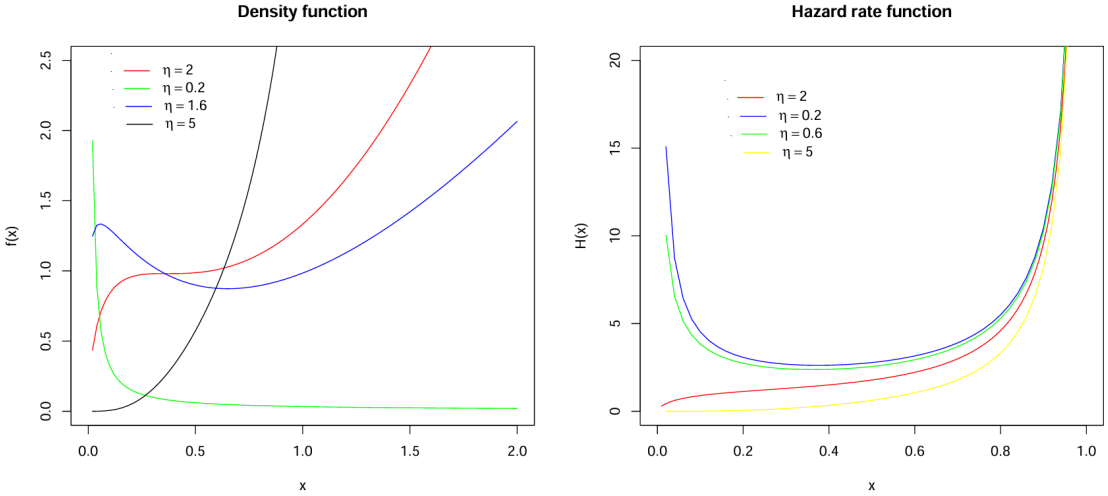

***Figure 4**: Log Xgamma distribution*

3.5 Discrete Xgamma distribution

Maiti et al. (2018) developed the Discrete Xgamma distribution, called as dxgamma-I to model the discrete real life situations and authors continued to strive to derive the statistical properties, estimation and real life application for the Discrete Xgamma distribution. CDF and HRF are:

$$G(x) = 1-\frac{1-logp-logp(x+1)+(log(p))^2/2(x+1)^2}{1-log(p)}p^{x+1}; x = 0,1,2.. \tag{15}$$

and

$$H(x) = \frac{[1+p-e^p(1+2p+p^2/2)+p(1-e^{-p}(1+p))x+(p^2)/2(1-e^{-p})x^2]\frac{e^{-px}}{1+p}}{\frac{1-log(p)-log(p)(x+1)+(log(p))^2/2(x+1)^2}{1-log(p)}p^{x+1}} \tag{16}$$

Graphical presentation of the probability mass function of the same placed in Figure 5.

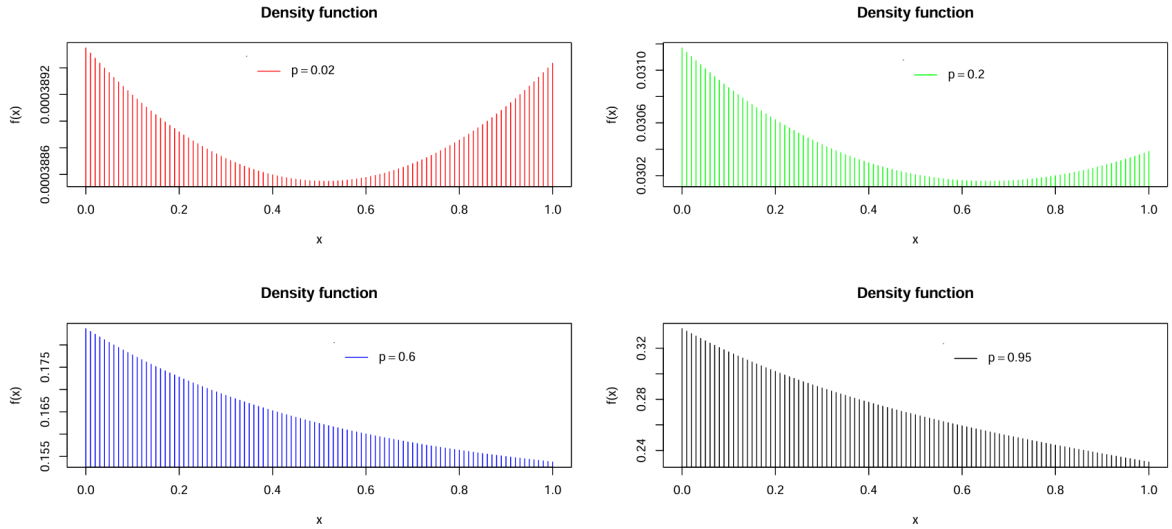

***Figure 5**: Discrete Xgamma distribution*

3.6 Wrapped Xgamma distribution

Hazem and Sen (2019) described the circular version of XGD and denoted as Wrapped Xgamma distribution. Wrapped xgamma distribution is very useful to model the circular data. CDF and HRF are:

$$G(x) = \left(1 - \frac{1+\lambda(1+\eta)+\frac{\eta^2\lambda^2}{2}}{1+\lambda}e^{-\eta\lambda}\right)\frac{1}{1-e^{-2\pi\lambda}} + \frac{2\pi\lambda}{\lambda+1}(1-(1+\eta\lambda))e^{-\eta\lambda}\frac{e^{-2\pi\lambda}}{(1-e^{-2\pi\lambda})^2}$$
$$+\frac{2\pi^2\lambda^2}{\lambda+1}(1-e^{-\eta\lambda})\frac{e^{-2\pi\lambda}(1+e^{-2\pi\lambda})}{(1-e^{-2\pi\lambda})^3}; 0 \le x < 2\pi, \lambda > 0$$

and

$$H(x) = \frac{\frac{\lambda^2 e^{-\eta\lambda}}{(\lambda+1)(1-e^{-2\pi\lambda})}\left(1+\frac{\lambda\eta^2}{2}+2\pi\lambda\left[(\pi-\eta)e^{-2\pi\lambda}+(\eta+\pi)\right]\frac{e^{-2\pi\lambda}}{(1-e^{-2\pi\lambda})^2}\right)}{1-\left(1-\frac{1+\lambda(1+\eta)+\frac{\eta^2\lambda^2}{2}}{1+\lambda}e^{-\eta\lambda}\right)\frac{1}{1-e^{-2\pi\lambda}}+U_1+\frac{2\pi^2\lambda^2}{\lambda+1}(1-e^{-\eta\lambda})\frac{e^{-2\pi\lambda}(1+e^{-2\pi\lambda})}{(1-e^{-2\pi\lambda})^3};0\le x<2\pi,\lambda>0} \quad (17)$$

where, $U_1 = \frac{2\pi\lambda}{\lambda+1}(1-(1+\eta\lambda))e^{-\eta\lambda}\frac{e^{-2\pi\lambda}}{(1-e^{-2\pi\lambda})^2}$

Measurements in direction is common in science and real life data observations. The popular encounter of such data sets might be related to direction of flight of a bird, orientation of certain animals, direction of magnetic field in a place, etc., could be two dimensional or three dimensional depending on the nature of measurements. To support the behavior of HRF, authors have used glaciologist data which mathched to HRF of Wrapped Xgamma distribution.

3.7 Inverse Xgamma distribution

Inverse transformation is used on XGD to develop the Inverse Xgamma distribution and it is developed by Yadav et al. (2019). Following are the CDF and HRF are:

$$G(x) = \left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x}; x>0, \eta>0 \quad (18)$$

and

$$H(x) = \frac{\frac{\eta^2}{1+\eta x^2}\frac{1}{}\left(1+\frac{\eta}{2x^2}\right)e^{-\eta/x}}{1-\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)x^2}\frac{1}{}\right]e^{-\eta/x}} \quad (19)$$

Yadav et al. (2018) plotted the HRF for various η and found that found that it is suitable for non monotone failure pattern which often occurs in clinical trials and reliability studies. Graphical presentation of the HRF of the same placed in Figure 6.

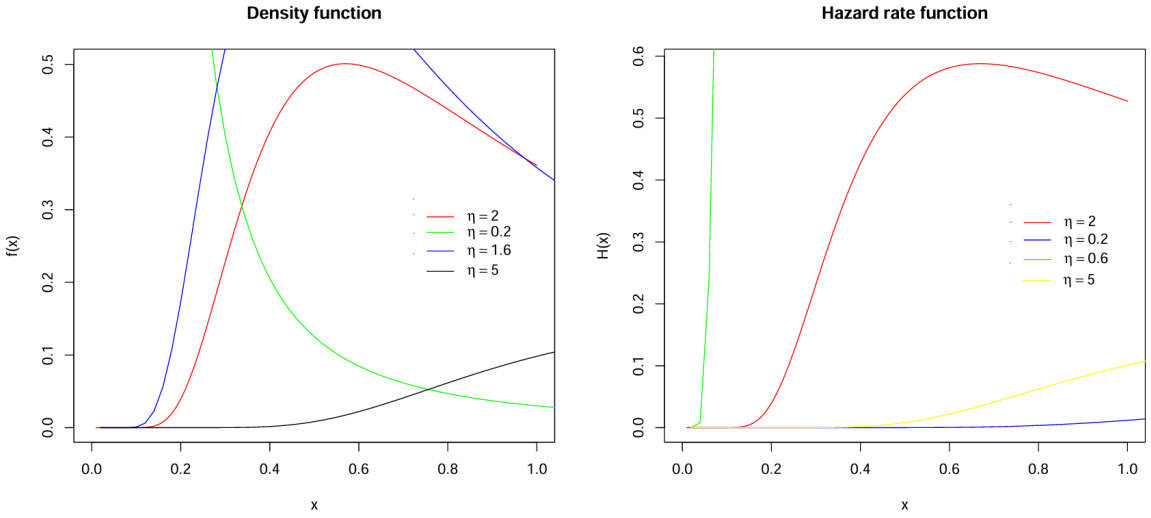

***Figure 6**: Inverse Xgamma distribution*

3.8 Quasi Xgamma-Geometric distribution

It is introduced by Sen et al. (2019) and Quasi Xgamma-Geometric distribution is the result of applying compound technique geometric distribution. CDF and HRF of Quasi Xgamma-Geometric distribution are:

$$G(x) = \left[1 - \frac{e^{-\theta x}\left(1+\alpha+\theta x+\frac{\theta^2}{2}x^2\right)}{(1+\alpha)}\right]\left[1 - \frac{pe^{-\theta x}\left(1+\alpha+\theta x+\frac{\theta^2}{2}x^2\right)^{-1}}{(1+\alpha)}\right],$$

$$x > 0, \theta > 0, \alpha > 0. \tag{20}$$

and

$$H(x) = \frac{\theta\left(\alpha+\frac{\theta^2}{2}x^2\right)}{\left(1+\alpha+\theta x+\frac{\theta^2}{2}x^2\right)}\left[1 - \frac{pe^{-\theta x}\left(1+\alpha+\theta x+\frac{\theta^2}{2}x^2\right)^{-1}}{(1+\alpha)}\right]^{-1} \tag{21}$$

The HRF of the Quasi Xgamma-Geometric distribution can be constant, decreasing, increasing, decreasing-increasing, upside down bathtub or bathtub failure rate shapes. Graphical presentation of the HRF of the same placed in Figure 7.

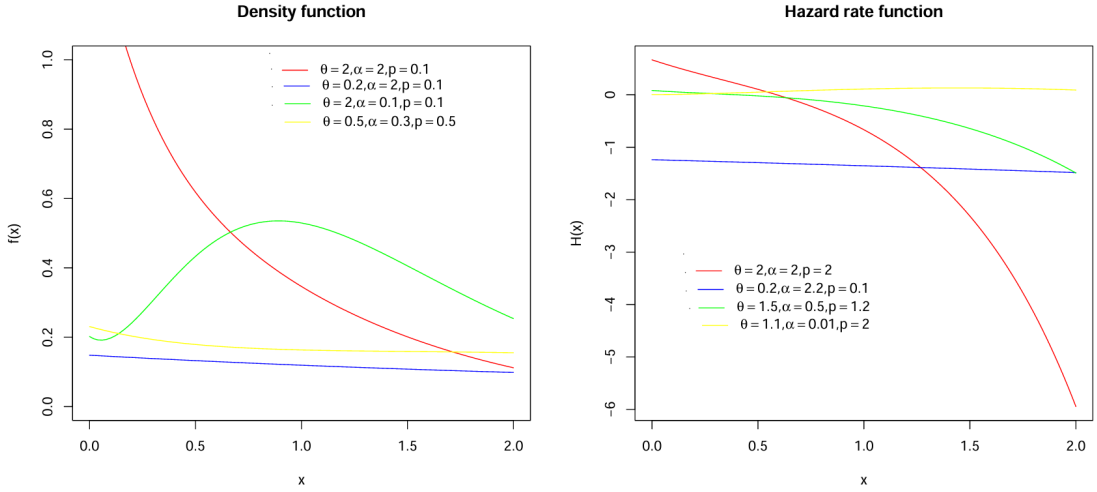

***Figure** 7: Quasi Xgamma-Geometric distribution*

3.9 Xgamma family

Cordeiro et al. (2020) described Xgamma family and developed the censored regression modelling. They introduced the special submodels like Xgamma lindley distribution, Xgamma weibull distribution and Xgamma Burr XII distribution. Glass fiber data, Diabetic retinopathy study and Leukemia data are used to prove the applicability of special submodels of Xgamma family and made the point that HRF of several submodels of Xgamma family are coincide with the real life situations.

3.10 Xgamma Weibull distribution

A new extension of weibull distribution is derived and named as Xgamma Weibull distributions [see, Yousof et al. (2020)] and it can be obtained by just doing the integration of Xgamma distribution with range 0 to $-\log[1 - G_W(x; b)]$. CDF and HRF are:

$$G(x) = 1 - \left[1 + \eta + \eta x^{\lambda} + \frac{\eta^2 x^{2\lambda}}{2}\right]\frac{e^{-\eta x^{\lambda}}}{1+\eta}; x > 0, \lambda > 0, \eta > 0 \tag{22}$$

and

$$H(x) = \frac{\frac{\lambda\eta^2}{1+\eta}x^{\lambda-1}e^{-\eta x^{\lambda}}\left(1+\frac{\eta x^{2\lambda}}{2}\right)}{\left[1+\eta+\eta x^{\lambda}+\frac{\eta^2 x^{2\lambda}}{2}\right]\frac{e^{-\eta x^{\lambda}}}{1+\eta}} \tag{23}$$

Increasing, decreasing and bathtub shape HRF can be accommodate with implementation of Xgamma Weibull distribution and the same trends of HRF are seen into the used data of manufacturing field. Graphical presentation of the HRF of the same placed in Figure 8.

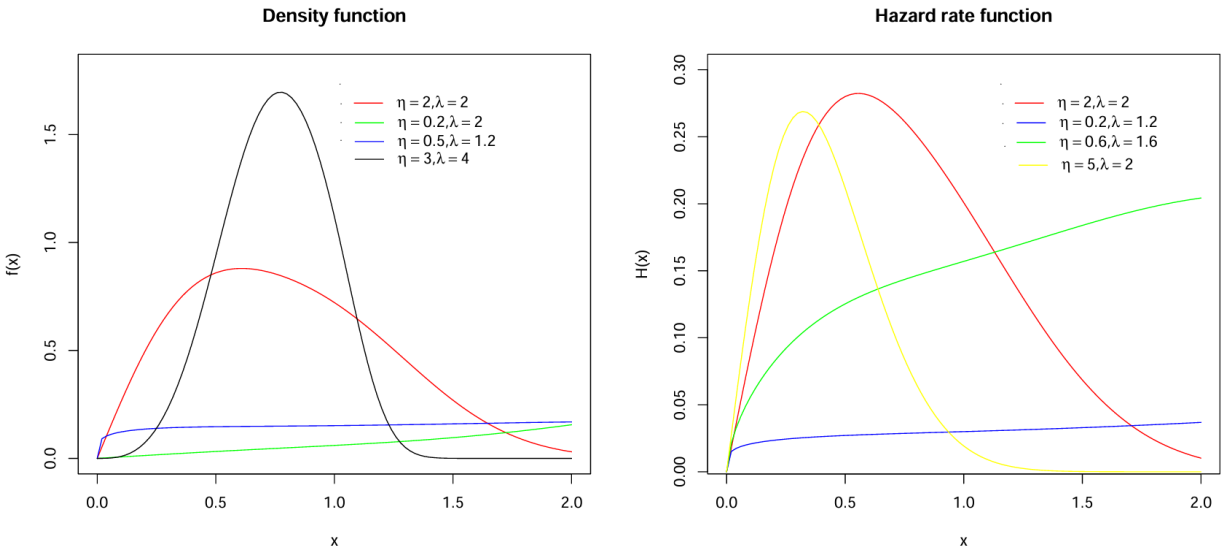

***Figure 8**: Xgamma Weibull distribution*

3.11 Half-logistic Xgamma distribution

Bantan et al. (2020) provided the flexible version of XGD by introducing one more parameter and called as Half-Logistic Xgamma distribution. To model the right skewed, uni-modal and J shaped data and CDF, HRF of its are:

$$G(x) = 2[1 - \mathcal{A}(x,\eta)]^{\lambda}(1 + [1 - \mathcal{A}(x,\eta)]^{\lambda})^{-1}; x > 0, \eta > 0, \lambda > 0 \tag{24}$$

and

$$H(x) = \frac{2\lambda\eta^2(1+\eta x^2/2)e^{-\eta x}[1-\mathcal{A}(x,\eta)]^{\lambda-1}/(1+\eta)(1+[1-\mathcal{A}(x,\eta)]^{\lambda})^2}{1-2[1-\mathcal{A}(x,\eta)]^{\lambda}(1+[1-\mathcal{A}(x,\eta)]^{\lambda})^{-1}} \tag{25}$$

where, $\mathcal{A}(x,\eta) = (1 + \eta + \eta x + \eta^2 x^2/2/(1 + \eta))e^{-\eta x}$.

HRF of Half-Logistic Xgamma distribution is able to capture the increasing, decreasing and semibathtub hazard rate scenarios, authors have shown these trends of HRF with the help of three different data sets of the manufacturing, medical field. Graphical presentation of the HRF of the same placed in Figure 9.

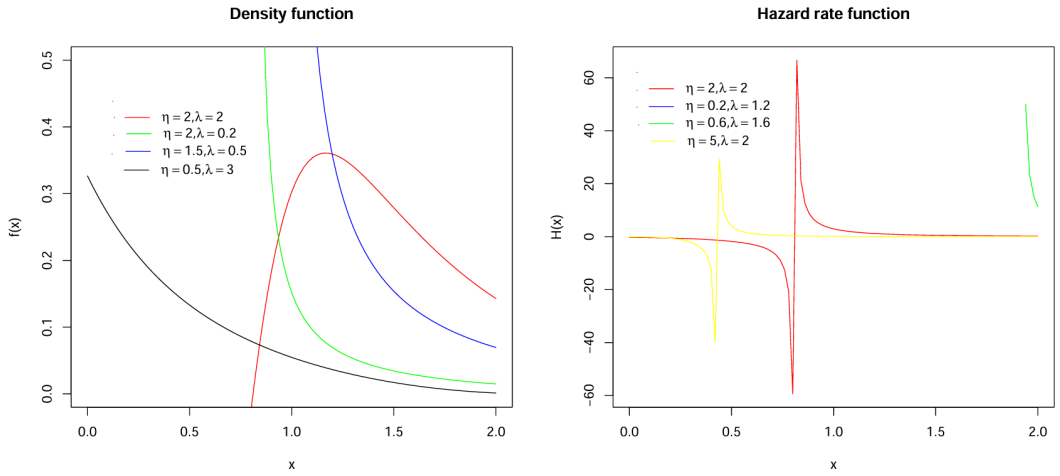

***Figure 9**: Half-logistic Xgamma distribution*

3.12 Lindley-Quasi Xgamma distribution

Hassan et al. (2020) published a paper introduced a new distribution, Lindley-Quasi Xgamma distribution. Authors dealt with properties, estimation and application of the Lindley-Quasi Xgamma distribution. CDF, HRF are:

$$G(x) = \frac{1}{2(\eta+\lambda)^2}[U_2], \quad x > 0,\ \eta > 0,\ \lambda > 0 \tag{26}$$

and

$$H(x) = \frac{2\eta e^{-\eta x}\{(\lambda+\eta)\left(\lambda+\frac{x^2\eta^2}{2}\right)+\eta(\eta-1)(1+\lambda x)\}}{1-\frac{1}{2(\eta+\lambda)^2}[U_2]} \tag{27}$$

Where, $U_2 = (\lambda + \eta)\{2\lambda + 2 - (2\lambda + x^2\eta^2 + 2\eta x + 2)e^{-\eta x}\} + 2(\eta - 1)\{\eta + \lambda - (\eta + \lambda\eta x + \lambda)e^{-\eta x}\}$ To justify the HRF, authors have used two data sets from the reccuring time of kidney infection and flood peaks of wheaten river. Graphical presentation of the HRF of the same placed in Figure 10.

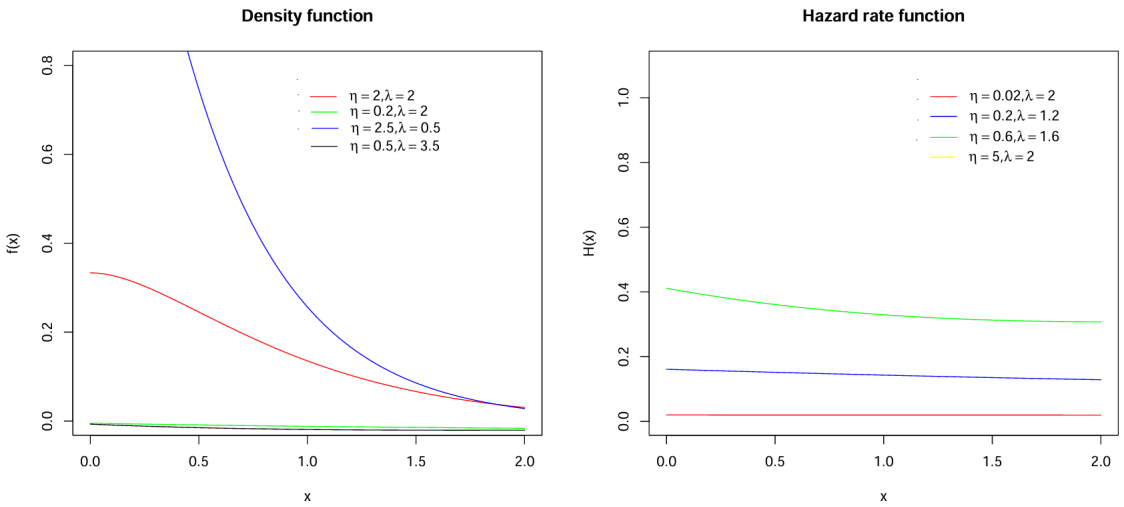

***Figure 10**: Lindley-Quasi Xgamma distribution*

3.13 Discrete Quasi Xgamma distribution

It is discretization of XGD and developed by Mazucheli et al. (2019). Authors proposed alternatives to model under and over dispersed datasets. Authors provided the probability mass function (PMF) when x is continuous random variable, following is the PMF for the support (−inf, +inf) and (0,inf) of x random variable.

$$PMF1 = \frac{f_X(y,\eta)}{\sum_{j=-\infty}^{\infty} f_X(j,\eta)}; y \in Z \tag{28}$$

and

$$PMF2 = \frac{f_X(y,\eta)}{\sum_{j=0}^{\infty} f_X(j,\eta)}; y \in Z + \tag{29}$$

where y is continuous random variable. Authors derived both type of PMF for different range and computed several statistical properties with applications of both. HRF suitability for real life scenario shown through two real life examples: Corn Borers data and outbreaks of strikes.

3.14 Poisson Xgamma distribution

Para et al. (2020) showed the importance of Poisson Xgamma distribution and it is a discrete model for count data. Authors showed that it beats so many known old and new discrete distributions. CDF and HRF are:

$$G(x) = \frac{\eta^2+(\eta x)^2/2+5\eta x^2/2+5\eta^2+x\eta+4\eta+1}{(1+\eta)^{x+4}}; x = 0,1,2,3\ldots, \eta > 0 \tag{30}$$

and

$$H(x) = \frac{\frac{\eta^2}{2(1+\eta)^{x+4}}[2(1+\eta)^2+\eta(x+1)(x+2)]}{1-\frac{\eta^2+(\eta x)^2/2+5\eta x^2/2+5\eta^2+x\eta+4\eta+1}{(1+\eta)^{x+4}}} \tag{31}$$

Two data sets: epileptic seizure counts and accidents to 647 women working on high explosive shells in 5 weeks are the representation of the matching the HRF of Poisson Xgamma distribution to the real life situation. Graphical presentation of the probability mass function of the same placed in Figure 11.

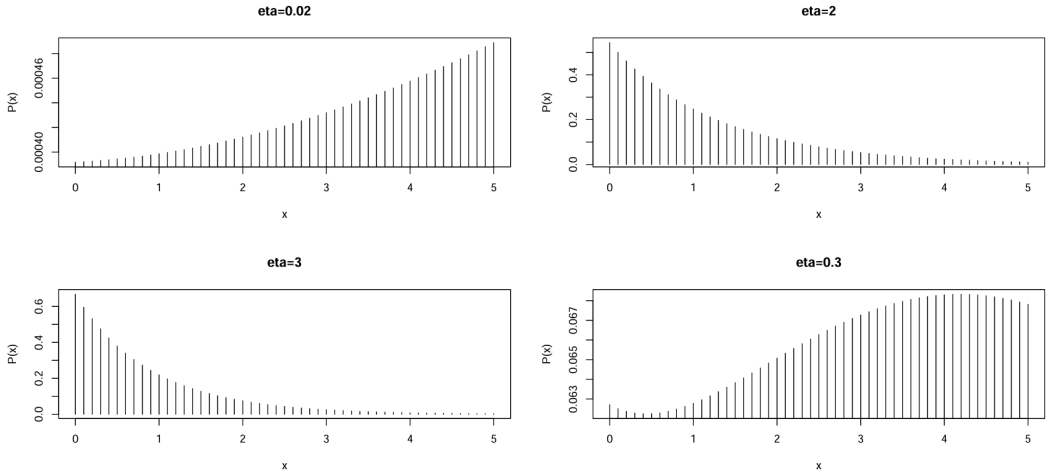

***Figure 11**: Poisson Xgamma distribution*

3.15 Exponentiated Xgamma distribution

Exponentiated version of Xgamma distribution is introduced by Yadav et al. (2021) and proved that this new version of XGD is very useful in some real situations. CDF and HRF are:

$$G(x) = \left[1 - \frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)} e^{-\eta x}\right]^{\lambda}; x > 0, \eta > 0, \lambda > 0 \tag{32}$$

and

$$H(x) = \frac{\frac{\lambda\eta^2}{1+\eta}\left[1-\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)\frac{e^{-\eta x}}{1+\eta}\right]^{(\lambda-1)}\left(1+\frac{\eta x^2}{2}\right)e^{-\eta x}}{1-\left[1-\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}\right]^{\lambda}} \tag{33}$$

For $\lambda \geq 1, \eta \geq 1$,it follows the pattern of increasing failure rate, decreasing failure rate when $\lambda < 1, \eta < 1$ and the pattern of bathtub shaped hazard rate may be traced for $\lambda < 1, \eta < 1$. Graphical presentation of the HRF of the same placed in Figure 12.

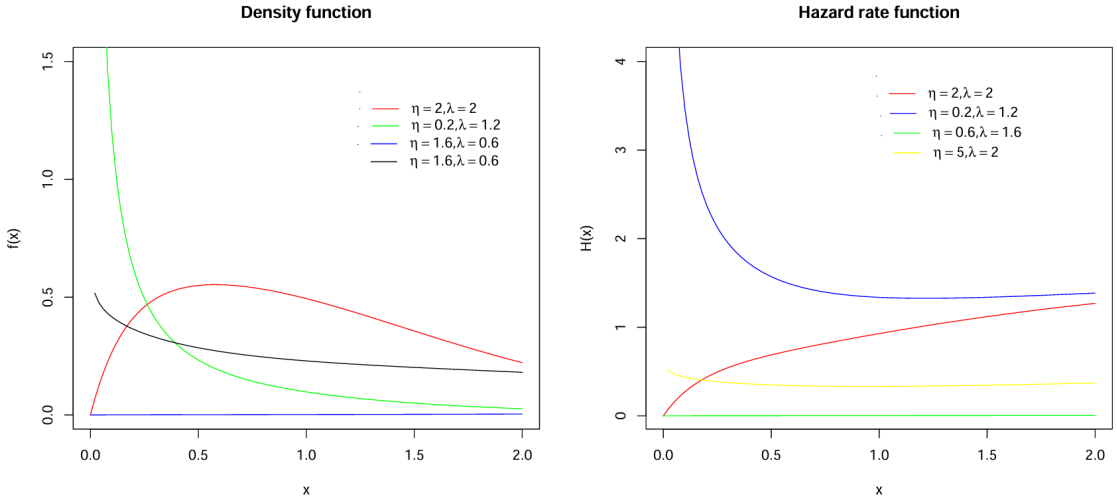

***Figure 12**: Exponentiated Xgamma distribution*

3.16 Weighted Lindley-Quasi Xgamma distribution

Wani and Shafi (2021) developed the Weighted Lindley-Quasi Xgamma distribution and applied the model to on two data sets. CDF and HRF are:

$$G(x) = \frac{(\eta+\lambda)2\eta(\gamma(c+1,\eta x)+\gamma(c+3,\eta x))+2(\eta-1)(\eta\gamma(c+1,\eta x)+\lambda\gamma(c+2,\eta x))}{c!(\eta+\lambda)(2\lambda+(c+1)(c+2))+2(\eta-1)((\eta+\lambda(c+1)))};$$
$$x > 0, \eta > 0, \lambda > 0 \tag{34}$$

and

$$H(x) = \left(\frac{\eta^{c+1}x^c\left((\lambda+\eta)(2\lambda+x^2\eta^2)+2\eta(\eta-1)(1+\lambda x)\right)e^{-\lambda x}}{U_6-U_7}\right)$$

Where, U_6 is $c!\ ((\eta+\lambda)(2\lambda+(c+1)(c+2))+2(\eta-1)(\eta+\lambda(c+1)))$ and U_7 is $((\lambda+\eta)(2\lambda\gamma(c+1,\eta x)+\gamma(c+3,\eta x))+2(\eta-1)(\eta\gamma(c+1,\eta x)+\lambda\gamma(c+2,\eta x)))$.

Hazard rate of it, revealed that model possesses non-decreasing hazard rate and it can be also seen that hazard rate becomes constant as value of x increases. Graphical presentation of the HRF of the same placed in Figure 13.

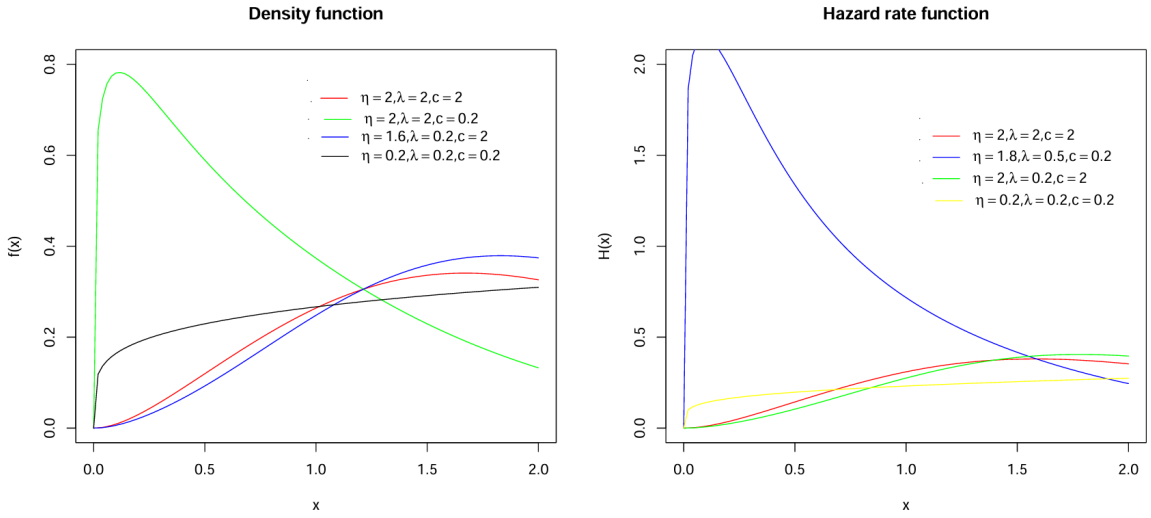

***Figure 13**: Weighted Lindley-Quasi Xgamma distribution*

3.17 Three Parameter Xgamma frechet distribution

Ibrahim et al. (2021) derived Three Parameter Xgamma frechet distribution and computed statistical properties of it. CDF and HRF are:

$$G(x) = \left[1 - \frac{1}{1+\eta}\left(1 - e^{-\left(\frac{a}{x}\right)^b}\right)^{\eta}\left(1 + \eta - \eta\log\left[1 - e^{-\left(\frac{a}{x}\right)^b}\right] + \frac{1}{2}\eta^2\left\{\log\left[1 - e^{-\left(\frac{a}{x}\right)^b}\right]\right\}^2\right)\right],$$
$$x > 0, \eta > 0, a > 0, b > 0. \tag{35}$$

and

$$H(x) = \frac{\frac{\eta}{1+\eta} b a^b x^{-(b+1)} e^{-\left(\frac{a}{x}\right)^b}\left[1-e^{-\left(\frac{a}{x}\right)^b}\right]^{\eta-1}\left(\eta+\frac{1}{2}\eta^2\left\{\log\left[1-e^{-\left(\frac{a}{x}\right)^b}\right]\right\}^2\right)}{1-\left[1-\frac{1}{1+\eta}\left(1-e^{-\left(\frac{a}{x}\right)^b}\right)^{\eta}\left(1+\eta-\eta\log\left[1-e^{-\left(\frac{a}{x}\right)^b}\right]+\frac{1}{2}\eta^2\left\{\log\left[1-e^{-\left(\frac{a}{x}\right)^b}\right]\right\}^2\right)\right]} \tag{36}$$

Authors used real data sets to prove the validity of developed model as literature full with new models. Considered data set for this model are from the manufacturing and medical field and they showed that HRF is monotonically increasing. Graphical presentation of the HRF of the same placed in Figure 14.

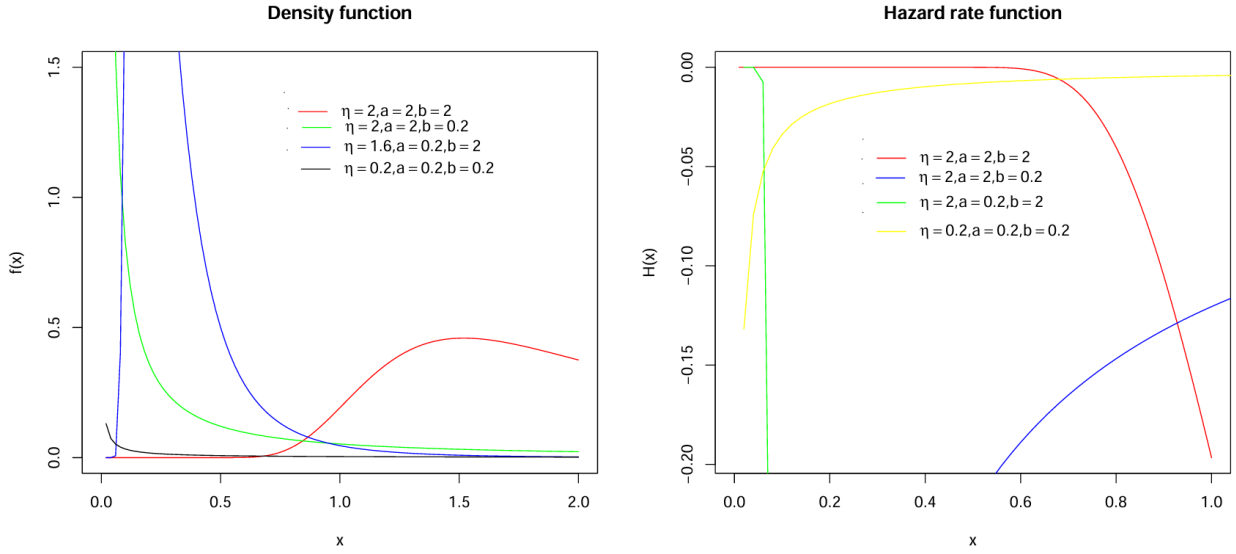

***Figure 14**: Three Parameter Xgamma Frechet distribution*

3.18 Alpha Power Transformed Xgamma distribution

Shukla et al. (2022) introduced two parameter XGD and made tis flexible version of XGD and called it as Alpha Power Transformed Xgamma distribution. CDF and HRF are:

$$G(x) = \begin{cases} \frac{\lambda^{1-u_0}-1}{\lambda-1}; \ \lambda > 0, \lambda \neq 1 \\ 1 - u_0; \ \alpha = 1 \end{cases} \tag{37}$$

Where, $u_0 = \frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)} e^{-\eta x}$

$$H_{APTXGD}(x) = \begin{cases} \frac{\frac{\log\lambda}{\lambda-1}\frac{\eta^2}{(1+\eta)} u_{01}\lambda^{1-u_0}}{1-\left[\frac{\lambda^{1-u_0}-1}{\lambda-1}\right]} \\ ;\lambda > 0, \lambda \neq 1 \\ \frac{\frac{\eta^2}{(1+\eta)}\left(1+\frac{\eta}{2}x^2\right)e^{-\eta x}}{1-[1-u_0]} ; \alpha = 1 \end{cases} \quad (38)$$

where, $u_{01} = \left(1+\frac{\eta}{2}x^2\right)e^{-\eta x}$. Authors found that HRF takes increasing and decreasing shape by using Glaser (1980) description about it. Authors justified the application of this by model four data sets of different field. Graphical presentation of the HRF of the same placed in Figure 15.

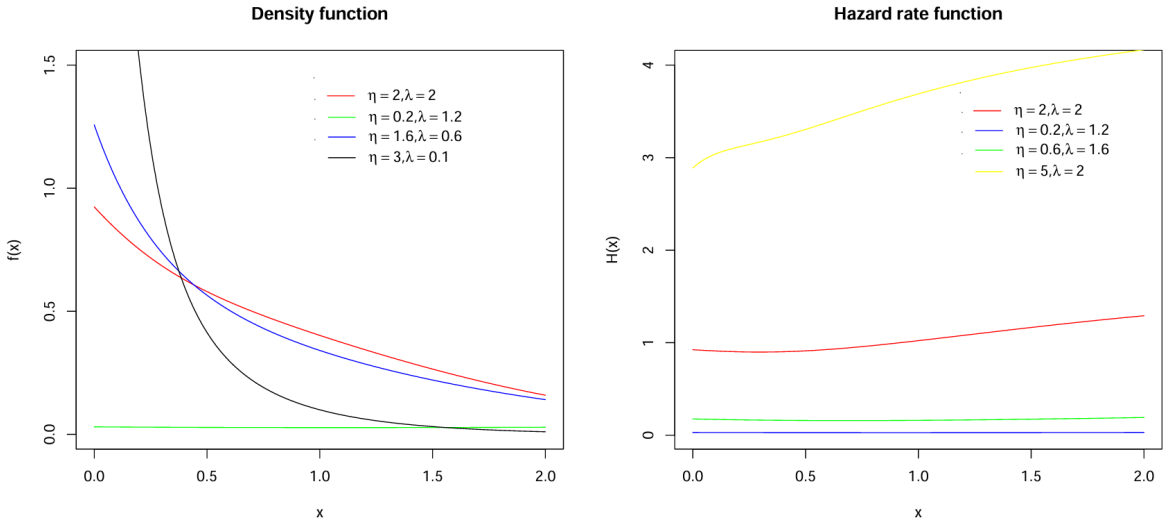

***Figure 15**: Alpha Power Transformed Xgamma distribution*

3.19 Flexible extension of Xgamma distribution

New extension of XGD based on Marshall-Olkin family of distribution is developed by Tripathi et al. (2022). They mentioned the different statistical properties of it. CDF and HRF are:

$$G(x) = 1 - \frac{\lambda\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}}{\left[1-\bar{\lambda}\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}\right]} \quad (39)$$

and

$$H(x) = \frac{\frac{\lambda\eta^2}{(1+\eta)}\frac{\left(1+\frac{\eta}{2}x^2\right)e^{-\eta x}}{\left[1-\bar{\lambda}\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}\right]^2}}{1-\frac{\lambda\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}}{\left[1-\bar{\lambda}\frac{\left(1+\eta+\eta x+\frac{\eta^2 x^2}{2}\right)}{(1+\eta)}e^{-\eta x}\right]}} \quad (40)$$

HRF of it can take every possible shape, i.e, increasing, decreasing and bathtub shape for the chosen value of the parameters. Authors used four real data sets to explore the areas of applications of it. Graphical presentation of the HRF of the same placed in Figure 16.

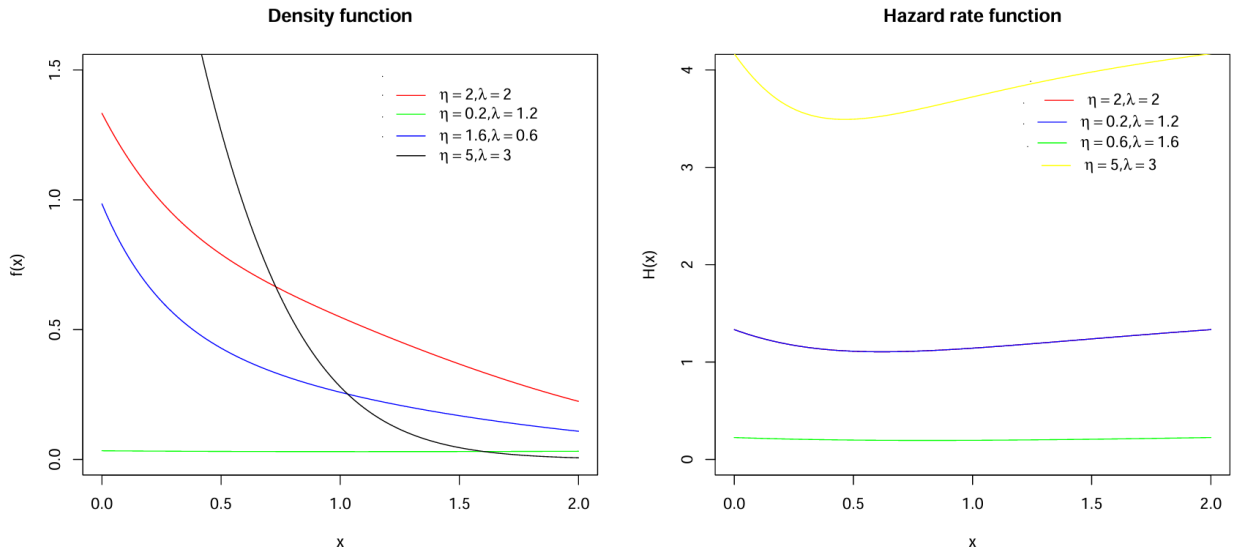

***Figure 16**: Flexible extension of Xgamma distribution*

3.20 Transmuted Inverse Xgamma distribution

Tripathi and Mishra (2022) developed the transmuted version of inverse xgamma distribution and they derived some of the properties for Transmuted Inverse Xgamma distribution. CDf and HRF are:

$$G(x) = (1+\lambda)\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x} - \lambda\left(\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x}\right)^2;$$
$$x>0, \eta>0, |\lambda| \le 1 \tag{41}$$

and

$$H(x) = \frac{(1+\lambda-2\lambda\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x})\frac{\eta^2}{1+\eta}\frac{1}{x^2}\left(1+\frac{\eta}{2x^2}\right)e^{-\eta x}}{1-(1+\lambda)\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x}+\lambda\left(\left[1+\frac{\eta}{\eta+1}\frac{1}{x}+\frac{\eta^2}{2(1+\eta)}\frac{1}{x^2}\right]e^{-\eta/x}\right)^2} \tag{42}$$

This model is applicable to the situation where researchers having keen interest to apply the proposed model for modeling of the data with increasing and deacreasing trend of HRF. Graphical presentation of the HRF of the same placed in Figure 17.

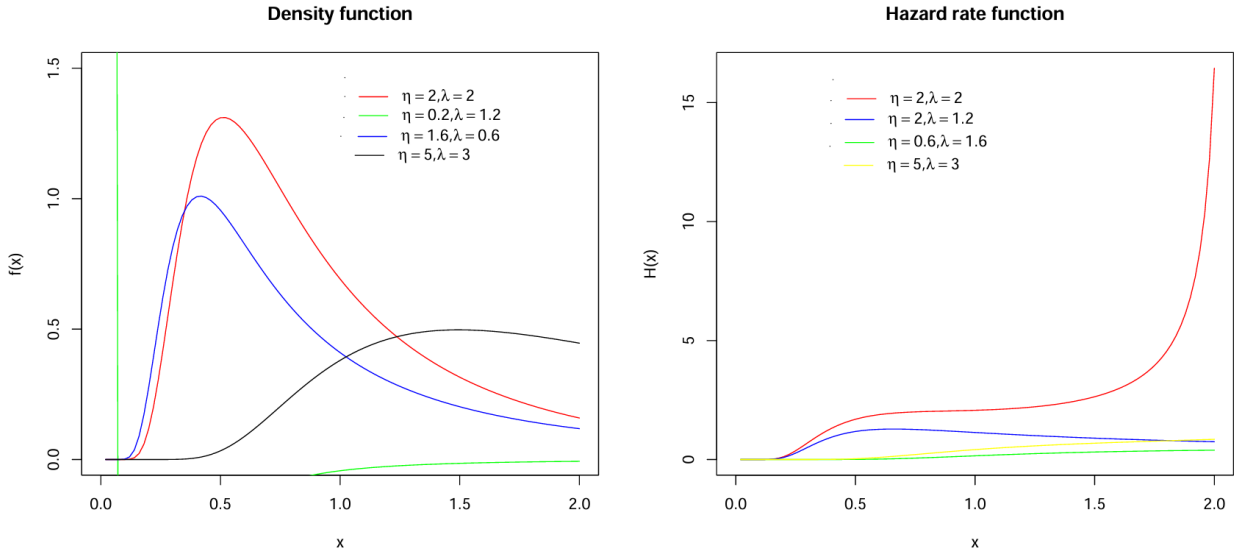

***Figure 17**: Transmuted Inverse Xgamma distribution*

3.21 Unit Xgamma distribution

Hashmi et al. (2022) presented a evolved version of XGD, called as unit Xgamma distribution. Unit Xgamma distribution is based on unit interval and it is good for modeling of such data sets. CDF and HRF are:

$$G(x) = 1 - \frac{1+\eta+\eta\left(\frac{x}{1-x}\right)+\frac{\eta^2}{2}\left(\frac{x}{1-x}\right)^2}{1+\eta}\, e^{-\eta\left(\frac{x}{1-x}\right)}, \quad 0<x<1, \eta>0 \tag{43}$$

and

$$H(x) = \frac{\eta^2\left\{1+\frac{x^2\eta}{2(1-x)^2}\right\}}{(1-x)^2\left[1+\eta+\frac{x\eta}{1-x}+\frac{x^2\eta^2}{2(1-x)^2}\right]} \tag{44}$$

It is suitable for negatively skewed data and analyze the data having increasing hazard rate function. Also to support the real life scenario which can be model by using Unit Xgamma distribution, authors used real data set is used for the validation of the same. Graphical presentation of the HRF of the same placed in Figure 18.

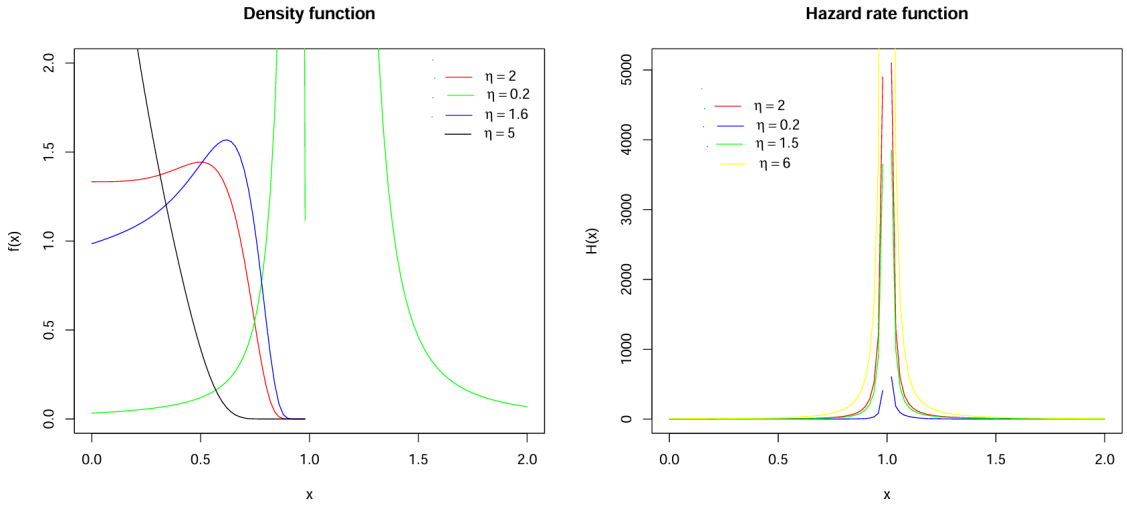

***Figure 18**: Unit Xgamma distribution*

3.22 Size Biased Lindley-Quasi Xgamma distribution

Wani et al. (2022) introduced Size Biased Lindley-Quasi Xgamma distribution and applied it to survival times data. CDF and HRF are:

$$G(x) = 1 - \left[1 + \frac{(2(\eta-1)(\eta+2\lambda+x\eta\lambda)+(\eta+\lambda)(2\lambda+6+3x\eta+x^2\eta^2))\eta x}{2(\eta^2+\lambda^2+3\eta\lambda+2\eta+\lambda)}\right]e^{-\eta x};$$
$$x > 0, \eta > 0, \lambda > 0 \tag{45}$$

and

$$H(x) = \frac{2\eta^2(\eta(\eta-1)(x+\lambda x^2)+(\eta+\lambda)(\lambda x+\frac{\eta^2x^3}{2}))e^{-\eta x}}{[U_4+(2(\eta-1)(\eta+2\lambda+x\eta\lambda)+(\eta+\lambda)(2\lambda+6+3x\eta+x^2\eta^2))\eta x]e^{-\eta x}} \tag{46}$$

where, U_4 is $2(\eta^2 + \lambda^2 + 3\lambda\eta + 2\eta + \lambda)$.

Developed model have non -decreasing hazard rate function and authors proved the necessity of this model over other models when HRF is decreasing in real life phenomena. Graphical presentation of the HRF of the same placed in Figure 19.

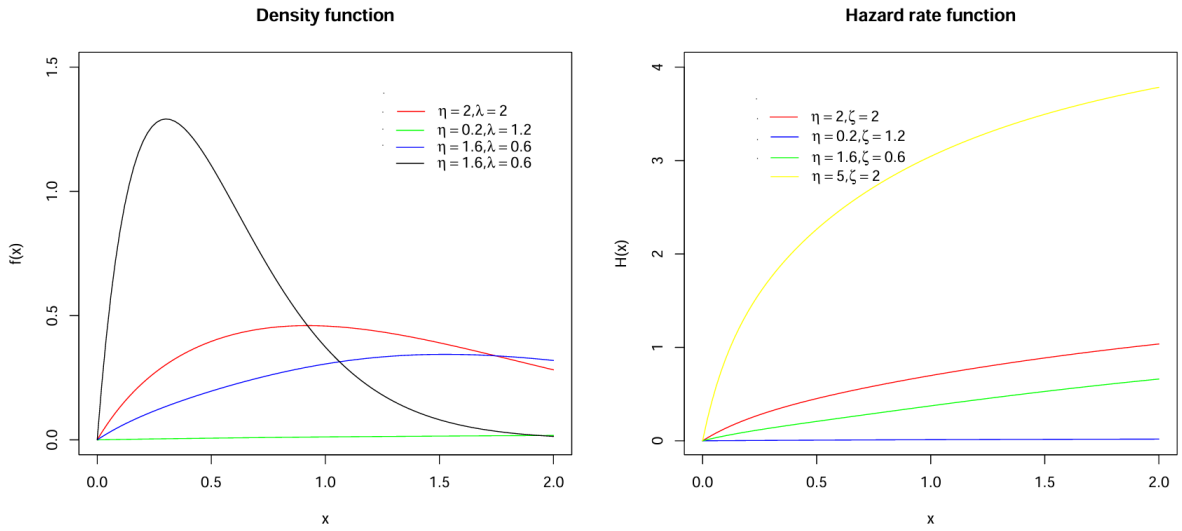

***Figure 19**: Size Biased Lindley Quasi Xgamma distribution*

3.23 Bivariate Xgamma distribution

Abulebda et al. (2022) developed Bivariate Xgamma distribution. Authors used FGM coupla to develop bivariate version of XGD. CDF and HRF are:

$$G(x_1,x_2)=U_5\left[1+\delta\left(\frac{\left(1+\eta_1+\eta_1x_1\frac{\eta_1^2x_1^2}{2}\right)}{(1+\eta_1)}e^{-\eta_1x_1}\right)\left(\frac{\left(1+\eta_2+\eta_2x_2\frac{\eta_2^2x_2^2}{2}\right)}{(1+\eta_2)}e^{-\eta_2x_2}\right)\right] \quad (47)$$

where, U_5 is $\left[1-\frac{\left(1+\eta_1+\eta_1x_1\frac{\eta_1^2x_1^2}{2}\right)}{(1+\eta_1)}e^{-\eta_1x_1}\right]\left[1-\frac{\left(1+\eta_2+\eta_2x_2\frac{\eta_2^2x_2^2}{2}\right)}{(1+\eta_2)}e^{-\eta_2x_2}\right]$.

HRFs of the Bivariate Xgamma distribution, η_1 and η_2 can be calculated by following equations:

$$\eta_1=\frac{-\partial}{\partial x_1}\log S(x_1,x_2)$$

and

$$\eta_2=\frac{-\partial}{\partial x_2}\log S(x_1,x_2)$$

where $S(x_1,x_2)$ is the survival function of Bivariate Xgamma distribution. HRF of Bivariate Xgamma distribution helps to researcher for modeling the increasing, decreasing and bathtub scenario. To demonstrate the use of the Bivariate Xgamma distribution, authors considered UEFA Champion's League data set and represents the time(in minutes) of the first kick goal scored by any team(X1) and time of first goal of any types scored by the home team(X2).

3.24 Xgamma exponential distribution

To Xgamma exponential distribution is introduced by Yadav et al. (2022) as an alternative to the one-parameter exponential distribution. The beauty of this distribution has been fully justified based on its monotone and bathtub-shaped hazard rate. CDF and HRF are given below:

$$G(x)=1-\frac{1}{2}e^{-\lambda x}\left[2+\lambda x+\frac{1}{2}(\lambda x)^2\right],\quad x>0,\lambda>0 \quad (48)$$

and,

$$H(x)=\frac{\lambda\left[1+\frac{1}{2}(\lambda x)^2\right]}{2+\lambda x+\frac{1}{2}(\lambda x)^2}. \quad (49)$$

Xgamma exponential distribution have the nature of increasing, decreasing and bathtub shape of HRF. Graphical presentation of the HRF of the same placed in Figure 20.

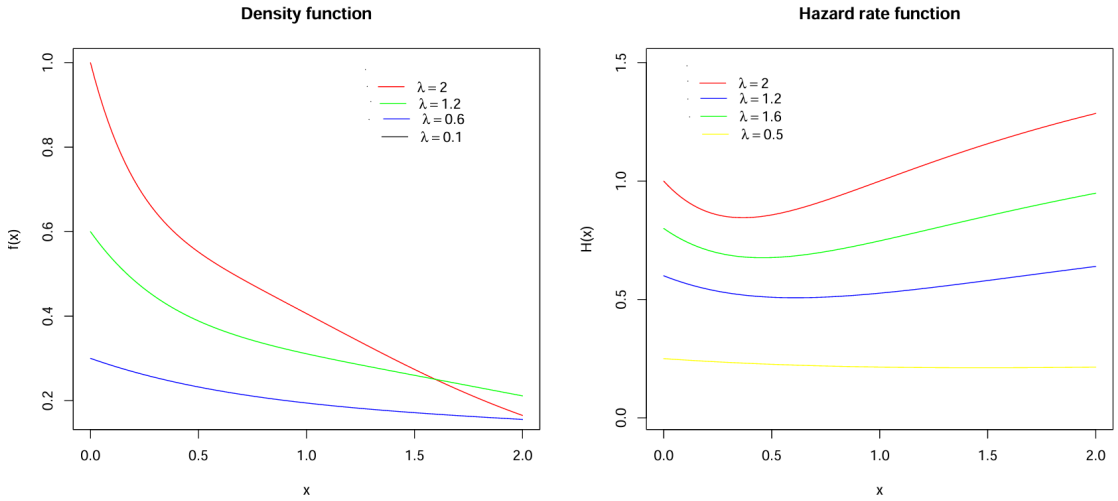

***Figure 20**: Xgamma Exponential distribution*

3.25 Weighted Xgamma exponential distribution

Yadav et al. (2023) introduced a weighted version of the Xgamma-exponential distribution following a similar methodological framework of weighted distributions. Their comprehensive study presented a thorough characterization of this new model along with diverse practical applications. The PDF and CDF of Weighted Xgamma exponential distribution are derived by employing a weight function $w(x)=x^{\eta}$ with the Xgamma exponential distribution as the baseline distribution. CDF and HRF are:

$$G(x)=\frac{2I_L(\eta+1,\lambda x)+I_L(\eta+3,\lambda x)}{\eta!(\eta^2+3\eta+4)},\quad x>0, \quad (50)$$

and

$$H(x) = \frac{\frac{2\lambda^{\eta+1}x^{\eta}e^{-\lambda x}\left[1+\frac{(\lambda x)^2}{2}\right]}{\eta!(\eta^2+3\eta+4)}}{1-\frac{2I_L(\eta+1,\lambda x)+I_L(\eta+3,\lambda x)}{\eta!(\eta^2+3\eta+4)}} \tag{51}$$

where I_L is the lower incomplete gamma function. This model is particularly noteworthy due to its capability to accommodate various hazard rate shapes including monotone as well as bathtub-shaped forms, making it highly flexible and applicable to a broad spectrum of real-world scenarios. Additionally, when η is set to 1 or 2, the model reduces to the length-biased and area-biased versions of Weighted Xgamma exponential distribution, respectively. Graphical presentation of the HRF of the same placed in Figure 21.

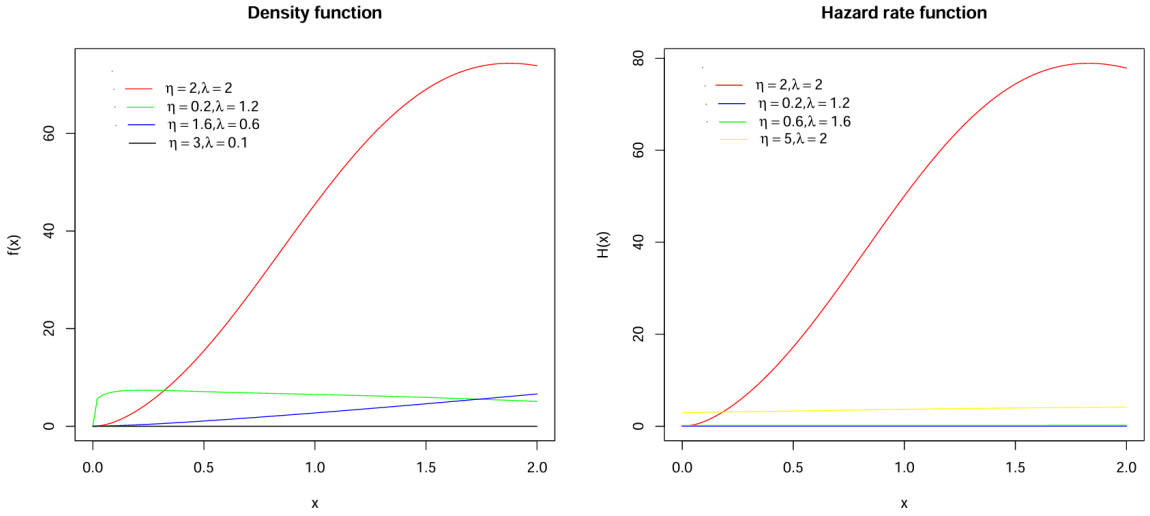

***Figure 21**: Weighted Xgamma Exponential distribution*

3.26 Truncated version of Xgamma distribution

Sen et al. (2024) proposed truncated version of Xgamma distribution and various estimation methods are discussed for the developed distribution. CDF is:

$$G(x) = \frac{F(x)-F(\alpha)}{F(\beta)-F(\alpha)}; \alpha \le x \le \beta \tag{52}$$

where F(.) denote the CDF of baseline distribution and HRF is:

$$H(x) = \frac{f(x)/(F(\beta)-F(\alpha))}{1-\frac{F(x)-F(\alpha)}{F(\beta)-F(\alpha)}} \tag{53}$$

Application of truncated version of Xgamma distribution is shown by strength data of glass of aircraft window and that is the proof of HRF of truncated version of Xgamma distribution matched with real life situation.

3.27 Improved Extension of Xgamma distribution

Improved extension of Xgamma distribution is introduced by Alomair et al. (2024) and it is called as Power Quasi Xgamma distribution. CDF and HRF are:

$$G(x) = 1 - \left[\frac{1+\eta+\lambda x^{\alpha}+(\lambda^2 x^{2\alpha}/2)}{1+\eta}\right] e^{-\lambda x^{\alpha}}; x > 0, \eta > 0, \lambda > 0, \alpha > 0 \tag{54}$$

and

$$H(x) = \frac{\frac{\lambda\alpha x^{\alpha-1}}{1+\eta}(\eta+(\lambda^2 x^{2\alpha}/2))e^{-\lambda x^{\alpha}}}{\frac{1+\eta+\lambda x^{\alpha}+(\lambda^2 x^{2\alpha}/2)}{1+\eta}e^{-\lambda x^{\alpha}}} \tag{55}$$

HRF of this can take every possible shape such as increasing, decreasing and bathtub which is essential for the real life application and authors used two data sets to show the application of the proposed distribution. Graphical presentation of the HRF of the same placed in Figure 22.

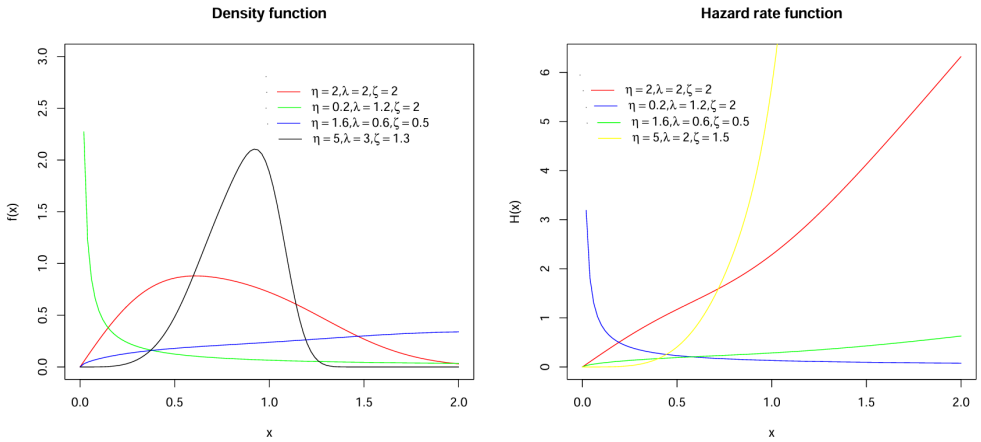

***Figure 22**: Power Quasi Xgamma distribution*

3.28 Generalized Inverse Xgamma distribution

Tripathi et al. (2025) derived the generalized form of inverse xgamma distribution which is having two parameter and more flexible than XGD. CDF and HRF are:

$$G(x) = \left(1 + \frac{\eta^2}{2(\eta+1)}\frac{1}{x^{(2\alpha)}} + \frac{\eta}{\eta+1}\frac{1}{x^{\alpha}}\right)e^{-\eta/x^{\alpha}}; \quad x > 0, \alpha > 0, \theta > 0, \tag{56}$$

and

$$H(x) = \frac{\frac{\alpha\eta^2}{1+\eta}\frac{1}{x^{(\alpha+1)}}\left(1+\frac{\eta}{2x^{2\alpha}}\right)e^{-\theta/x^{\alpha}}}{1-\left(1+\frac{\eta^2}{2(\eta+1)}\frac{1}{x^{(2\alpha)}}+\frac{\eta}{\eta+1}\frac{1}{x^{\alpha}}\right)e^{-\eta/x^{\alpha}}} \tag{57}$$

Generalized Inverse Xgamma distribution is positively skewed and uni-modal distribution, also HRF is increasing and reaches to a peak after that declined slowly, which indicates that the model possesses the hump or upside-down bathtub property of hazard rate. Graphical presentation of the HRF of the same placed in Figure 23.

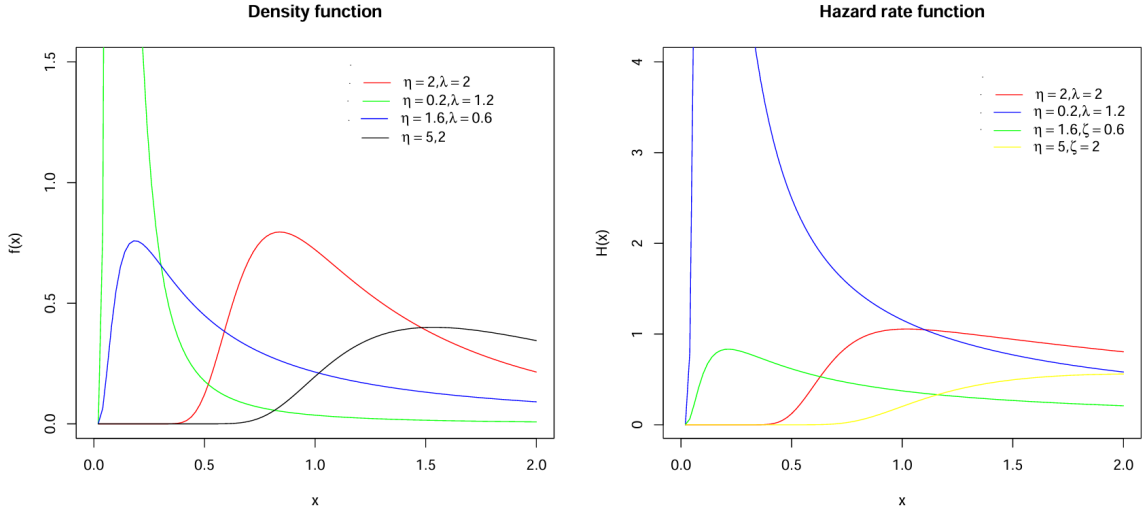

***Figure 23**: Generalized Inverse Xgamma distribution*

3.29 Power Xgamma distribution

If X follows the new power xgamma distribution [see, Boudjerda (2025)] which is generated by taking the transformation by $X=(1/Y)^{\beta}$ where Y follows one parameter XGD. CDF and HRF of Power Xgamma distribution this is given below.

$$G(x) = \frac{1+\eta+\eta x^{\lambda}+\frac{\eta^2}{2}x^{2\lambda}}{1+\eta}e^{-\eta x^{\lambda}}, \quad x > 0,\ \eta > 0,\ \lambda > 0 \tag{58}$$

and

$$H(x) = \frac{\frac{\eta^2\lambda}{1+\eta}x^{\lambda-1}\left(\frac{\eta^2}{2}x^{2\lambda}+1\right)e^{-\eta x^{\lambda}}}{\frac{1+\eta+\eta x^{\lambda}+\frac{\eta^2}{2}x^{2\lambda}}{1+\eta}e^{-\eta x^{\lambda}}} \tag{59}$$

This probability distribution modeled the situation of both monotonic and non-monotonic HRF which is common behavior in real life. Moreover Boudjerda (2025) claimed that in some situation it gives better result than XGD, gamma, Weibull, gamma-Lindley, Lomax and Lindley distributions. Graphical presentation of the HRF of the same placed in Figure 24.

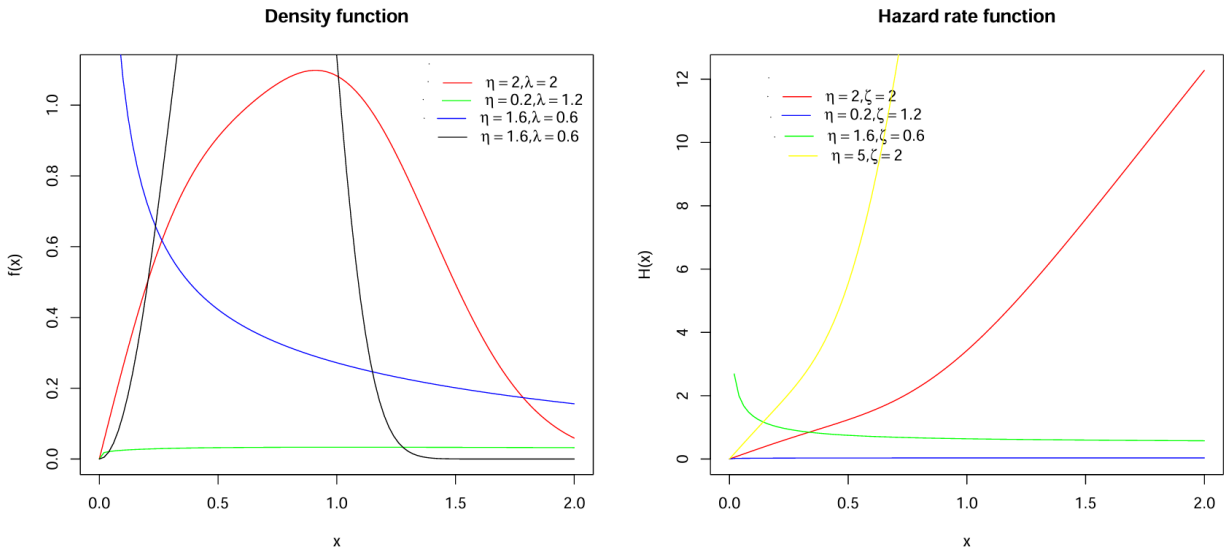

***Figure 24**: Power Xgamma distribution*

4. Application of Various Xgamma family of distributions

Real-life applications form a crucial component of any statistical study because they provide empirical validation for the proposed models and methodologies. In the context of the XGD and its various generalizations, real data analyses demonstrate not only the goodness of fit but also the practical interpretability of the underlying distributional assumptions in diverse applied settings. These applications help to bridge the gap between theoretical development and practical utility, thereby strengthening the credibility and relevance of the proposed models. In this study, several data sets that have been used in the literature to validate the XGD and its extensions are compiled and presented. For each data set, the observations together with the corresponding bibliographic reference are reported in this section, enabling readers to trace the original source and the specific modeling context in which XGD-type distributions have been employed. The selected data sets originate from a wide range of application areas, including medical studies (such as survival times or remission durations), engineering and reliability experiments (such as component or system lifetimes), and other industrial or environmental contexts where modeling of positive continuous data is essential.

An important feature of the considered data sets is their ability to exhibit different types of hazard rate behavior, which is one of the main motivations for employing flexible lifetime distributions like XGD and its generalizations. Collectively, these data sets display increasing, decreasing, bathtub-shaped, inverted-bathtub, and other non-monotone hazard rate patterns, thereby providing a rich testbed for assessing the performance of competing models. By showing that XGD-based distributions can successfully accommodate such diverse hazard structures, the applications underscore the flexibility and robustness of this family of models for practical reliability and survival analysis.

Data 1: Following data used by XGD, Sen et al. (2016) and it shows the application of XGD version in virulent tubercle bacilli. 72 guinea pigs infected with virulent tubercle bacilli and data is:

10, 33, 44, 56, 59, 72, 74, 77, 92, 93, 96, 100, 100, 102, 105, 107, 107, 108, 108, 108, 109, 112, 113, 115, 116, 120, 121, 122, 122, 124, 130, 134, 136, 139, 144, 146, 153, 159, 160, 163, 163, 168, 171, 172, 176, 183, 195, 196, 197, 202, 213, 215, 216, 222, 230, 231, 240, 245, 251, 253, 254, 254, 278, 293, 327, 342, 347, 361, 402, 432, 458, 555.

Data 2: Data values represents the Fatigue lifetime of 23 deep-groove ball bearings For detail of data, see, Weighted Xgamma distribution, Sen et al. (2017).

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.48, 51.84, 51.96, 54.12, 55.56, 67.80, 68.64, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40

Data 3: Monitoring and modelling the cancer decease data is big task in medical industry. Sen and Chandra (2017) used random sample of 128 bladder cancer patients to show the modeling of cancer decease datawith the help of Quasi Xgamma distribution.

0.08, 2.09, 3.48, 4.87, 6.94, 8.66, 13.11, 23.63, 0.20, 2.23, 0.26, 0.31, 0.73,0.52,4.98,6.97,9.02, 13.29, 0.40, 2.26, 3.57, 5.06, 7.09, 11.98, 4.51, 2.07,0.22,13.8,25.74,0.50,2.46,3.64,5.09,7.26,9.47, 14.24, 19.13, 6.54, 3.36,0.82,0.51,2.54,3.70,5.17,7.28,9.74,14.76,26.31,0.81,1.76,8.53,6.93, 0.62, 3.82, 5.32, 7.32, 10.06, 14.77, 32.15, 2.64, 3.88, 5.32, 3.25, 12.03,8.65,0.39,10.34,14.83,34.26, 0.90,2.69, 4.18, 5.34, 7.59, 10.66, 4.50,20.28,12.63,0.96,36.66,1.05,2.69,4.23,5.41,7.62,10.75, 16.62,43.01, 6.25, 2.02, 22.69, 0.19, 2.75, 4.26, 5.41, 7.63, 17.12, 46.12, 1.26, 2.83, 4.33, 8.37, 3.36, 5.49, 0.66, 11.25, 17.14, 79.05, 1.35, 2.87, 5.62, 7.87, 11.64,17.36,12.02,6.76,0.40, 3.02,4.34,5.71, 7.93, 11.79, 18.1, 1.46, 4.40, 5.85, 2.02, 12.07

Data 4: Stress-rupture life of kevlar 49/epoxy strands are recorded in terms of given observations and it is used by Sen et al. (2018) to discussed the real life application of Quasi Xgamma Poisson distribution.

0.01, 0.01, 0.02, 0.02, 0.02, 0.03, 0.03, 0.04, 0.05, 0.06, 0.07, 0.07, 0.08, 0.09, 0.09, 0.10, 0.10, 0.11, 0.11, 0.12, 0.13, 0.18, 0.19, 0.20, 0.23, 0.24, 0.24, 0.29, 0.34, 0.35, 0.36, 0.38, 0.40, 0.42, 0.43, 0.52, 0.54, 0.56, 0.60, 0.60, 0.63, 0.65, 0.67, 0.68, 0.72, 0.72, 0.72, 0.73, 0.79, 0.79, 0.80, 0.80, 0.83, 0.85, 0.90, 0.92, 0.95, 0.99, 1.00, 1.01, 1.02, 1.03, 1.05, 1.10, 1.10, 1.11, 1.15, 1.18, 1.20, 1.29, 1.31, 1.33, 1.34, 1.40, 1.43, 1.45, 1.50, 1.51, 1.52, 1.53, 1.54, 1.54, 1.55, 1.58, 1.60, 1.63, 1.64, 1.80, 1.80, 1.81, 2.02, 2.05, 2.14, 2.17, 2.33, 3.03, 3.03, 3.34, 4.20, 4.69, 7.89

Data 5: Altunand and Hamedani (2018) dealt with the comparison of two different algorithms called SC16 and P3 for estimating unit capacity factors to implement the application of Log Xgamma distribution.

0.853,0.759,0.866,0.809,0.717,0.544,0.492,0.403,0.344,0.213,0.116,0.116, 0.092,0.070,0.059, 0.048,0.036,0.029,0.021,0.014,0.011,0.008, 0.006

Data 6: Sen et al. (2019) proved the importance of Quasi Xgamma-Geometric distribution by using remission times(in months) of bladder cancer patients and the observations of remission times are given below:

0.08, 2.09, 3.48, 4.87, 6.94, 8.66, 13.11, 23.63, 0.20, 2.23, 0.26, 0.31, 0.73,0.52,4.98,6.97,9.02, 13.29, 0.40, 2.26, 3.57, 5.06, 7.09, 11.98, 4.51, 2.07,0.22,13.8,25.74,0.50,2.46,3.64,5.09,7.26,9.47, 14.24, 19.13, 6.54, 3.36,0.82,0.51,2.54,3.70,5.17,7.28,9.74,14.76,26.31,0.81,1.76,8.53,6.93, 0.62, 3.82, 5.32, 7.32, 10.06, 14.77, 32.15, 2.64, 3.88, 5.32, 3.25, 12.03,8.65,0.39,10.34,14.83,34.26, 0.90,2.69, 4.18, 5.34, 7.59, 10.66, 4.50,20.28,12.63,0.96,36.66,1.05,2.69,4.23,5.41,7.62,10.75, 16.62,43.01, 6.25, 2.02, 22.69, 0.19, 2.75, 4.26, 5.41, 7.63, 17.12, 46.12, 1.26, 2.83, 4.33, 8.37, 3.36, 5.49, 0.66, 11.25, 17.14, 79.05, 1.35, 2.87, 5.62, 7.87, 11.64,17.36,12.02,6.76,0.40, 3.02,4.34,5.71, 7.93, 11.79, 18.1, 1.46, 4.40, 5.85, 2.02, 12.07, 21.73, 2.07, 3.36,6.93,8.65,12.63,22.69

Data 7: Given observations are the survival time (in weeks) of patient suffering from acute Myelogeneous and this positive skewed data handeled by Sen et al. (2019) to illustrate the importance of Quasi Xgamma-Geometric distribution.

65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65, 56, 65, 17, 7, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43

Data 8: To show the superiority of Xgamma Weibull distribution over other exiting version of XGD, Yousof et al. (2020) used a data set and data represents the strength of 1.5 cm glass fibres, measured at National physical laboratory, England.

0.55, 0.93, 1.25, 1.36, 1.49, 1.52, 1.58, 1.61, 1.64, 1.68, 1.73, 1.81, 2.00, 0.74, 1.04, 1.27, 1.39, 1.49, 1.53, 1.59, 1.61, 1.66, 1.68, 1.76, 1.82, 2.01, 0.77, 1.11, 1.28, 1.42, 1.50, 1.54, 1.60, 1.62, 1.66, 1.69, 1.76, 1.84, 2.24, 0.81, 1.13, 1.29, 1.48, 1.50, 1.55, 1.61, 1.62, 1.66, 1.70, 1.77, 1.84, 0.84, 1.24, 1.30, 1.48, 1.51, 1.55, 1.61, 1.63, 1.67, 1.70, 1.78, 1.89.

Data 9: Bantan et al. (2020) showed the application of Half-logistic Xgamma distribution to the manufacturing industry and data represents 100 observations of breaking stress of carbon fibers (in Gba).

3.7, 3.11, 4.42, 3.28, 3.75, 2.96, 3.39, 3.31, 3.15, 2.81, 1.41, 2.76, 3.19, 1.59, 2.17, 3.51, 1.84, 1.61, 1.57, 1.89, 2.74, 3.27, 2.41, 3.09, 2.43, 2.53, 2.81, 3.31, 2.35, 2.77, 2.68, 4.91, 1.57, .65, 2.12, 1.8, 0.85, 4.38, 2, 1.18, 1.71, 1.17, 2.17, 0.39, 2.79, 1.08, 2.73, 0.98, 1.73, 1.59, 1.92, 2.38, 3.56, 2.55, 3.22, 3.39, 4.9, 1.69, 3.11, 3.6, 2.05, 1.61, 2.03, 2.48, 1.25, 2.48, 1.12, 2.88, 2.87, 3.19, 1.87, 2.95, 2.67, 4.2, 2.85, 2.55, 2.17, 2.97, 3.68, 0.81, 1.22, 5.08, 1.69, 3.68, 4.70, 2.03, 2.82, 2.50, 3.22, 3.15, 1.47, 5.56, 1.84, 1.36, 2.59, 2.83, 2.56, 3.33, 2.93, and 2.97.

Data 10: Following data set having 63 observation and represents gauge length of 10mm and this data sets is used by Bantan et al. (2020) for application purpose.

1.901, 2.132, 2.203, 2.228, 2.257, 2.350, 2.361, 2.396, 2.397, 2.445, 2.454, 2.474, 2.518, 2.522, 2.525, 2.532, 2.575, 2.614, 2.616, 2.618, 2.624, 2.659, 2.675, 2.738, 2.740, 2.856, 2.917, 2.928, 2.937, 2.937, 2.977, 2.996, 3.030, 3.125, 3.139, 3.145, 3.220, 3.223, 3.235, 3.243, 3.264, 3.272, 3.294, 3.332, 3.346, 3.377, 3.408, 3.435, 3.493, 3.501, 3.537, 3.554, 3.562, 3.628, 3.852, 3.871, 3.886, 3.971, 4.024, 4.027, 4.225, 4.395, and 5.020.

Data 11: Death time (in weeks) of patient cancer of the tongue with an aneuploid DNA profile modeled by using Half-logistic Xgamma distribution [see, Bantan et al. (2020)].

1, 3, 3, 4, 10, 13, 13, 16, 16, 24, 26, 27, 28, 30, 30, 32, 41, 51, 61, 65, 67, 70, 72, 73, 74, 77, 79, 80, 81, 87, 87, 88, 89, 91, 93, 93, 96, 97, 100, 101, 104, 104, 108, 109, 120, 131, 150, 157, 167, 231, 240, and 400.

Data 12: Following data is having 58 recurrence times to infection, at the point of insertion of the catheter and this data is used by Hassan et al. (2020) [see, Lindley-Quasi Xgamma distribution].

8, 16, 23, 22, 28, 447, 318, 30, 12, 24, 245, 7, 9, 511, 30, 53, 196, 15, 154, 7, 333, 141, 96, 38, 536, 17, 185, 177, 292, 114, 15, 152, 562, 402, 13, 66, 39, 12, 40, 201, 132, 156, 34, 30, 2, 25, 130, 26, 27, 58, 43, 152, 30, 190, 119, 8, 78, 63

Data 13: 72 exceedances of flood peaks (in m3/s) of the Wheaton River for the year 1958-1984 are reported in following data set and it is used by Hassan et al. (2020) to justify the development of Lindley-Quasi Xgamma distribution.

1.7, 2.2, 14.4, 1.1, 0.4, 20.6, 5.3, 0.7, 13.0, 12.0, 9.3, 1.4, 18.7, 8.5, 25.5, 11.6, 14.1, 22.1, 1.1, 2.5, 14.4, 1.7, 37.6, 0.6, 2.2, 39.0, 0.3, 15.0, 11.0, 7.3, 22.9, 1.7, 0.1, 1.1, 0.6, 9.0, 1.7, 7.0, 20.1, 0.4, 14.1, 9.9, 10.4, 10.7, 30.0, 3.6, 5.6, 30.8, 13.3, 4.2, 25.5, 3.4, 11.9, 21.5, 27.6, 36.4, 2.7, 64.0, 1.5, 2.5, 27.4, 1.0, 27.1, 20.2, 16.8, 5.3, 9.7, 27.5, 2.5, 27.0, 1.9, 2.8

Data 14: Exponentiated Xgamma distribution developed by Yadav et al. (2021) and demonstrated the application by using thirty successive values of March precipitation in Minneapolis/St Paul.

0.77, 1.74, 0.81, 1.2, 1.95, 1.2, 0.47, 1.43, 3.37, 2.2, 3, 3.09, 1.51, 2.1, 0.52, 1.62, 1.31, 0.32, 0.59, 0.81, 2.81, 1.87, 1.18, 1.35, 4.75, 2.48, 0.96, 1.89, 0.9, 2.05.

Data 15: Yadav et al. (2021) introduced new version of XGD, Exponentiated Xgamma distribution and used this model for modelling of new cases of Covid-19 in Italy during 31st May 2020 to 30th June 2020.

334, 200, 319, 322, 177, 519, 270, 197, 280, 283, 202, 380, 163, 347, 337, 301, 210, 329, 332, 251, 264, 224, 221, 113, 190, 296, 255, 175, 174, 126, 142.

Data 16: For application purpose of Weighted Lindley-Quassi Xgamma distribution, Wani and Shafi (2021) utilized the data set of tensile strength measures in GPa of 69 carbon fibres tested under tension at gauge lengths of 20mm.

1.312,1.314,1.479,1.552,1.700,1.803,1.861,1.865,1.944,1.958,1.966,1.997,2.006,2.021,2.027,
2.055,2.063,2.098,2.140,2.179,2.224,2.240,2.253,2.270,2.272,2.274,2.301,2.301,2.359,2.382,
2.382,2.426,2.434,2.435,2.478,2.490,2.511,2.514,2.535,2.554,2.566,2.570,2.586,2.629,2.633,
2.642,2.648,2.684,2.697,2.726,2.770,2.773,2.800,2.809,2.818,2.821,2.848,2.880,2.954,3.012,
3.067,3.084,3.090,3.096,3.128, 3.233, 3.433, 3.585, 3.585

Data 17: Alpha Power transformed Xgamma distribution [see, Shukla et al. (2022)] is very useful to deal with the data which is observed under accelerated life test, following data is the sample failure time (in minutes) of 15 electronic components.

1.4,5.1,6.3,10.8,12.1,18.5,19.7,22.2,23,30.6, 37.3,46.3,53.9,59.8,66.2"

Data 18: Following data is generated from clean up monitoring wells by using vinyl chloride which is a volatile organic compound. Shukla et al. (2022) applied developed Alpha power transformed Xgamma distribution to model the following data.

5.1,1.2,1.3,0.6,0.5,2.4,0.5,1.1,8.0,0.8,0.4,0.6,0.9,0.4,2.0,0.5,5.3,3.2,2.7,2.9,2.5,2.3,1.0,
0.2,0.1,0.1,1.8,0.9,2.0,4.0,6.8,1.2,0.4,0.2

Data 19: In recent time modelling of corona-virus cases distribution is the key interest of researchers and in same fashion, Alpha Power transformed Xgamma distribution [see, Shukla et al. (2022)] used to model the corona-virus cases distribution among the fifteen countries viz.,France, Italy, Spain, US, Germony, UK, Turkey, Iran, Russia, China, Brazil, Canada, Belgium, Netherlands and Switzerland.

5.37,6.56,7.61,32.83,5.24,5.06,3.65,3.03 2.89,2.74,2.10,1.57,1.55,1.27,0.97

Data 20: Following observation represents the water capacity month-wise from the Shasta reservoir in California in the month of February from 1991 to 2010. Sharqa et al. (2022) modeled the water capicity with the help of Unit Xgamma distribution.

0.338936, 0.431915, 0.759932, 0.724626, 0.757583, 0.811556, 0.785339, 0.783660, 0.815627, 0.847413, 0.768007, 0.843485, 0.787408, 0.849868, 0.695970, 0.842316, 0.828689, 0.580194, 0.430681 and 0.742563.

Data 21: Survival times (in months) of 46 patients of melanoma (non-censored data) is reported in following data and Wani et al. (2022) implemented Size Biased Lindley-Quasi Xgamma distribution to mentioned data.

3.25, 3.50, 4.75, 4.75, 5.00, 5.25, 5.75, 5.75, 6.25, 6.50, 6.50, 6.75, 6.75, 7.78 ,8.00, 8.50, 8.50, 9.25, 9.50, 9.50, 10.00, 11.50, 12.5, 13.25, 13.5, 14.25, 14.50, 14.75, 15.00, 16.25, 16.25, 16.50, 17.5, 21.75, 22.50, 24.50, 25.50, 25.75, 27.50, 29.50, 31.00, 32.50, 34.00, 34.50, 35.25, 58.50

Data 22: Truncated data of strength of glass of aircraft window is given below and to tackle this truncated data, Truncated version of Xgamma distribution is developed by Sen et al. (2024).

18.83, 20.80, 21.657, 23.03, 23.23, 24.05, 24.321, 25.5, 25.52, 25.80, 26.69, 26.770, 26.78, 27.05, 27.67, 29.90, 31.11, 33.20, 33.73, 33.76, 33.890, 34.76, 35.75, 35.91, 36.98, 37.08, 37.09, 39.58, 44.045, 45.29, 45.381

Data 23: Generalized Nnverse Xgamma distribution introduced by Tripathi et al. (2025) and placed the real life application section by using the data of survival times (in days)guinea pigs with different doses of tubercle bacilli, and this data is reported below:

12, 15, 22, 24, 24, 32, 32, 33, 34, 38, 38, 43, 44, 48, 52, 53, 54, 54, 55, 56, 57, 58, 58, 59, 60, 60, 60, 60, 61, 62, 63, 65, 65, 67, 68, 70, 70, 72, 73, 75, 76, 76, 81, 83, 84, 85, 87, 91, 95, 96, 98, 99, 109, 110, 121, 127, 129, 131, 143, 146, 146, 175, 175, 211, 233, 258, 258, 263, 297, 341, 341, 37

Data 24: To prove the applicability of Generalized Inverse Xgamma distribution [see, Tripathi et al. (2025)] utilized the data of 44 patients suffering from head and neck cancer disease and were treated using combined radio therapy and chemotherapy, and this data is placed below:

12.20, 23.56, 23.74, 25.87, 31.98, 37, 41.35, 47.38, 55.46, 58.36, 63.47, 68.46, 78.26, 74.47, 81.43, 84, 92, 94, 110, 112, 119, 127, 130, 133, 140, 146, 155, 159, 173, 179, 194, 195, 209, 249, 281, 319, 339, 432, 469, 519, 633, 725, 817, 1776

Data 25: Newly version of Power Xgamma distribution derived by Boudjerda (2025) and it is used to model the repair times for an airborne communication transceiver.

0.50, 0.60, 0.60, 0.70, 0.70, 0.70, 0.80, 0.80,1.00, 1.00, 1.00, 1.00, 1.10, 1.30, 1.50, 1.50, 1.50, 1.50, 2.00,2.00, 2.20, 2.50, 2.70, 3.00, 3.00, 3.30, 4.00, 4.00, 4.50, 4.70,5.00, 5.40, 5.40, 7.00, 7.50, 8.80, 9.00, 10.20, 22.00, 24.50

Data 26: Boudjerda (2025) formulated the new Power Xgamma distribution and incorporated this newly developed distribution to develop the times of failure of fatigue fracture of Kevlar 373/epoxy.

1.2985, 1.3211, 1.3503, 1.3551, 1.4595, 1.4880, 1.5728, 1.5733, 1.7083, 1.7263, 1.7460, 1.7630, 1.7746, 1.8275, 0.0251, 0.0886, 0.0891, 0.2501, 0.3113, 0.3451, 0.4763, 0.5650, 0.5671, 0.6566, 0.6748, 0.6751, 0.6753, 0.7696, 0.8375, 0.8391, 0.8425, 0.8645, 0.8851, 0.9113, 0.9120, 0.9836, 1.0483, 1.0596, 1.0773, 1.1733, 1.2570, 2.2100, 3.7455, 3.9143, 4.8073, 5.4005, 5.4435, 5.5295, 6.5541, 9.0960, 2.2460, 2.2878, 2.3203, 2.3470, 2.3513, 2.4951, 2.5260, 2.9911, 3.0256, 3.2678, 3.4045, 3.4846, 3.7433, 1.2766, 1.8375, 1.8503, 1.8808, 1.8878, 1.8881, 1.9316, 1.9558, 2.0048, 2.0408, 2.0903, 2.1093, 2.1330.

Fitting of all data sets for XGD are reported in Table 2 and results of this table indicate that authors developed better versions of XGD where they claimed, new versions of XGD are better than the XGD. Comparison between the new versions of XGD and XGD are done based on different measures such as AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion), KS value and p-value.

***Table 2**: Model fitting summary of data sets for XGD*

Data Set	Estimate	LL	AIC	BIC	KS value	p-value
1	0.01678	-425.5916	853.1832	855.4599	0.10096	0.45530
2	0.04071	-113.9656	229.9312	231.0667	0.13232	0.81550
3	0.28606	-425.1691	852.3382	855.1902	0.18487	0.00031
4	1.69785	-104.1007	210.2015	212.8166	0.08271	0.49420
5	4.75336	-6.371815	10.74363	9.608136	0.21181	0.25350
6	0.28259	-449.0842	900.1684	903.0736	0.17829	0.00037
7	0.06475	-176.3484	354.6967	356.1932	0.35452	0.00049
8	1.33761	-85.80016	173.6003	175.7435	0.41961	4.634e-10
9	0.86076	-183.4777	368.9554	371.5605	0.28607	1.559e-07
10	0.76304	-122.2032	246.4064	248.5495	0.45868	6.143e-12
11	0.03533	-286.5651	575.1302	577.0814	0.18236	0.06294
12	0.02397	-377.3694	756.7389	758.7993	0.40089	1.602e-08
13	0.20448	-266.4647	534.9293	537.206	0.25620	0.00015
14	1.18998	-44.57353	91.14705	92.54825	0.22263	0.10220
15	0.01167	-187.9499	377.8997	379.3337	0.20905	0.11480
16	0.91375	-121.5597	245.1194	247.3535	0.43273	1.198e-11
17	0.10031	-64.91847	131.8369	132.545	0.15700	0.80000
18	1.03129	-56.48505	114.9701	116.4965	0.13838	0.53300
19	0.42634	-43.55587	89.11175	89.8198	0.25012	0.25870
20	2.32157	-14.23687	30.47375	31.46948	0.42648	0.00081
21	0.17419	-168.2985	338.5971	340.4257	0.11814	0.54210
22	0.09375	-122.2735	246.5469	247.9809	0.32455	0.00208
23	0.03095	-391.7607	785.5215	787.7981	0.16467	0.04030
24	0.01319	-303.383	608.7659	610.5501	0.28070	0.00147
25	0.54521	-101.9971	205.9941	207.683	0.22053	0.04087
26	1.03319	-126.326	254.6521	256.9828	0.14741	0.06621

5. Conclusions

In this paper, a rigorous review work is done for XGD. Evolution of XGD is discussed in detail. Different forms of XGD are systematically summarized in a tabular format tabular form with 4 columns: serial number, year, authors and model. CDF and HRF are useful to characterize any probability distribution. Hence, we reported the CDF and HRF of all developed model for baseline XGD are described with their mathematical formula to summarize all development or

extension of XGD. Also all the data sets of various fields are reported which are used by authors for validation of several extended or modified XGD distribution. This study serves as a valuable resource for researchers, providing a comprehensive understanding of the wide applicability and acceptance of XGD. It also offers a foundation for developing new versions of XGD by highlighting the existing advancements. Furthermore, the findings provide industrial practitioners with multiple options to select from XGD and its variants, as the review establishes that XGD and its extensions are suitable for a broad range of data types and application domains. In future, researchers can use these different models in industry and also find gap in the presented literature, and may develop new version XGD which is not attempted by authors in past.

References

Abulebda, M., Pathak, A. K., Pandey, A., & Tyagi, S. (2022). On a Bivariate XGamma Distribution Derived from Copula. Statistica, 82(1), 15-40.

Al-Mofleh, H., & Sen, S. (2019). The wrapped xgamma distribution for modeling circular data appearing in geological context. arXiv preprint arXiv:1903.00177.

Alomair, A. M., Babar, A., Ahsan-ul-Haq, M., & Tariq, S. (2025). An improved extension of Xgamma distribution: Its properties, estimation and application on failure time data. Heliyon, 11(2).

Altun, E., & Hamedani, G. G. (2018). The log-xgamma distribution with inference and application. Journal de la Société Française de Statistique, 159(3), 40-55.

Bantan, R., Hassan, A. S., Elsehetry, M., & Kibria, B. M. (2020). Half-logistic xgamma distribution: Properties and estimation under censored samples. Discrete Dynamics in Nature and Society, 2020.

Boudjerda, K. (2025). A new power xgamma distribution: statistical properties, estimation methods and application. Statistics, Optimization & Information Computing, 13(2), 716-740.

Cordeiro, G. M., Altun, E., Korkmaz, M. C., Pescim, R. R., Afify, A. Z., & Yousof, H. M. (2020). The xgamma family: Censored regression modelling and applications. Revstat-Statistical Journal, 18(5), 593-612.

Hashmi, S., Ahsan-ul-Haq, M., Zafar, J., & Khaleel, M. A. (2022). Unit Xgamma distribution: its properties, estimation and application: Unit-Xgamma distribution. Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences, 59(1), 15-28.

Hassan, A., Wani, S. A., & Shafi, S. (2020). Pak. j. statist. 2020 vol. 36 (1), 73-89 lindley-quasi xgamma distribution: Properties and applications. Pak. J. Statist, 36(1), 73-89.

Maiti, S. S., Dey, M., & Sarkar, S. (2018). Discrete xgamma distributions: properties, estimation and an application to the collective risk model. Journal of Reliability and Statistical Studies, 117-132.

Mazucheli, J., Bertoli, W., Oliveira, R. P., & Menezes, A. F. (2020). On the discrete quasi Xgamma distribution. Methodology and Computing in Applied Probability, 22(2), 747-775.

Mohamed, M. I., Handique, L., Chakraborty, S., Butt, N. S., & Yousof, H. M. (2021). A new three-parameter Xgamma Fréchet distribution with different methods of estimation and applications. Pakistan Journal of Statistics and Operation Research, 291-308.

Para, B. A., Jan, T. R., & Bakouch, H. S. (2020). Poisson Xgamma distribution: A discrete model for count data analysis. Model Assisted Statistics and Applications, 15(2), 139-151.

Sen, S., & Chandra, N. (2017). The quasi xgamma distribution with application in bladder cancer data. Journal of data science, 15(1), 61-76.

Sen, S., Afify, A. Z., Al-Mofleh, H., & Ahsanullah, M. (2019). The quasi xgamma-geometric distribution with application in medicine. Filomat, 33(16), 5291-5330.

Sen, S., Alizadeh, M., Aboraya, M., Ali, M. M., Yousof, H. M., & Ibrahim, M. (2024). On truncated versions of xgamma distribution: Various estimation methods and statistical modelling. Statistics, Optimization & Information Computing, 12(4), 943-961.

Sen, S., Chandra, N., & Maiti, S. S. (2017). The weighted xgamma distribution: Properties and application. Journal of Reliability and Statistical Studies, 10(1), 43-58.

Sen, S., Korkmaz, M. C., & Yousof, H. M. (2018). The quasi xgamma-Poisson distribution. Theory and Applications of Statistics and Information, 18(3), 65-76.

Sen, S., Maiti, S. S., & Chandra, N. (2016). The xgamma distribution: statistical properties and application. Journal of Modern Applied Statistical Methods, 15(1), 38.

Shukla, S., Yadav, A. S., Rao, G. S., Saha, M., & Tripathi, H. (2022). Alpha Power Transformed Xgamma Distribution and Applications to Reliability, Survival and Environmental Data. Journal of Scientific Research, 66(3).

Tripathi, H., & Mishra, S. (2022). The transmuted inverse xgamma distribution and its statistical properties. International Journal of Mechanical Engineering, Kalahari Journals , 7.

Tripathi, H., Yadav, A. S., Saha, M., & Kumar, S. (2025). Generalized Inverse Xgamma Distribution: Properties, Estimation and Its Applications To Survival Data. Thailand Statistician, 23(3), 573-597.

Tripathi, H., Yadav, A. S., Saha, M., & Shukla, S. (2022). A New Flexible Extension of Xgamma Distribution and its Application to COVID-19 Data. Nepal Journal of Mathematical Sciences, 3(1), 11-30.

Wani, S. A., & Shafi, S. (2021). Generalized Lindley-Quasi Xgamma distribution. Journal of Applied Mathematics, Statistics and Informatics, 17(1), 5-30.

Wani, S. A., Hassan, A., Sheikh, B. A., & Dar, S. A. (2022). Size Biased Lindley-Quasi Xgamma Distribution Applicable to Survival Times. Thailand Statistician, 20(2), 452-474.
Yadav, A. S, Shukla, S., Jaiswal, N., Singh, S. K., & Koley, D. (2023). Development and Estimation of Weighted Xgamma Exponential Distribution with Applications to Lifetime Data. Statistica, 83(4), 247–269. https://doi.org/10.6092/issn.1973-2201/16940
Yadav, A. S., Maiti, S. S., & Saha, M. (2021). The inverse xgamma distribution: statistical properties and different methods of estimation. Annals of Data Science, 8(2), 275-293.
Yadav, A. S., Saha, M., Tripathi, H., & Kumar, S. (2021). The Exponentiated XGamma Distribution: a New Monotone Failure Rate Model and Its Applications to Lifetime Data. Statistica, 81(3), 303-334.
Yadav, A. S., Shukla, S., Goual, H., Saha, M., & Yousof, H. M. (2022). Validation of Xgamma exponential model via Nikulin–Rao–Robson goodness-of-fit test under complete and censored samples with different methods of estimation. Statistics, Optimization & Information Computing, 10(2), 457–483.
Yousof, H. M., Korkmaz, M. Ç., & Sen, S. (2021). A new two-parameter lifetime model. Annals of Data Science, 8(1), 91-106.

Declaration

Conflict of Study: The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.
Ethical Approval: Not Applicable
Consent to Participate: Not Applicable
Acknowledgement: Not Applicable

SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.1

Modelling Wind Speed at Ikeja Station using Skewed Statistical Distributions

Olalude, G. A. [1], Afolabi, S. A. [2*], Olaleye, O. A.[1], Folorunso, S. A. [3]

[1]Department of Statistics, Federal Polytechnic, Ede, Nigeria

[2]Department of Mathematics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

[3]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

* Corresponding Email: afolabisa63@gmai.com

Received: 06 October 2025 / Revised: 12 November 2025 / Accepted: 03 December 2025 / Published online: 10 December 2025

ABSTRACT

Wind speed, a key atmospheric parameter, results from the movement of air from high- to low-pressure regions driven primarily by temperature variations. This study modelled the wind speed (m/s) of Ikeja, Lagos State, using data from 2000 to 2020 through statistical techniques such as descriptive statistics, data visualization, and goodness-of-fit (GOF) tests, including chi-square, Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises analysis. Three positively skewed distributions, Gamma, Weibull, and Log-normal, were evaluated. Descriptive analysis indicated that the dataset was predominantly right-skewed (Skp > 0). The GOF results show that the Weibull distribution provides the best representation of the wind speed data (p=0.02), followed by the distributions of Log-normal and Gamma. The Weibull parameter ($\alpha > 1$) further confirmed its suitability for the data. The findings suggest that Ikeja may experience higher wind speeds in the future, emphasizing the need for precautionary measures to mitigate potential damage to infrastructure and property. This study provides valuable insights for meteorologists and urban planners in anticipating and managing climate-related risks in the city.

Keywords*:* Climate change; Data visualization; Meteorologists; Precautionary measures; Temperature variations

1. Introduction

Wind energy is a promising and abundant form of green energy available across many regions of the world. Its growing importance stems from factors like the decline in fossil fuel reserves, increasing global energy demand, and the urgent need to address climate change (Masseran, 2015). The negative impacts of climate change have prompted many nations to adopt sources of renewable energy, such as wind, resulting in the rapid global expansion of wind power generation (Chang et al., 2015). Among its key advantages are zero fuel costs, minimal price volatility, and enhanced energy security. However, its major drawback lies in the intermittent and unpredictable nature of wind (Afanasyeva et al., 2016). Excluding large hydropower, wind energy remains the most widely utilised renewable source for electricity generation, having grown nearly 300-fold since 1990. China, the United States, Germany, Spain, and India lead global developments in the sector (Abbasi & Abbasi, 2016).

A study was conducted by (Afolabi, 2020) on modelling water pollutants using Weibull and Lognormal distributions, analyzing thirteen pollutants from two major rivers in Oyo State in evaluating water safety. The findings revealed that not all pollutants exhibited right-skewed behaviour, and the distribution of Lognormal provided a better fit than the Weibull distribution based on the tests of GOF. Accurate wind distribution modelling requires multi-year data to capture variability and temporal patterns. To reduce the cost and time of processing long-term datasets, statistical distribution functions, notably probability density functions (PDFs), are widely applied to describe wind speed characteristics, with parameters commonly estimated using the Maximum Likelihood Estimation (MLE) method. Traditional models like Weibull, Lognormal, Gamma, and GEV distributions, along with mixture forms like Weibull-Lognormal and GEV-Lognormal, have proven effective in characterizing wind speed variability (Kollu et al., 2012). The integration of both wind speed and direction enhances modelling accuracy and adaptability under changing climatic conditions (Murphy et al., 2025).

The challenges encountered via wind energy development in West Africa are largely attributed to inadequate measurement data, limited assessment studies, and inconsistent wind resource classification across the region. Since wind resources are highly site-specific, accurate national wind profiling requires assessments across multiple locations. In Nigeria, wind resource evaluation has progressed through several developmental stages, with various policy initiatives reflecting the government's commitment to harnessing wind energy for electricity generation (Ajayi, 2013). Scholars such as (Akgül et al., 2016) emphasised that the Weibull distribution is widely used for modelling wind speed data, but it may not always yield accurate results. Their study applied the distribution of Inverse Weibull (IW) to wind speed modelling in Turkey and found that parameter estimation using MLE and Modified Maximum Likelihood (MML) methods produced more accurate results than the traditional Weibull model at most stations.

Weibull distribution is applied to wind speed data from 2000–2023 across selected African stations, revealing significant regional variability in wind characteristics and power density, and highlighting the continent's strong potential for wind energy development and the need for region-specific renewable energy policies (Aweda & Samson, 2024). The wind speed analysis at two altitudes ($50m$ and $100m$) in a tropical coastal location was carried out by (Ogunjo, 2025) using multifractal and Weibull statistical methods, revealing that multifractality is influenced by land-sea breezes and demonstrating the Weibull distribution's superior fit for characterizing wind behaviour essential for modelling and energy applications.

Recent studies have advanced the modelling of wind speed by introducing more flexible statistical frameworks capable of capturing asymmetry, skewness, and heavy-tailed behaviour in environmental data. (Chaturvedi et al., 2025) made a proposition of the generalized positive exponential family of distributions as a robust alternative to traditional models, demonstrating improved performance in representing the variability and non-normal characteristics of wind speed data. Similarly, (Lencastre et al., 2024) evaluated wind-speed distributions beyond the conventional Weibull model and showed that alternative skewed and heavy-tailed distributions can offer superior fit under diverse wind regimes. Building on these recent developments, the present study applies skewed statistical models to wind speed data from Ikeja Station, emphasizing their suitability for characterizing non-normality and enhancing the accuracy of wind-speed modelling.

2. Methodology

This section outlines the methodologies adopted to achieve the research objectives, focusing on Exploratory Data Analysis (EDA) and the application of skewed probability distributions for wind speed modelling. The EDA approach was employed to examine the structure and characteristics of the wind speed data, identify trends and patterns, and evaluate statistical properties such as central tendency, dispersion, and skewness.

2.1 Descriptive Statistics

Minimum

$$min(X) = min\{x_1, x_2 \dots, x_n\} \tag{1}$$

Maximum

$$min(X) = max\{x_1, x_2 \dots, x_n\} \tag{2}$$

Range

$$\text{Range} = \max(X) - \min(X) \tag{3}$$

Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

Standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{5}$$

Skewness

$$\gamma_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{(X_i - \bar{X})^3}{s^3} \tag{6}$$

The arithmetic mean is the $\bar{x}$, the total number of observations is n, and the standard deviation raised to the third power is s^3, with X being the wind speed of interest. The histogram of wind speeds in Lagos State is presented to examine the data distribution. Additionally, the density plot and the Cullen and Frey plots are provided to assess the plausible distributions of the wind speed data.

2.2 Modelling the Distribution of Wind Speeds

The three distributions, which are Gamma, Weibull and Log-normal, considered in this study are presented below. The mathematical frameworks of each statistical distribution were provided to understand the computational aspect of each distribution.

Distribution of Gamma

The PDF - probability density function of the gamma distribution for wind speed (x), with two parameters α and β is expressed as:

$$f(x;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x},x>0 \quad (7)$$

Where $\alpha,\ \beta,x$ are the parameters of the shape with rate or the scale of the gamma distribution, and the wind speed values, respectively.

Distribution of Weibull

The PDF – probability density function of the Weibull distribution for the wind speeds (x), with two parameters α and β is defined as:

$$f(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}e^{-(x/\beta)^{\alpha}},x>0 \quad (8)$$

Where α, β, and x are the parameters of the shape with the scale of the Weibull distribution, and wind speed values, respectively.

Distribution of Log-Normal

The PDF – probability density function of the log-normal distribution for wind speeds (x), with two parameters μ and σ is given by:

$$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\},x>0 \quad (9)$$

Where σ, μ, and x are the parameters of the shape with the scale of the Log-normal distribution and wind speeds, respectively.

2.3 Parameter Estimation

For this study, both MLE – Maximum Likelihood Estimation and MOM – Method of Moment were employed.

2.3.1 Maximum Likelihood Estimation (MLE)

Likelihood Function of Gamma Distribution

$$L(\alpha,\beta\mid x_1,\dots,x_n)=\prod_{i=1}^{n}\frac{\beta^{\alpha}}{\Gamma(\alpha)}x_i^{\alpha-1}e^{-\beta x_i}$$

$$=\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{n}\left(\prod_{i=1}^{n}x_i^{\alpha-1}\right)exp\left(-\beta\sum_{i=1}^{n}x_i\right)$$

The log-likelihood function is given thus;

$$\ell(\alpha,\beta)=n\alpha\ln\beta-n\ln\Gamma(\alpha)+(\alpha-1)\textstyle\sum_{i=1}^{n}\ln x_i-\beta\sum_{i=1}^{n}x_i \quad (10)$$

Likelihood Function of Weibull Distribution

$$L(\alpha,\beta\mid x_1,\dots,x_n)=\prod_{i=1}^{n}\frac{\alpha}{\beta}\left(\frac{x_i}{\beta}\right)^{\alpha-1}exp\left[-\left(\frac{x_i}{\beta}\right)^{\alpha}\right]$$

$$=\left(\frac{\alpha}{\beta}\right)^{n}\left(\prod_{i=1}^{n}x_i^{\alpha-1}\right)\beta^{-n(\alpha-1)}exp\left(-\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha}\right)$$

The function of log-likelihood is expressed as;

$$\ell(\alpha,\beta)=n(\ln\alpha-\ln\beta)+(\alpha-1)\textstyle\sum_{i=1}^{n}\ln x_i-\alpha n\ln\beta-\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^{\alpha} \quad (11)$$

Likelihood Function of Lognormal Distribution

$$L(\mu,\sigma\mid x_1,\dots,x_n)=\prod_{i=1}^{n}\frac{1}{x_i\sigma\sqrt{2\pi}}exp\left[-\frac{(\ln x_i-\mu)^2}{2\sigma^2}\right]$$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n}\prod_{i=1}^{n}x_i^{-1}exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i-\mu)^2\right]$$

The function of the log-likelihood is given as;

$$\ell(\mu,\sigma)=-n\ln(\sigma\sqrt{2\pi})-\textstyle\sum_{i=1}^{n}\ln x_i-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i-\mu)^2 \quad (12)$$

All likelihoods were optimized numerically using Newton–Raphson algorithms in R (fitdistrplus package).

2.3.2 Method of Moments (MOM)

Gamma MOM is expressed as:

$$\alpha=\frac{\bar{x}^2}{s^2},\beta=\frac{\bar{x}}{s^2}. \quad (13)$$

Weibull MOM is solved numerically via:

$$\frac{s}{\bar{x}} = \sqrt{\frac{\Gamma(1+2/\alpha)}{\Gamma(1+1/\alpha)^2} - 1}. \quad (14)$$

Lognormal MOM is given as;

$$\hat{\mu} = \frac{1}{n}\sum \ln x_i, \hat{\sigma}^2 = \frac{1}{n}\sum (\ln x_i - \hat{\mu})^2. \quad (15)$$

Parameter estimation was performed in R programming using the **fitdistrplus** package, which applies numerical maximum-likelihood optimization.

2.4 Tests of Goodness of Fit (GOF)

The tests of GOF are used to assess whether it is rational to make an assumption that a given random sample was drawn or came from a specified distribution. They operate as a form of setting hypotheses, in which the hypotheses of null and alternative are defined as:

H_0: The sample data do not follow a specified distribution.

H_1: The sample data do follow a specified distribution.

2.4.1 Test of Hypothesis

With X represents the amount of wind speed, the hypothesis tests are formulated below;

$H_0: X \sim /\text{Gamma}(\alpha, \beta)\ vs. H_1: X \sim \text{Gamma}(\alpha, \beta)$.

$H_0: X \sim /\text{Weib}(\alpha, \beta)\ vs. H_1: X \sim \text{Weib}(\alpha, \beta)$.

$H_0: X \sim /\text{Lognorm}(\mu, \sigma)\ vs. H_1: X \sim \text{Lognorm}(\mu, \sigma)$.

2.4.2 Measures of Goodness of Fit

The following GOF measures were applied:

- Chi-square test.
- Kolmogorov-Sminorv (KS).
- Anderson-Darling (AD).
- Cramer-von Mises (CvMS)

Chi-square Test Statistic and P-value

Let O_i be the observed count and E_i be the expected count in the bin i:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (16)$$

Where k is the number of bins.

Degrees of freedom: $df = k - p - 1$**.** Where p is the number of parameters estimated.

P-value: Following the chi-square statistic above:

$$p\text{-value} = 1 - F_{\chi^2_{df}}(\chi^2) \quad (17)$$

Where the chi-square CDF - cumulative distribution function is $F_{\chi^2_{df}}$ with degrees of freedom as df.

Binning approach: Bins were constructed using Sturges' rule, $k = 1 + 3.3\log_{10}(n)$.

3. Empirical Results

A descriptive analysis of wind speed in Ikeja, Lagos, was conducted to examine its statistical characteristics. The results are summarized in Table 1. Wind speed in Ikeja is right-skewed (skewness = 0.818), with a mean of 9.062 m/s and a standard deviation of 3.127 m/s. The wind speeds spanned from 4.05 m/s to 18.15 m/s over the study period. The 21-year dataset reveals fluctuating patterns, suggesting that wind speeds in Ikeja display stochastic characteristics with a pronounced skewed distribution. Figure 1 presents the histogram and cumulative distribution of wind speed.

Table 1*: Descriptive Statistics of Wind Speeds (m/s) in Ikeja, Lagos*

Statistics	Wind Speeds (m/s)
Mean	9.062
Standard Deviation	3.127
Skewness	0.818
Minimum	4.05
Maximum	18.15

The histogram (Figure 2) shows a positively skewed distribution with a long, thin tail. Most wind speed observations in Ikeja, Lagos, fall between 4 m/s and 12 m/s, with fewer values ranging from 12.1 m/s to 18.15 m/s. Figure 2 also presents the kernel density estimate, the Normal Q-Q plot, and the Cullen and Frey plot for wind speed in Ikeja.

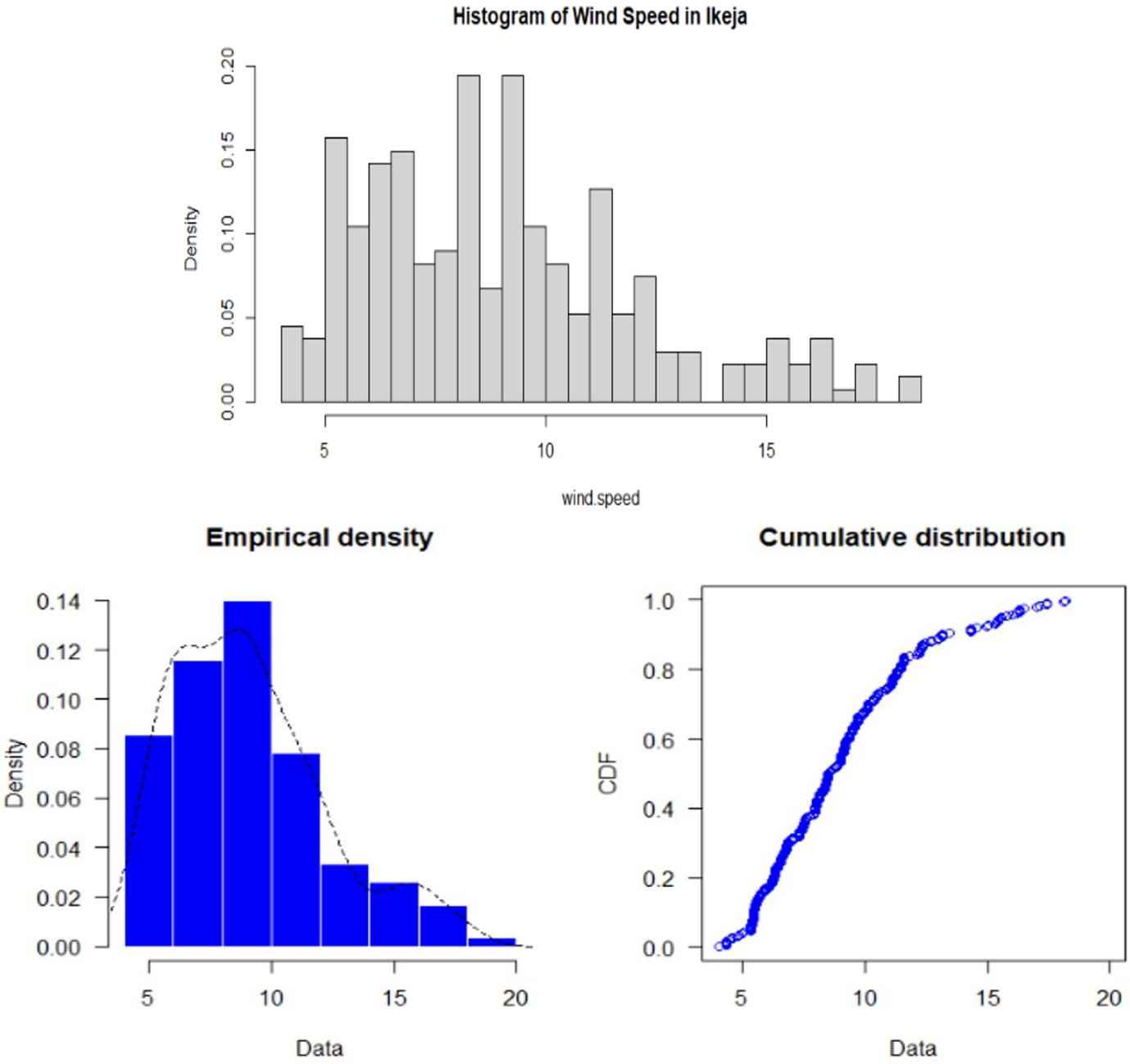

***Figure 1**: Wind Speed Variation in Ikeja*

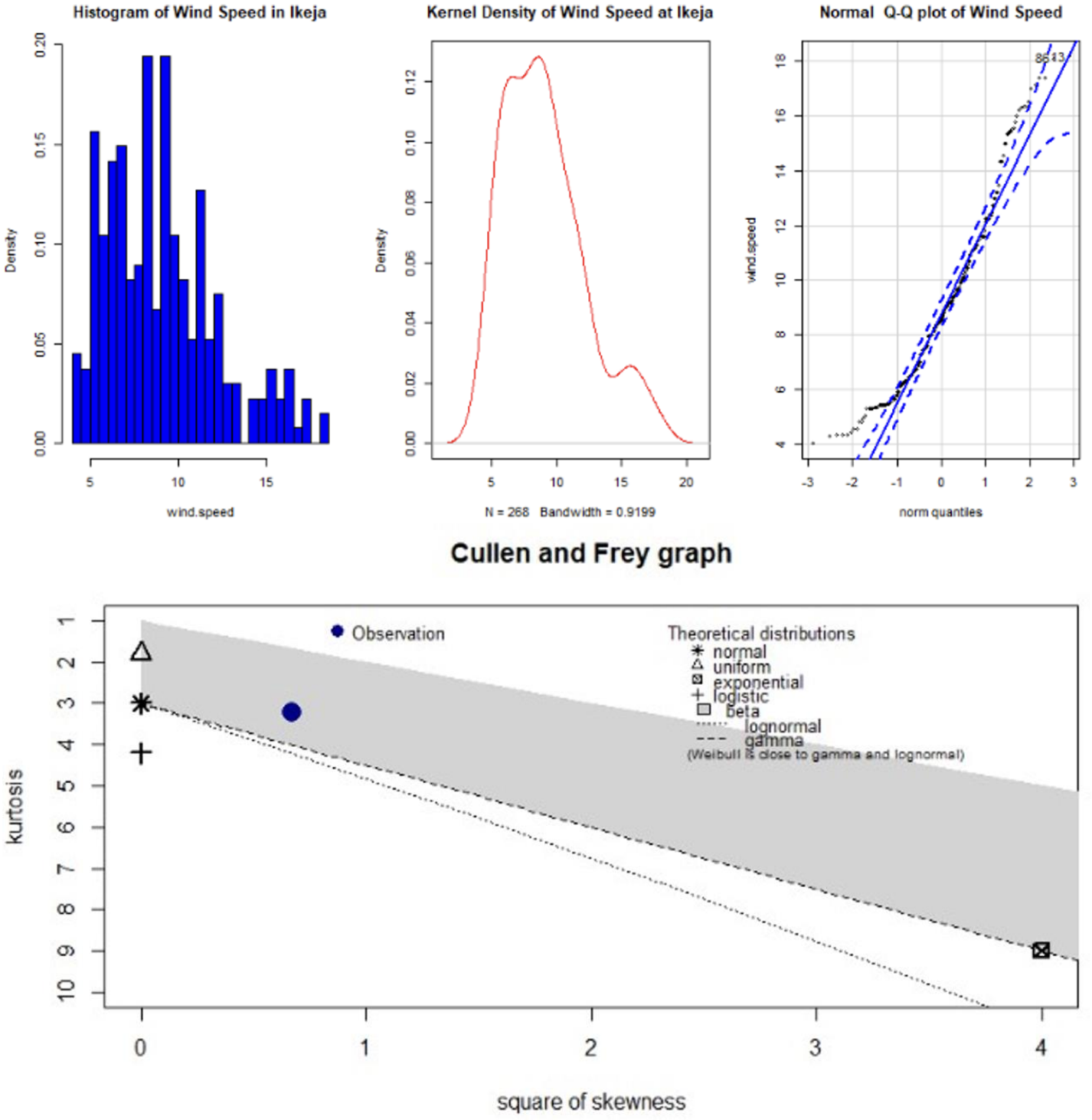

Figure 2: *Wind Speed Exploratory Data Analysis*

The plots established the appropriate suggestion of the data patterns in terms of skewness and suggest the likely fitted distributions as shown in the Cullen and Frey plot. The Normal Q-Q plot shows that the data deviated from normality, indicating that the distribution of normal is not suitable for consideration. It is then claimed that rightly skewed distributions should be adopted for modelling the data of wind speed. Having justified the selections of the distribution, the analysis proceeds with fitting the Weibull, gamma, and lognormal distributions.

3.1 Wind Speed Distribution Models

Three different continuous distribution models were fitted to the wind speed (m/s) data in Ikeja, Lagos, using two methods of estimation: Maximum Likelihood Estimation (MLE) and the Method of Moments (MOM). The estimated values of the distribution parameters for the Gamma, Weibull, and Lognormal distributions are presented in Table 2. Varying parameter values were obtained from the two estimation methods. Figure 3 compares the Weibull, Gamma, and Lognormal fits to the wind speed data using four diagnostic plots: density, Q–Q, CDF, and P–P. Together, they

show how closely each distribution matches the observed data. Among the fitted models, the distribution of Weibull provides the best fit to the wind speed data compared to the Gamma and Lognormal distributions.

***Table 2**: Estimated Parameters of the Gamma, Weibull and Lognormal Distributions*

Methods	Parameters	Gamma	Weibull	Lognormal
MLE	Shape = α(S.E)	39.783 (12.231)	9.017 (1.657)	-
	Rate = β(S.E)	4.198 (1.298)	-	-
	Scale = β(S.E)	-	10.0524 (0.254)	-
	MeanLog = μ (S.E)	-		2.237 (0.00357)
	sdLog = σ (S.E)	-		0.164 (0.025)
MOM	Shape = α	44.566	9.012	-
	Rate = β	4.701	-	-
	Scale = β		10.044	-
	MeanLog = μ			2.238
	sdLog = σ			0.149

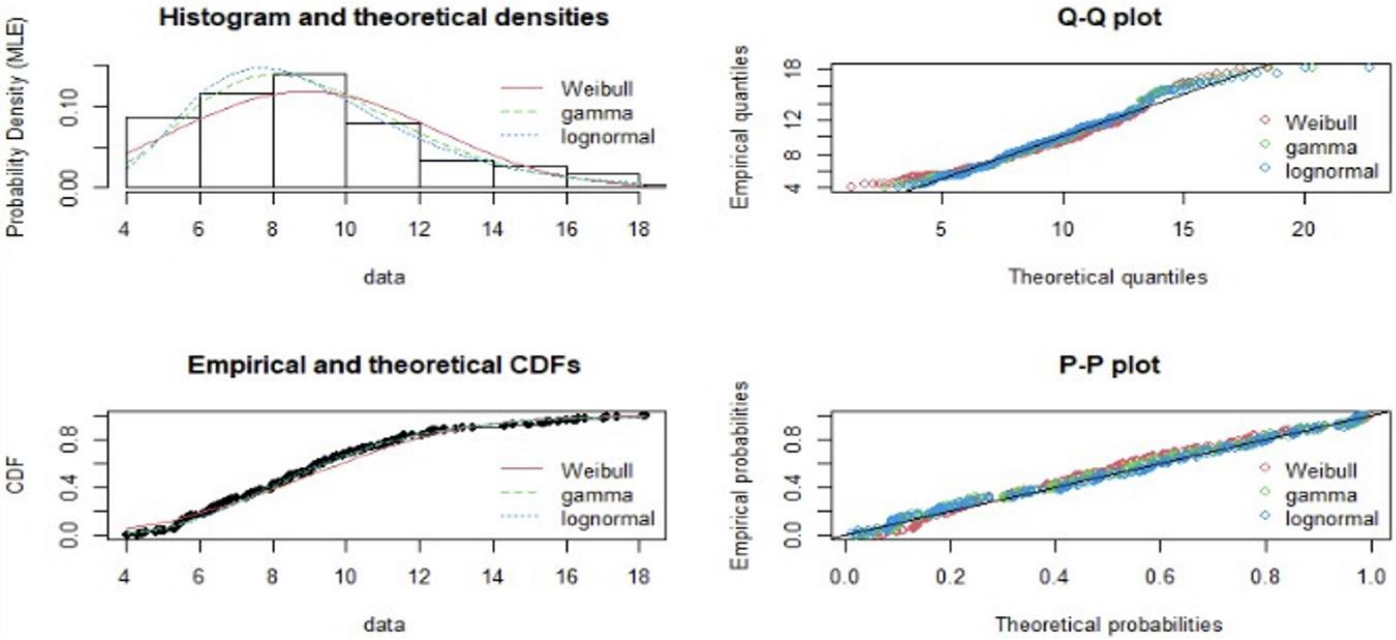

***Figure 3:** Model Diagnostics for Weibull, Gamma, and Lognormal Wind Speed Fits*

3.2 Efficiency of Estimates and Distributions

To obtain a better estimator of the MOM and MLE estimator for the wind speed (m/s) in Ikeja, Lagos, Nigeria, the following statistical tools were used based on the minimum value(s). Table 3 shows that the Weibull Distribution best described the Wind Speed (m/s) of Ikeja, Lagos State, followed by the Log-Normal Distribution, then the Gamma Distribution.

***Table 3**: Performance Comparison of MLE and MOM Estimation Methods*

Distribution	**Statistics**							**Remark**
Gamma	Log-likelihood	AIC	BIC	KS - value	CvMS	ADS	MSE	
MLE	-296.6432	**598.316**	**604.062**	0.0749	**0.1826**	**2.235**	3.724	**Good**
MOM	**-297.9629**	598.824	604.680	**0.0696**	0.1834	2.371	**3.491**	
$Gam(\alpha,\beta)$	-297.3031	598.570	604.371	**0.0723**	0.183	2.303	3.608	
Weibull								
MLE	-283.3128	**580.725**	**586.391**	0.0347	**0.0201**	0.1848	**3.628**	**BEST**
MOM	**-283.4161**	580.813	586.453	**0.2846**	0.0213	**0.10**	**3.628**	
$Weib(\alpha,\beta)$	-283.3645	**580.769**	**586.422**	0.160	**0.0207**	**0.1424**	3.628	
Log-Normal								
MLE	-302.1687	**605.347**	**612.141**	0.0943	0.311	**1.8214**	**3.421**	**Better**
MOM	**-302.705**	608.131	615.289	**0.0867**	**0.300**	2.153	3.331	
$Lnorm(\mu,\sigma)$	**-302.4369**	606.739	613.715	0.0905	0.3055	1.9872	**3.376**	

As revealed in Table 3, the distribution of Weibull offers the most accurate representation of wind speeds (m/s) in Ikeja, Lagos State, with the Log-Normal distribution next in performance, and the Gamma distribution performing the least.

3.3 Chi-square Goodness of Fit

Sequel to the results of the efficiency, it is also of interest to obtain a measure of goodness of fit to ascertain a better distributional pattern of the wind speed in Ikeja, Lagos, Nigeria. Table 4 provides the results obtained using the chi-square GOF for the Gamma, Weibull, and Lognormal distributions under both MLE - maximum likelihood and MOM - method of moments estimation. For each model, the chi-square statistics, degrees of freedom, and corresponding p-value are shown. Across all cases, the p-values are below the 0.05 significance level, indicating that the test of chi-square rejects the null hypothesis for each distribution. Although all models are rejected by this test, the Weibull distribution shows comparatively smaller chi-square values and higher p-values than the Gamma and Lognormal distributions, suggesting relatively better agreement with the observed wind-speed data.

***Table 4**: Chi-Square GOF Results for the Distribution*

Distributions	Methods	$\alpha - level$ = 0.05		
		χ^2 - value	df	p-value
Gamma	MLE	17.635	3	0.001
	MOM	17.937	3	0.0005
		17.786		0.00075
Weibull	MLE	16.867	3	0.0007
	MOM	15.020	3	0.002
		15.944		0.00135
Log-Normal	MLE	18.319	3	0.0004
	MOM	18.243	3	0.0004
		18.281		0.00007

The result in Table 4 establishes that among the three distributions applied to the fitness of the wind speed (m/s), the Weibull distribution is the best. In other words, the fitness of the wind speed in Ikeja, Lagos, might be arranged as $\boldsymbol{Weib}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) \geq \boldsymbol{Lnorm}(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\sigma}) \geq \boldsymbol{Gam}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$, which is also consistent with Figure 4.

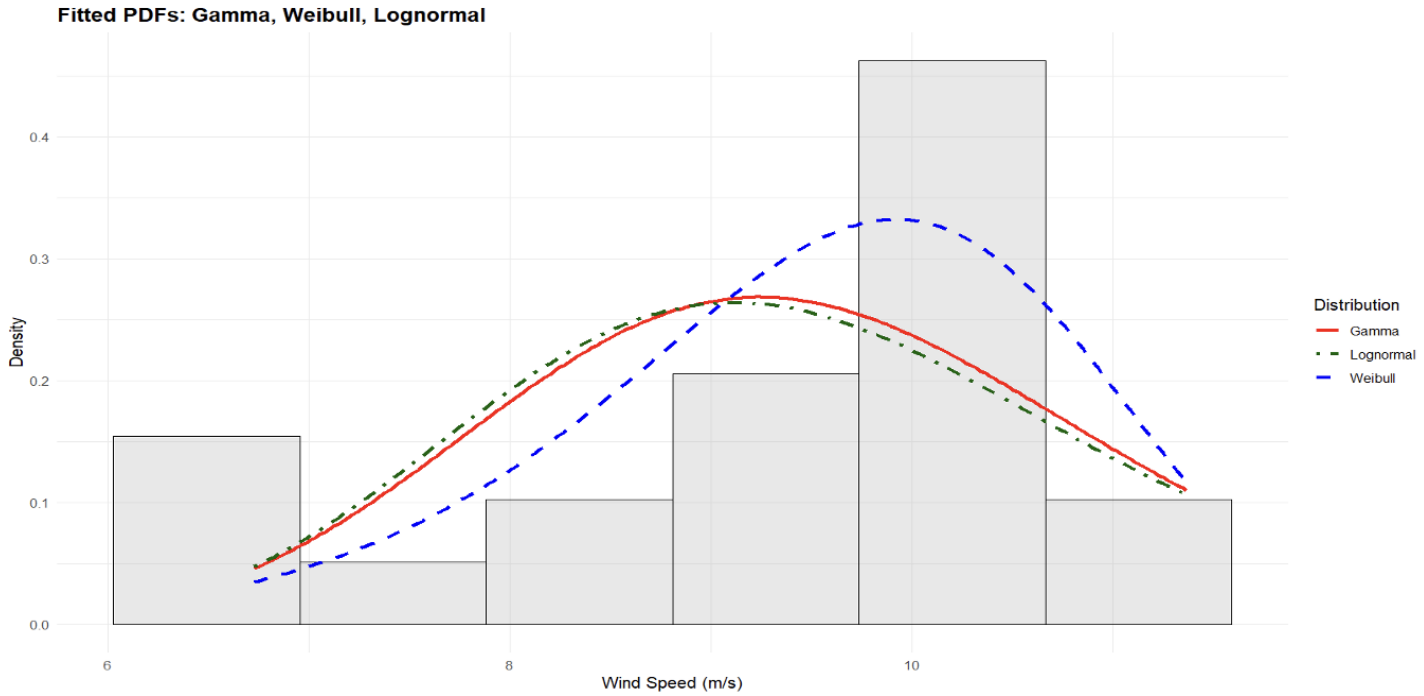

***Figure 4**: Fitted Distribution Curves*

The overlaid density plot in Figure 4 shows that the Weibull model (blue dashed curve) most closely follows the histogram of wind speeds, particularly around the modal region and in the upper tail, indicating that it is the most fitted distribution. The Lognormal model (green dot–dash) provides a moderate fit, while the Gamma model (red solid) shows the largest discrepancies, especially near the peak and at higher wind speeds.

4. Conclusion

In this study, distributions of Gamma, Weibull, and Lognormal were fitted to wind speed data at Ikeja, Lagos State. The results indicate that the Weibull distribution provides the best overall fit under both Maximum Likelihood Estimation – MLE and the Method of Moments - MOM, with both methods producing similar parameter estimates. The Lognormal

distribution gives a reasonable fit under MLE, although MOM performs moderately for this model. For the Gamma distribution, MLE yields acceptable results, whereas MOM performs fairly. Overall, the Weibull distribution is recommended as the most suitable model for wind speed in Ikeja, outperforming both the Lognormal and Gamma distributions.

The distribution of the Weibull is widely utilized in wind engineering because its shape parameter $\boldsymbol{\alpha}$ describes wind variability, while the scale parameter $\boldsymbol{\beta}$ determines energy potential through wind power density. Therefore, the Weibull model adopted in this study provides a practical basis for wind resource assessment and turbine selection in urban environments like Ikeja. This study is restricted to just fitting three models, Gamma, Weibull, and Lognormal, to the wind speed in Ikeja, Lagos State. However, looking at the display of the Cullen and Frey graph in Figure 2, it was observed that the Beta distribution could be considered in comparison with the Weibull distribution. Future study could consider fitting a Beta distribution for further justification of this study's findings.

REFERENCES

Abbasi, S., & Abbasi, T. (2016). Impact of wind-energy generation on climate: A rising spectre. *Renewable and Sustainable Energy Reviews*, *59*, 1591-1598.

Afanasyeva, S., Saari, J., Kalkofen, M., Partanen, J., & Pyrhönen, O. (2016). Technical, economic and uncertainty modelling of a wind farm project. *Energy Conversion and Management*, *107*, 22-33.

Afolabi, S. A. (2020). Modelling of Water Pollutants with Weibull and Lognormal Distribution: A Case Study of Oyo State, Nigeria.

Ajayi, O. (2013). Sustainable energy development and environmental protection: The case of five West African Countries. *Renew. Sustain. Energy Rev*, *26*, 532-539.

Akgül, F. G., Şenoğlu, B., & Arslan, T. (2016). An alternative distribution to Weibull for modeling the wind speed data: Inverse Weibull distribution. *Energy Conversion and Management*, *114*, 234-240.

Aweda, F., & Samson, T. (2024). Statistical Analysis of wind speed characteristics using the Weibull distribution at selected African stations. *J. Sustain. Energy*, *3*(3), 187-197.

Chang, T.-J., Chen, C.-L., Tu, Y.-L., Yeh, H.-T., & Wu, Y.-T. (2015). Evaluation of the climate change impact on wind resources in Taiwan Strait. *Energy Conversion and Management*, *95*, 435-445.

Chaturvedi, A., Bhatti, M. I., Bapat, S. R., & Joshi, N. (2025). Modeling wind speed data using the generalized positive exponential family of distributions. *Modeling Earth Systems and Environment*, *11*(2), 98.

Kollu, R., Rayapudi, S. R., Narasimham, S., & Pakkurthi, K. M. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of energy and environmental engineering*, *3*(1), 27.

Lencastre, P., Yazidi, A., & Lind, P. G. (2024). Modeling wind-speed statistics beyond the Weibull distribution. *Energies*, *17*(11), 2621.

Masseran, N. (2015). Markov chain model for the stochastic behaviors of wind-direction data. *Energy Conversion and Management*, *92*, 266-274.

Murphy, E., Huang, W., Bessac, J., Wang, J., & Kotamarthi, R. (2025). Joint modeling of wind speed and wind direction through a conditional approach. *Environmetrics*, *36*(3), e70011.

Ogunjo, S. (2025). Wind energy characterization using multifractal formalism at two different altitudes in a tropical country. *Modeling Earth Systems and Environment*, *11*(1), 63.

Declaration

Conflict of Study: The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution Statement: Olalude, G.A., Olaleye, O. A., and Folorunso, S. A. conceived the idea and designed the research; then Afolabi S. A. analyzed and interpreted the data and wrote the paper.

Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.

Consent to Publish: All authors have agreed to publish this Journal.

Ethical Approval: Not Applicable

Consent to Participate: All participants were made aware that their responses would be kept anonymous and confidential and that the data would only be utilized for the study.

Acknowledgement: The author is thankful to the editor and reviewers for their valuable suggestions to improve the quality and presentation of the paper.

Author(s) Bio / Authors' Note

***Gbenga Adelekan Olalude**:*

***G.A.O** is a Chief Lecturer at the Department of Mathematics and Statistics, Federal Polytechnic Ede, Osun State, Nigeria. He holds a Master's degree in Statistics. His main areas of interest include time-series analysis, Bayesian inference, and mathematical statistics. He has published more than 50 research papers in international/national journals and conferences. He is a member of the Nigerian*

Statistical Association, the American Statistical Association, and the Professional Statisticians Society of Nigeria. Email: olaludelekan@yahoo.com

Saheed Abiodun Afolabi:

***S.A.A** is a PhD student and a Graduate Assistant (GA) at the College of Computing and Mathematics, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He recently completed his master's degree in applied statistics at the same KFUPM, Saudi Arabia, with a distinction after his undergraduate and graduate program at the University of Ibadan, Nigeria, with a first-class honour through National Diploma in Statistics at the Federal Polytechnic Ede, Osun State, Nigeria. He is a member of different statistical societies like the Nigerian Statistical Association (NSA), Professional Statisticians Society of Nigeria (PSSN), International Biometric Society (IBS), Royal Statistical Society (RSS), and American Statistical Association (ASA); and a research collaborator and volunteer at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis (UI-LISA). Email:* afolabisa63@gmail.com

O. A. OLALEYE:

***O.A.O** (B.Sc., M. Tech, Ph.D) is a Principal Lecturer in the Department of Statistics Federal Polytechnic, Ede, Osun state, Nigeria and an External Examiner in the Department of General Studies at Federal Polytechnic, Ayede, Oyo state, Nigeria, West Africa. He researches more in Fluid dynamics and related fields and has scores of publications along this line, both as a lead author and as a coauthor. He is also a Reviewer for about 20 journal bodies with about 50 manuscripts reviewed by him. Email:* hezekiaholaleye@gmail.com

Dr Serifat A. Folorunso:

***S.A.F** is a statistician and data scientist whose research focuses on the application of advanced statistical and machine learning methods to environmental, economic and health data. She holds a PhD in Statistics and an MSc in Applied Data Science from Teesside University, United Kingdom. Her recent works explore the use of skewed and flexible distributions for modelling environmental variables, including wind speed, water pollution, waste management, and climate-related parameters. Dr Folorunso has authored and co-authored several peer-reviewed publications and actively contributes to interdisciplinary research promoting data-driven solutions for sustainable development. Email:* Serifatf005@gmail.com

Appendix

Calculation of the p-value of a Chi-square Distribution

A. Manual Approach:

Computing the **p-value (*p*)** for a chi-square statistic, one of the approaches is to use the survival function of the distribution of chi-square.

$$p = 1 - F_{\chi_\nu^2}(x)$$

Scenario

- $x = 15.020$ is the chi-square value,
- $df = 3$ is the degree of freedom,
- $F_{\chi_\nu^2}(x)$ is the cumulative distribution function (CDF).

Step-by-Step Mathematical Computation

$$p = 1 - F_{\chi_3^2}(15.020)$$

The chi-square CDF expression for $df = 3$ is given thus:

$$F_{\chi_3^2}(x) = \frac{\gamma\left(\frac{3}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{3}{2}\right)}$$

Where $\gamma(s,t)$ is the lower incomplete gamma function, and $\Gamma(s)$ is the gamma function.

Computing the intermediate values,

- $\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2} \approx 0.88623.$
- $x/2 = 15.020/2 = 7.510$

Computing the lower incomplete gamma function,

- $\gamma\left(\frac{3}{2}, 7.510\right) \approx 0.88533$

Note: This value is taken from standard mathematical tables/computer evaluation.

Computing the CDF, we have

$$F_{\chi_3^2}(15.020) = \frac{0.88533}{0.88623} \approx 0.99899$$

Compute the p-value, we have:

$$p = 1 - 0.99899$$
$$p \approx 0.00101$$

B. Technical Approach with R-software:

R-programming Code

```
p_value <- 1 - pchisq(15.020, df = 3)
```

R-programming Result

```
> p_value <- 1 - pchisq(15.020, df = 3)
> p_value
[1] 0.001799637 ≈ 0.002.
```

SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.2

Assessment of the performance of the Bayesian Forecasting for Vector ARMA Processes

Emad E. A. Soliman [1*], Sherif S. Ali [1]

[1]Department of Statistics, King Abdulaziz University, Jeddah 21589, Saudi Arabia

* Corresponding Email: ehegazy@kau.edu.sa

Received: 29 November 2025 / Revised: 03 December 2025 / Accepted: 04 December 2025 / Published online: 10 December 2025

ABSTRACT

Multivariate time series are widely observed in numerous domains. In economics, for example, you can monitor the annual savings in conjunction with the real interest rate. Such variables are jointly analyzed to understand the dynamic interactions that exist between them, thus improving the precision of forecasts. Enhanced forecasting is achievable when the series are examined together, especially when one series holds information about another one. An approximate Bayesian analytical method to estimate and forecast vector autoregressive moving average (Vector ARMA) processes was introduced by Shaarawy (1989). A basic goal for the research in hand is the numerical assessment of the proposed approach in tackling forecasting difficulties associated with Vector ARMA processes through a comprehensive simulation study. Furthermore, the research checks how the performance of the suggested method fluctuates when varying parameter values and time series lengths. The findings of the numerical study demonstrated that the methodology was effective in accurately forecasting future observations for Vector ARMA processes across various values of the parameter and different time series lengths.

Keywords*:* Vector autoregressive moving average processes; Multivariate t distribution; Bayesian forecasting; Prediction; Matrix Normal-Wishart prior; Jeffreys' prior

1. Introduction

Multivariate time series are encountered in various areas of scientific and applied research. Regarding economics, researchers can investigate the relationships among production cost, product's sales price, sales volume, and advertising expenses using a four-dimensional framework. In meteorology studies, a bivariate technique may be used to estimate daily maximum temperatures in addition to humidity. According to the works of Tiao and Box (1981), Liu (2009), Box et al. (2016), Chatfield (2019), and Wei (2019), the above mentioned and other examples are examined and forecasted collectively using multivariate models, which seeks to reveal the interaction between different phenomena in order to improve forecasts' goodness.

In a variety of application areas, one encounters several time series data sets that can be successfully analyzed, modeled, and forecasted using the class of multivariate (vector) autoregressive moving average models, or VARMA models. The ultimate goal of this analysis is forecasting (or prediction), which is thought to be the most essential stage in the analysis of time series. The used prediction approach determines how reliable the forecasts are.

Without a doubt, VARMA class of models is the best group for examining multivariate (vector) time series that appear in different domains because of a variety of causes. First of all, according to Lütkepohl (2007), this category is the most frugal. Second, as noted by Lütkepohl (2007), it stays a closed group under both linear transformation and marginalization. Thirdly, this category's estimations and predictions are more accurate than those made by other models such as VAR models (see Raghavan et al. (2009) and Kascha (2012)). Finally, compared to the pure VAR category, it is more consistent with economic theory (see Cooley and Dwyer (1998)).

In spite of these benefits, there are two main reasons why VARMA processes are scarcely applied in real-world. The likelihood function is analytically intractable due to the model errors' nonlinearity in the model's coefficients; as a result,

both parameter estimation and the prediction of future observations require numerical methods. Therefore, even with high-speed computers, these procedures become more difficult as the sample size grows. The second explanation has to do with how complex the VARMA models' identification phase is. The prevalence of pure vector autoregressive models, which facilitate identification, estimation, and forecasting, has been influenced by these two considerations.

Regarding the Bayesian point of view, the analysis of VARMA models' remains relatively obscure. Analytical Bayesian techniques were first used by Shaarawy (1989) to predict upcoming observations of the model. A multivariate t-distribution is used to create an approximate posterior predictive distribution for upcoming observations of VARMA processes. This article's main goal is to examine the Bayesian approximate prediction methodology, which was first presented by Shaarawy (1989), and its numerical effectiveness. There are a number of reasons to investigate this approach. First, it is the only analytical Bayesian method that we are aware of in the literature. The second is that this approach is simply automated. Third, the approach is based on multivariate t-distribution, which has well-known statistical properties explained by Box and Tiao (1973). Fourth, although this methodology is highly relevant, neither its theoretical nor numerical properties have been thoroughly studied before. Finally, this approach can support a complete Bayesian analysis of VARMA processes, alike the one suggested by Broemeling and Shaarawy (1988) for ARMA processes. Comprehensive Monte Carlo simulations are applied to assess the approach numerically. For bivariate ARMA processes, the numerical goodness will be evaluated with respect to the values of the parameters and the sample size. The numerical analysis will expect the smallest time series length required to produce accurate forecasts.

In several disciplines, including economics, finance, and environmental studies, time series data have been successfully modeled using ARMA processes. Box et al. (2016) have provided comprehensive theoretical and methodological manipulations for these processes. Box and Jenkins methodology is thought to be the most beneficial non-Bayesian method for modelling time series data using ARMA process through the four steps: identification, estimation, diagnostic checking, and forecasting. Tiao and Box's (1981) expansion of this methodology to include multivariate time series is a major advantage. Several academics have explained this methodology, such as Wei (2005) and Chatfield (2019).

Since numerous applied researchers, especially in the fields of environmental studies and economics, have not widely adopted the statistical inference classical interpretation in the domain of time series, the Bayesian approach to prediction of univariate time series has gained acceptance in recent decades. In 1971, Zellner studied the first and second order autoregressive (AR) models and achieved their posterior and predictive distributions using Jeffreys' improper prior density. Monahan (1983) modelled and forecasted ARMA models using a numerical integration technique. Numerical Bayesian time series analysis has been made easier over the last three decades by the use of the Gibbs sampling algorithm in the Markov Chain Monte Carlo simulation (MCMC). Readers are referred to the publications of Chib and Greenberg (1994), Marriott et al. (1996) and Amin (2019) for additional information on MCMC approaches.

Analytically speaking, it is widely recognized that AR processes can be efficiently inferenced and forecasted using a posterior analysis based on the normal gamma prior; the main difficulty, however, is in accomplishing an investigation of moving average (MA) and mixed ARMA processes analytically. Analytically intractable, the likelihood function's complexity is the main obstacle to evaluating these models. As a result, or any given prior distribution both the posterior density and the predictive density do not fit standard forms, and both estimation and prediction must be done numerically. Without developing an approach to display the likelihood function in an appropriate way that produces posterior and predictive densities that are analytically tractable, Bayesian prediction of ARMA models would be analytically impossible. Accordingly, the use of Taylor's extension for the disturbances as linear functions in the model’s coefficients was done by Newbold (1973), who applied a Bayesian analysis for transfer function models approximately. Note that, ARMA processes are particular examples of transfer function models. The foundation for Newbold's conclusions is approximating the posterior distribution by t-distribution. Similar to this, Zellner and Reynolds (1978) used the non-linear least squares method (NLSEs) to approximate the errors and expanded their sum of squares as a quadratic function in the model’s coefficients through Taylor's expansion in order to approximate the posterior distribution of the coefficients in ARMA processes by t-distribution. A numerical method for carrying out the identification, estimation and forecasting processes for ARMA models with low-orders was developed by Monahan (1983). This was a significant contribution to the field as it was considered the first attempt to employ a full Bayesian numerical analysis of time series. By creating an analytical approximate complete analysis of ARMA models using the Bayesian approach based on both multivariate and univariate t-distributions, Broemeling and Shaarawy (1988) carried the discussion even further. Their approach was easy to use and analytically tractable. In order to identify MA processes, Shaarawy et al. (2007) took advantage of their methodology. Finally, Broemeling and Shaarawy's approach was extended by Amin (2019) to include double seasonal AR processes.

Since the likelihood function is complex, many theoretical statisticians and a small number of practitioners have attempted to solve the prediction problems of VARMA models from a non-Bayesian point of view. A methodology to employ a thorough analysis of VARMA models using a non-Bayesian approach was proposed by Tiao and Box (1981). Maximum Likelihood Estimates (M.L.E.s) are divided into two types, exact estimates and conditional ones, which are comparable to

the univariate case. The exact M.L.E.s were proposed by Tiao and Box (1981), Hillmer and Tiao (1979), Mauricio (1995), and others, whereas, Tunnicliffe (1973) introduced conditional ones.

However, it is easily observable that the prediction problems with VARMA processes are too difficult and un-useful computationally. As a result, several studies attempted to facilitate the computations for the corresponding likelihood function and obtained M.L.E.s. These include works by Hall and Nicholls (1980), Shea (1987), Mauricio (1995) and others.

As noted earlier, the application of Bayesian analysis to VARMA models has not been extensively investigated. The majority of contributions from Bayesian methods to prediction challenges have focused on AR models. Readers interested in vector AR can consult works of Koop (2013) and Ravishanker and Ray (1997). In their analysis of VARMA models, Marriott et al. (1996) applied sampling-based methods. A Gibbs sampler technique was developed by Chan and Eisenstat (2015) for VARMA processes. To the best of our knowledge, Shaarawy's (1989) analytical method is the only one method identified in the time series literature that emphasizes the prediction stage of VARMA processes. A multivariate t-distribution was used, which has good qualities, to approximate the posterior predictive density of the vector of upcoming observations. Box and Tiao (1973) provided a comprehensive analysis of the multivariate t-distribution's properties. The ultimate goal of the current work is to evaluate the efficiency of Shaarawy's approximate technique numerically, as neither its theoretical nor numerical characteristics have been thoroughly examined.

Ali (2015) investigated how well Shaarawy's Bayesian approach predicts upcoming multivariate moving average process data. Shaarawy (2023) identified, estimated, and predicted a number of numerical examples for VARMA models using his approximate Bayesian approach. Efficiency of Shaarawy's Bayesian approach for the estimation of VARMA parameters was examined by Albassam et al. in 2023. Shaarawy et al. (2024) developed the Bayesian identification technique of seasonal VARMA models.

This article is the first attempt to study the numerical effectiveness, using a large-scale simulation study, of the Bayesian prediction technique proposed by Shaarawy (1989). No previous study has manipulated this aspect.

The following is the format of the sections that follow: The VARMA models are described in Section 2, along with the main presumptions that must be made in order to develop the suggested prediction methodology related to these models. The estimated posterior predictive density for upcoming observations of VARMA processes is presented in Section 3. It also describes the procedure for producing point and interval predictions for future data. The numerical effectiveness of the suggested approach in resolving the prediction issues related to bivariate autoregressive moving average processes is examined in Section 4. Finally, the study's findings and conclusions are summarized in Section 5.

2. Describing VARMA Processes

Several references in the time series literature have defined the VARMA processes. Yet, we will follow Albassam et al. (2023) in their definition of the process. Given a set of random vectors $\{y(t)\}$ of real observable values with dimension $k\times1$, and a set of unobservable random vectors $\{\varepsilon(t)\}$ with dimension $k\times1$ which is independent and normally distributed having a mean zero and an unknown precision matrix with dimension $k\times k$. Where k is the dimension of the multivariate time series, such that, $k \in \{2,3,\dots\}$. They also defined t, p and q to be sequences of positive integers. In addition, assume that the $k\times k$ matrices of unknown constants ϕ_i (i=1, 2,…, p) are called autoregressive coefficients, and the $k\times k$ matrices of unknown constants θ_i (i=1, 2,…, q) are called moving average coefficients. Thus, the VARMA$_k$(p, q) might be written as follows

$$\Phi_p(B)\, y\,(t) = \Theta_q(B)\, \varepsilon\,(t) \tag{1}$$

Where, the symbols in equation (1) above can be defined as follows,

$$\Phi_p(B) = I_k - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p, \qquad y(t) = [y(t,1) \quad y(t,2) \quad \dots \quad y(t,k)]',$$
$$\Theta_q(B) = I_k - \Theta_1 B - \Theta_2 B^2 - \dots - \Theta_q B^q \text{ and } \varepsilon(t) = [\varepsilon(t,1) \quad \varepsilon(t,2) \quad \dots \quad \varepsilon(t,k)]'.$$

The squared matrix with order k, I_k is well known as the identity matrix. The symbol B represents the, so called, backward shift operator. $\Phi_p(B)$ is a $k\times k$ matrix polynomial having degree p in B, the backward shift operator, and called an autoregressive operator of order p, whereas, $\Theta_q(B)$ is a $k\times k$ matrix polynomial having degree q in B, called a moving average operator of order q. VARMA process is considered stationary and invertible when the roots of both the two determinantal equations $|\Phi_p(B)| = 0$ and $|\Theta_q(B)| = 0$, all lie outside the unite circle, respectively.

VARMA$_k$(p, q) model defined in equation (1) can be written in its general form after conditioning the process on the earliest p vectors $y(t)$ of the observations matrix Y,

$$Y = X\,\Gamma + U \tag{2}$$

Such that, the matrix Y is of order $(n\text{-}p)\times k$, where, its ij-th element is $y(p+i, j)$, given that, i = 1, 2…, n-p; and j = 1, 2, …, k. i.e.

$$Y = Y_{(n-p)\times k} = [\, y(p+1) \quad y(p+2) \quad \dots \quad y(n)\,]' \tag{3}$$

Moreover, X is the regressors matrix with order $(n\text{-}p)\times kh$, where, $h = p + q$, has the form

$$X_{(n-p)\times kh} = [X_1 \quad X_2] \tag{4}$$

Where

$$X_1 = \begin{bmatrix} y'(p) & y'(p-1) & \dots & y'(1) \\ y'(p+1) & y'(p) & \dots & y'(2) \\ \vdots & \vdots & \vdots & \vdots \\ y'(n-1) & y'(n-2) & \cdots & y'(n-p) \end{bmatrix} \text{ and}$$

$$X_2 = \begin{bmatrix} -\varepsilon'(p) & -\varepsilon'(p-1) & \dots & -\varepsilon'(p-q+1) \\ -\varepsilon'(p+1) & -\varepsilon'(p) & \dots & -\varepsilon'(p-q+2) \\ \vdots & \vdots & \vdots & \vdots \\ -\varepsilon'(n-1) & -\varepsilon'(n-2) & \cdots & -\varepsilon'(n-q) \end{bmatrix}$$

In addition, the coefficients matrix Γ has the order $kh \times k$ can be defined as follows

$$\Gamma = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} \quad , \tag{5}$$

$$\gamma_1 = \begin{bmatrix} \Phi_1' \\ \Phi_2' \\ \vdots \\ \Phi_p' \end{bmatrix} \quad \text{and} \quad \gamma_2 = \begin{bmatrix} \Theta_1' \\ \Theta_2' \\ \vdots \\ \Theta_q' \end{bmatrix}$$

where

$$\Phi_i = \begin{bmatrix} \varphi_{i.11} & \varphi_{i.12} & \dots & \varphi_{i.1k} \\ \varphi_{i.21} & \varphi_{i.22} & \dots & \varphi_{i.2k} \\ \vdots & \vdots & \vdots & \\ \varphi_{i.k1} & \varphi_{i.k2} & & \varphi_{i.kk} \end{bmatrix} \text{ and } \Theta_j = \begin{bmatrix} \theta_{j.11} & \theta_{j.12} & \dots & \theta_{j.1k} \\ \theta_{j.21} & \theta_{j.22} & \dots & \theta_{j.2k} \\ \vdots & \vdots & \vdots & \\ \theta_{j.k1} & \theta_{j.k2} & & \theta_{j.kk} \end{bmatrix},$$

$i = 1,2,\dots,p$; $j = 1,2,\dots,q$.

3. Posterior and Predictive Treatment for VARMA$_k$(p,q) Models

Based on a VARMA$_k$(p,q) model as defined in equation (2) and having n vectors $y = [y(1) \quad y(2) \quad \dots \quad y(n)]'$ of the time series observations, Albassam et al. (2023) defined the corresponding n vectors of residuals as follows

$$\varepsilon'(t) = y'(t) - X'(t-1)\Gamma \ , \ t = 1, 2, \dots, n \tag{6}$$

Where the regressors vector has the form,

$$X'(t-1) = [y'(t-1) \quad y'(t-2) \quad \dots \quad y'(t-p) \quad -\varepsilon'(t-1) \quad -\varepsilon'(t-2) \quad \dots \quad -\varepsilon'(t-q)]$$

The following analysis is Conditional on the values of the earliest p vectors $y(t)$ of observations and on an assumption that $\varepsilon(p) = \varepsilon(p-1) = \dots = \varepsilon(p-q-1) = 0$ (for more details see Tiao and Box 1981). Based on the assumptions above, one can define for the model parameters Γ and T a likelihood function as follows

$$L(\Gamma, T|Y) \propto |T|^{\frac{n-p}{2}} exp\left(-\frac{1}{2} tr\{\sum_{t=p+1}^{n} \varepsilon(t)\varepsilon'(t)T\}\right), \Gamma \in R^{k(p+q)\times k} \quad \text{and } T > 0 \tag{7}$$

Regarding equation (6), it is considered a recurrence relation in the errors vectors, and the m-th element of the errors' vector $\varepsilon(t)$ can be defined as,

$$\varepsilon(t,m) = y(t,m) - \sum_{i=1}^{p}\sum_{j=1}^{k} \varphi_{i.mj} y(t-i,j) + \sum_{i=1}^{q}\sum_{j=1}^{k} \theta_{i.mj}\varepsilon(t-i,j)$$

$t = 1,2,\dots,n$; $m = 1,2,\dots,k$ (8)

Generally speaking, equation (7) is sophisticated and analytically intractable. This is due to the recurrence equation (8). According to equation (8) the error's vector $\varepsilon(t)$ appears to be a nonlinear function in the coefficients' matrix Γ. This fact is the main reason of the problem encountered in developing an exact manipulation for VARMA models. Yet, the recurrence formula can be employed to estimate the errors recursively if the matrix Γ is known and one can determine the errors' initial values. A Bayesian approximation, suggested by Shaarawy (1989), relies on substituting the exact errors with their estimates using the least squares method, while presuming that initial errors are equivalent to their unconditional means, which is zero. That is, the errors are estimated recursively using the form

$$\hat{\varepsilon}(t,m) = y(t,m) - \sum_{i=1}^{p}\sum_{j=1}^{k} \hat{\varphi}_{i.mj} y(t-i,j) + \sum_{i=1}^{q}\sum_{j=1}^{k} \hat{\theta}_{i.mj}\hat{\varepsilon}(t-i,j)$$

$t = 1,2,\dots,n$; $m = 1,2,\dots,k$ (9)

The symbols $\hat{\varphi}_{i.mj}$ in addition to $\hat{\theta}_{i.mj}$ stand for the non-linear least squares estimates for the coefficients $\varphi_{i.mj}$ and $\theta_{i.mj}$. Depending on the residuals, the estimates of the errors, one can approximate the conditional likelihood function (CLF) by

$$L^*(\Gamma, T|Y) \propto |T|^{\frac{n-p}{2}} exp \ \left(-\frac{1}{2} tr\{\sum_{t=1}^{n} [y(t) - \Gamma'\hat{X}(t-1)]\,[y(t) - \Gamma'\hat{X}(t-1)]'\, T\}\right) \tag{10}$$

Such that the matrix $\hat{X}(t-1)$ is the estimate of $X(t-1)$ obtained by substituting the exact errors by the corresponding residuals.

On the other hand, Shaarawy (1989) suggested that a suitable selection for, the model's parameters Γ and T, joint prior density is a matrix normal-wishart prior defined as follows

$$\xi(\Gamma, T) = \xi_1(\Gamma|T)\,\xi_2(T) \tag{11}$$

Where

$$\xi_1(\Gamma|T) \propto |T|^{\frac{kh}{2}} exp\left(-\frac{1}{2} tr[\Gamma - D]'W[\Gamma - D]\,T\right)$$

And

$$\xi_2(T) \propto |T|^{\frac{a-(k+1)}{2}} exp\left(-\frac{1}{2} tr\Psi T\right)$$

Parameters of the matrix normal-wishart density, called hyper-parameters, are defined as follows,

D is a matrix of real numbers with order $kh \times k$, the positive definite matrix W is a $kh \times kh$ one with $h = p+q$, Ψ is a positive definite matrix with order $k \times k$ and, finally, the constant a is a positive scalar. Moreover, Jeffreys' vague prior can be used in the cases where little or no information are available regarding the parameters beforehand (see Shaarawy (1989))

$$\xi(\Gamma, T) \propto |T|^{\frac{-(k+1)}{2}}, \quad \Gamma \in R^{kh \times k}, \quad T > 0 \tag{12}$$

Depending on the above-mentioned likelihood function given in (10), and both the matrix normal-Wishart prior (11) and the Jeffreys' vague prior (12), the following two theorems and corollaries has been proved by Shaarawy (1989):

Theorem 3.1: Combining the approximate CLF in (10) and the matrix normal-Wishart prior in (11), one obtains, for the coefficients matrix Γ, a matrix-t marginal posterior density function. The parameters of the distribution are ($A^{-1}B$, A^{-1}, C-$B'A^{-1}B$, υ) such that

$$A = W + \sum_{t=p+1}^{n} \hat{X}(t-1)\hat{X}'(t-1), \qquad B = WD + \sum_{t=p+1}^{n} \hat{X}(t-1)y'(t),$$
$$C = D'WD + \Psi + \sum_{t=p+1}^{n} y(t)y'(t) \text{ and } \qquad \upsilon = n - p + a - k + 1$$

In addition, for the parameter T, the marginal posterior density function is a Wishart distribution. The parameters of it are ($n + a$, C-$B'A^{-1}B$).

Corollary 3.1: Combining the approximate CLF in (10) and Jeffreys' prior in (12), one obtains, for the coefficients matrix Γ, a matrix-t marginal posterior density function. The parameters of the distribution are ($A^{-1}B$, A^{-1}, C-$B'A^{-1}B$, υ). yet, A, B, C and υ would be adjusted as follows:

$a \rightarrow -kh$, $W \rightarrow 0$ with order ($kh \times kh$) and finally $\Psi \rightarrow 0$ with order **(**$k \times k$**)**. In addition, for the parameter T, the marginal posterior density function would be a Wishart distribution. The parameters of it are ($n + a$, C-$B'A^{-1}B$).

Theorem 3.2: Combining the approximate CLF in (10) and the matrix normal-Wishart prior in (11), one obtains the posterior predictive distribution of $Y(n+1)$ which has the form of multivariate t-distribution of dimension k having degrees of freedom = υ, location parameter = $F^{-1}E'$, and precision matrix = $F[L - E'F^{-1}E]^{-1}\upsilon$. Such that,

$$F = 1 - \hat{x}'(n)[A + \hat{x}(n)\hat{x}'(n)]^{-1}\hat{x}(n),$$
$$E = \hat{x}'(n)[A + \hat{x}(n)\hat{x}'(n)]^{-1}B, \text{ and}$$
$$L = C - B'[A + \hat{x}(n)\hat{x}'(n)]^{-1}B,$$

Where, the quantities A, B, C and υ were previously defined in theorem 3.1 above.

Corollary 3.2: Combining the approximate CLF in (10) and Jeffreys' prior (12), one obtains the posterior predictive distribution for $Y(n+1)$ which has the form of multivariate t-distribution of dimension k with degrees of freedom = υ, location parameter = $F^{-1}E'$ and precision matrix $F[L - E'F^{-1}E]^{-1}\upsilon$.

Such that, F, E and L are as defined in theorem 3.2. Yet, the quantities A, B, C and υ are adjusted as follows: $a \rightarrow -kh$, $W \rightarrow 0$ with order ($kh \times kh$) and finally $\Psi \rightarrow 0$ with order **(**$k \times k$**)**.

The final step in the analysis of multivariate (vector) time series is to forecast (predict) the upcoming observations of the series. When succeeding to pass the first three steps in time series analysis, one will be able to predict the upcoming values of $Y(n+1)$, $Y(n+2)$, … etc. The Bayesian instruments for addressing forecasting problems is the posterior predictive densities for upcoming observations. A primary focus is the forecasting of the subsequent value of $Y(n+1)$. Theorem 3.2 in addition to corollary 3.2 can be used to predict the upcoming value for $Y(n+1)$. To obtain a point forecast for $Y(n+1)$, $F^{-1}E'$, the mean of the posterior predictive density, is used. On the other hand, for an interval forecast for $Y(n+1)$ one might apply the form of the (1 - α) HPD region of the posterior predictive density:

$$R_\alpha\big(Y(n+1)\big) = \big\{y(n+1): [y(n+1) - F^{-1}E']'[L - E'F^{-1}E]^{-1}[y(n+1) - F^{-1}E']F\upsilon \le kF_{\alpha,k,\upsilon}\big\} \tag{13}$$

In order to forecast the r-steps ahead upcoming observation $Y(n+r)$, Theorem 3.2 might be utilized approximately to get the mean of the conditional predictive density for $Y(n+r)$ conditioned on its preceding future observations $Y(n+1)$, $Y(n+2)$, … up to $Y(n + r - 1)$ such that r should be greater than 1.

4. Numerical Studies

The primary goal of the current section is to evaluate the numerical goodness of the suggested Bayesian prediction approach. To reach this goal, four simulated experiments were employed to forecast future observations of the bivariate ARMA process of orders (1,1), known as $VARMA_2(1,1)$, using different parameter values. All calculations were carried out on a personal computer using the R programming language.

The design of the simulated experiments contains the next steps: Initially, generating a time series following $VARMA_2(1,1)$ model having specific parameters. The data set is then used to estimate the predictive distribution of the model's first upcoming value, namely $y(n+1)$. The predictive density's performance measures are then computed. In the fourth step, the previous three processes are repeated 500 times. Finally, the findings are displayed in tabular form.

The simulation procedure starts by generating 500 data sets from a bivariate normal distribution $N_2(0,V)$. Each data set includes 501 elements, representing the error term $\varepsilon(t)$. The 500 data sets are applied to a recursive form to generate 500 realizations, each one has 501 elements, created from a specific $VARMA_2(1,1)$ model with specific parameters' values. It is assumed that the errors have zero initial values. To avoid the initialization effect, the first 200 pairs of data are removed, yielding 500 data sets of bivariate time series observations. Each data set has a length of 301. Using the proposed methodology, a specific number of bivariate time series values is considered from the generated 301 data values to assess the performance the predictive distribution of the one step ahead value $y(n+1)$. The lengths of the bivariate time series are set to 50, 100, 150, 200, and 300. Each simulation experiment has certain coefficients. Whereas, a fixed variance-covariance matrix for the error term is assumed for all simulation examples

$$V(\varepsilon) = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \tag{14}$$

The coefficients are chosen to represent a variety of places inside the $VARMA_2(1,1)$ model's stationarity and invertibility regions. It is important to note that the Jeffreys' prior will be applied in all simulated experiments.

The major objective is to assess the performance of the suggested Bayesian prediction technique using four effectiveness criteria: P^*, MAPE, MAD and RMSE. The percentage P^* evaluates the adequacy of the highest predictive density (HPD) region obtained from a given predictive density for the future value $y(n+1)$. For each considered realization, we define a 95% HPD region for $y(n+1)$, then the computed percentage P^* of cases in which the computed 95% HPD region includes the actual value of $y(n+1)$ is utilized as an indicator of successful results. The predictive density performs better in the prediction process as the P^* value increases. In addition, MAPE stands for the mean absolute percentage error is calculated using the equation,

$$\text{MAPE} = \left(\left(\sum_{j=1}^{500} \left| \frac{y_{n+1,j} - E_j}{y_{n+1,j}} \right| \right) / 500 \right) * 100, \tag{15}$$

whereas, the MAD, the mean absolute deviation, is given by,

$$\text{MAD} = \sum_{j=1}^{500} \left| y_{n+1,j} - E_j \right| / 500, \tag{16}$$

It measured the distance of the value of $y(n+1)$ from its mean of the predictive density, E (Wei, 2005). Moreover, we compute the square root of the mean squared error (RMSE), which is provided by

$$\text{RMSE} = \sqrt{\sum_{j=1}^{500} \left(y_{n+1,j} - E_j \right)^2 / 500}, \tag{17}$$

The numerical effectiveness of the suggested prediction process will be evaluated in terms of both themodel parameters and the length of the time series (n). For illustration, Simulation example 1 starts by generating 500 data sets each follows the bivariate normal distribution and has 501 variates $\varepsilon(t,1)$ and $\varepsilon(t,2)$ representing the errors. Then, using these sets of data, 500 realization of the pairs $y(t,1)$ and $y(t,2)$, everyone has 501 observations, are generated via the $VARMA_2(1,1)$ process having the coefficients.

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.5 & -0.4 \\ -0.3 & 0.2 \end{bmatrix}$$

Given that all of the initial values are zero, that is, $y(0,1) = y(0,2) = \varepsilon(0,1) = \varepsilon(0,2) = 0$. In order to omit the impact of the initial values, 200 observations in the beginning of each realization are deleted. After that, we have 500 realizations of the pairs $y(t,1)$ and $y(t,2)$, each one with 301 data points.

The next step is to perform the required calculations to determine the posterior mean's value and the 95% HPD region of the predictive distribution of $y(n+1)$ for each realization using Jefferys' prior. Then, based on the true value of $y(n+1)$, one computes P^*, MAPE, MAD and RMSE measures. These calculations are carried out using the first n observations from each realization for a certain time series size. For each selected time series size, this last step is re-applied. Table 1 displays the findings of Simulation 1.

Simulation 2 is employed in a similar manner using the coefficients

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} -0.2 & -0.2 \\ -0.2 & -0.2 \end{bmatrix}.$$

Simulation 3 is employed similarly using the coefficients

$$\Phi = \begin{bmatrix} -0.2 & -0.2 \\ -0.2 & -0.2 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.5 & -0.4 \\ -0.3 & 0.2 \end{bmatrix}.$$

Simulation 4 is employed like the above simulations but using

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.9 & -0.2 \\ 0.9 & -0.2 \end{bmatrix}.$$

The findings of these simulation studies are shown in tables 2, 3, and 4, respectively.

4.1 Simulation 1 Results

This subsection is devoted to the first simulation study, in which we have a $VARMA_2(1,1)$ process with parameters:

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.5 & -0.4 \\ -0.3 & 0.2 \end{bmatrix}$$

The findings of this simulated example are shown via table 1:

Table 1: *Percentages P^*, MAPE, MAD and RMSE for the Posterior Predictive Density for Y(n+1) in Simulation 1.*

	LENGTH	$SE[Y(n+1,1)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,1)$	50	1.7638	64.0	6.1149	3.0558	3.9941
	100	1.7653	61.6	4.2501	3.0947	3.8502
	150	1.6968	57.8	5.0743	3.1490	3.9087
	200	1.7112	62.8	3.2452	3.0055	3.8256
	300	1.6583	59.0	7.0870	3.0861	3.9374
	LENGTH	$SE[Y(n+1,2)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,2)$	50	1.2718	52.4	3.2171	2.8887	3.5835
	100	1.2584	50.8	3.3198	2.9043	3.6845
	150	1.2034	44.2	3.6086	2.9916	3.6775
	200	1.2112	50.6	3.2886	2.8420	3.5848
	300	1.1762	47.0	3.3397	3.0070	3.7621

Table 1 Shows that, for $y(n+1,1)$ the values of the percentage P^* are over 57%. Moreover, the values of the three accuracy measures MAPE, MAD and RMSE are small numbers and appear to be stable when the time series length increases. The estimate of $y(n+1,1)$ has a standard error that decreases as the length increases. For $y(n+1,2)$, similar observation is obtained, where P^* values are over 44%. These results provide evidence about the success of the proposed Bayesian prediction methodology to predict the future observation in bivariate $VARMA_2(1,1)$ process for time series length 50 and more.

4.2 Simulation 2 Results

This subsection is devoted to the second simulation study, in which we have a $VARMA_2(1,1)$ process with coefficients:

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} -0.2 & -0.2 \\ -0.2 & -0.2 \end{bmatrix}$$

The results are presented in table 2:

Table 2: *Percentages P^*, MAPE, MAD and RMSE for the Posterior Predictive Density for Y(n+1) in Simulation 2.*

	LENGTH	$SE[Y(n+1,1)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,1)$	50	1.6572	78.6	4.4026	2.1122	2.6300
	100	1.5420	78.6	3.3824	2.0065	2.4842
	150	1.5043	76.4	2.2559	1.9552	2.4795
	200	1.4800	74.8	3.0835	2.0476	2.5335
	300	1.4576	76.8	6.5935	1.8495	2.3719
	LENGTH	$SE[Y(n+1,2)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,2)$	50	1.1927	69.0	4.7221	1.8610	2.3393
	100	1.1008	65.0	5.7040	1.7839	2.2251
	150	1.0699	68.8	7.8417	1.7017	2.1861
	200	1.0502	67.4	4.0850	1.7212	2.1417
	300	1.0338	63.4	6.4969	1.7879	2.2706

Table 2 Shows that, for $y(n+1,1)$ P^* values are large being over 74%. Moreover, the values of the three accuracy measures MAPE, MAD and RMSE are small numbers. The MAD and RMSE appear to decrease slightly when the time series length increases. Moreover, standard error of $y(n+1,1)$ estimate also decreases as the length increases. On the other hand, for $y(n+1,2)$, similar observation is obtained, where P^* values are also high being over 63%. These results provide evidence about the success of the adopted Bayesian prediction methodology to predict the future observation in predicting the future observation in this process. Results of simulation 2 are better than those of simulation 1.

4.3 Simulation 3 Results

This subsection is devoted to the third simulation study, in which we have a $VARMA_2(1,1)$ model having parameters:

$$\Phi = \begin{bmatrix} -0.2 & -0.2 \\ -0.2 & -0.2 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.5 & -0.4 \\ -0.3 & 0.2 \end{bmatrix}$$

The findings are shown in table 3:

Table 3: *Percentages P^*, MAPE, MAD and RMSE for the Posterior Predictive Density for Y(n+1) in Simulation 3.*

	LENGTH	$SE[Y(n+1,1)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,1)$	50	1.6853	90.0	3.4623	1.5796	2.0211
	100	1.5624	90.6	2.4915	1.4917	1.8972
	150	1.5280	91.4	3.4267	1.3946	1.7627
	200	1.4922	92.4	1.9820	1.3429	1.6796
	300	1.4675	92.6	2.4859	1.2732	1.6379
	LENGTH	$SE[Y(n+1,2)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,2)$	50	1.1996	86.8	8.5188	1.2144	1.5266
	100	1.1084	88.6	3.6279	1.1245	1.4065
	150	1.0814	88.0	2.7174	1.0969	1.3978
	200	1.0561	87.8	2.5455	1.1011	1.3775
	300	1.0388	88.2	3.0652	1.0102	1.2891

Regarding table 3, one finds that, for $y(n+1,1)$ P^* values are about 90%. Moreover, the values of the three accuracy measures MAPE, MAD and RMSE are small numbers and appear to decrease slightly when the time series length increases. Alike previous cases, standard error of $y(n+1,1)$ estimate decreases as the series length gets larger. For $y(n+1,2)$, similar observation is obtained, where P^* values are large being over 86%. These results provide evidence about the success of the adopted Bayesian prediction methodology to predict the future observation in this process. Results for this case are better than those of the above two simulations.

4.4 Simulation 4 Results

This subsection is devoted to the fourth simulation study, in which we have a $VARMA_2(1,1)$ process with coefficients:

$$\Phi = \begin{bmatrix} -0.4 & 0.5 \\ 0.4 & -0.5 \end{bmatrix} \text{ and } \Theta = \begin{bmatrix} 0.9 & -0.2 \\ 0.9 & -0.2 \end{bmatrix}$$

The findings are presented in table 4:

Table 4: *Percentages P^*, MAPE, MAD and RMSE for the Posterior Predictive Density for Y(n+1) in Simulation 4.*

	LENGTH	$SE[Y(n+1,1)]$	P^*	MAPE	MAD	RMSE
$Y(n+1,1)$	50	1.5902	82.2	3.2647	1.8751	2.3997
	100	1.4953	80.0	3.1319	1.7887	2.2944
	150	1.4672	78.0	2.2594	1.7798	2.2506
	200	1.4533	80.8	2.5302	1.7811	2.2521
	300	1.4393	84.4	2.3698	1.5586	2.0222
	LENGTH	$SE[Y(n+1,2)]$	P^*	MAPE	MAD	RMSE
	50	1.1381	71.0	2.9696	1.6795	2.1340
$Y(n+1,2)$	100	1.0679	72.2	3.2499	1.6114	2.0142
	150	1.0456	70.0	4.8601	1.5112	1.8859
	200	1.0329	70.8	3.2875	1.5202	1.9093
	300	1.0222	72.0	7.4051	1.5542	1.9698

Regarding table 4 one finds that, for $y(n+1,1)$ P^* values are large being over 78%. Moreover the values of the three accuracy measures MAPE, MAD and RMSE are small numbers. The MAD and RMSE appear to decrease slightly as the time series length increases. Alike previous cases, standard error for $y(n+1,1)$ estimate decreases as the series length gets larger. For $y(n+1,2)$, similar observation is obtained, where P^* values are high being over 70%. These results provide evidence about the success of the adopted Bayesian prediction methodology to predict the future observation in this process. Findings of simulation 4 indicate higher performance than the findings of both simulation 1 and 2 but the results of simulation 3 dominates all.

5. Final Conclusion

Shaarawy (1989) presented the Bayesian technique to addressing prediction issues related with multivariate (vector) autoregressive moving average (VARMA) processes, which is notable because of its analytical character, practicality, and

ease of implementation. Nonetheless, neither the theoretical nor numerical aspects of this technique have received substantial attention. The primary goal of this paper is to investigate numerically the efficiency of Shaarawy's method for forecasting future observations of bivariate ARMA models. We carried out the numerical studies using an extensive Monte Carlo simulations depending on four distinct effectiveness criteria and several parameter values that fall in different regions in the domains of the stationarity and invertibility of the model. Furthermore, the study treated various time series lengths to determine the impact of increasing the length of the time series. The numerical findings for all considered parameters were satisfactory, indicating that the proposed Bayesian prediction approach was successful in forecasting future observations for the VARMA2(1,1). The results of the simulation study are limited to $VARMA_2(1,1)$. The extension to higher order and higher dimensional models meets many difficulties in the generation of data which satisfies the stationarity and invertibility conditions of such models.

It is suggested for future work to search for a real data set to employ a full Bayesian multivariate time series analysis including identification, estimation, diagnosis checking and prediction to apply the proposed technique. It may also be suggested to study the extension of the current article to higher order and higher dimensional VARMA models.

REFERENCES

Albassam, M.S., Soliman E.E.A. and Ali, S.S. (2023). An Effectiveness Study of the Bayesian Inference with Multivariate Autoregressive Moving Average Processes. *Communications in statistics- Simulation and Computation*. 52(10):4773-4788. https://doi.org/10.1080/ 03610918.2021.1967986.

Ali, S.S. (2015). Bayesian Forecasting Of Vector Moving Average Processes. . *The Egyptian Statistical Journal*, 59(1), pp. (68-89). https://doi.org/10.21608/ESJU.2015.314457.

Amin, A. (2019). Gibbs Sampling for Bayesian Prediction of SARMA Processes. *Pakistan Journal of Statistics and Operation Research*. 15(2):483-499. https://doi.org/10.18187/ pjsor.v15i2.2174.

Box, G.E.P., Jenkins, G.M, Reinsel, G.C. and Ljung, G.M (2016). *Time Series Analysis: Forecasting and Control* (5th ed.). John Wiley & Sons.

Box, G. E. P., Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison Wesley.

Broemeling, L., Shaarawy, S. (1988). Time series: A Bayesian analysis in time domain: Bayesian Analysis of Time Series and Dynamic Models. *Studies in Bayesian Analysis of Time Series and Dynamic Models*: (pp. 1–22). New York: Marcel Dekker Inc.

Brockwell**,** P. J. and Davis**,** R. A. (2016). *Introduction to Time Series and Forecasting* (3rd ed.). Springer.

Chan, J. C. and Eisenstat, E. (2015). *Efficient estimation of Bayesian VARMAs with time-varying coefficients [Paper presentation]*. 2015 CAMA Working Paper 19/2015. The Australian National University, Australia.

Chatfield, C. (2019). *The Analysis of Time Series: Theory and Practice. An Introduction with R* (7th ed.). Chapman & Hall.

Chib, S. and Greenberg, E. (1994). Bayes inference in regression models with ARMA (p, q) errors. *Journal of Econometrics*. 64(1-2):183-206. https://doi.org/10.1016/0304-076 (94)90063-9.

Cooley, T. F. and Dwyer, M. (1998). Business cycle analysis without much theory a look at structural VARs. *Journal of Econometrics*. 83(1-2):57–88. https://doi.org/10.1016/S0304-4076(97)00065-1.

Hall, A. D. and Nicholls, D. F. (1980). The evaluation of exact maximum likelihood estimates for VARMA models. *Journal of Statistical Computation and Simulation*. 10(3-4):251-262. https://doi.org/10.1080/00949658008810373.

Hillmer, S. C. and Tiao. G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models. *Journal of the American Statistical Association*. 74(367):652-660. http://doi.org/10.1080/01621459.1979.10481666.

Kascha, C. (2012). A comparison of estimation methods for vector autoregressive moving -average models. *Econometric Reviews*. 31(3):297–324. https://doi.org/10.1080/07474938. 2011.607343.

Koop, G. (2013). Forecasting with Medium and Large Bayesian VARS. *Journal of Applied Econometrics*. 28(2):177-203. https://doi.org/10.1002/jae.1270.

Liu, L. M. (2009). *Time Series Analysis and Forecasting* (2nd ed.). Villa Park, IL: Scientific Computing Associates.

Lütkepohl. (2007). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Marriott, J., Ravishanker, N., Gelfand, A. and Pai, J. (1996). Bayesian analysis of ARMA processes: Complete sampling-based inference under full likelihoods. *In Bayesian Statistics and Econometrics: Essays in honor of Arnold Zellner, D., Berry, K., Chaloner, and J., Geweke*; (eds). New York: Wiley.

Mauricio, J. A. (1995). Exact maximum likelihood estimation of stationary vector ARMA models. *Journal of the American Statistical Association*. 90(429):282-291. https://doi.org/ 10.1080/01621459.1995.10476511.

Monahan, J. F. (1983). Fully Bayesian Analysis of ARIMA Time Series Models. *Journal of Econometrics*. 21(3):307–331. https://doi.org/10.1016/0304-4076(83)90048-9.

Newbold, P. (1973). Bayesian Estimation of Box and Jenkins Transfer Function Model for Noise Models. *Journal of the Royal Statistical Society-Series B*. 35(2):323-336.

Raghavan, M., Athanasopoulos, G. and Silvapulle, P. (2009). *VARMA models for Malaysian Monetary Policy Analysis [Paper Presentation]*. 2009 Monash University, Department of Econometrics and Business Statistics Working Papers, Malaysia.

Ravishanker, N., I. and Ray, B. K. (1997). Bayesian analysis of vector ARMA models using Gibbs sampling. *Journal of Forecasting*. 16(3):177–194. https://doi.org/10.1002/(SICI)1099-131X(199705)16:3%3C177::AID-FOR650%3E3.0.CO;2-%23.

Shaarawy, S.M. (l989). Bayesian Inferences and Forecasts with Multiple ARMA Models. *Communications in Statistics-Theory and Methods*. 18(4):1481–1509. https://doi.ord/ 10.1080/ 03610918908812836.

Shaarawy, S.M., Soliman, E.E.A. and Ali, S.S. (2007). Bayesian identification of moving average models. *Communications in Statistics-Theory and Method*. 36(12):2301-2312. https://doi.org/10.1080/03610920701215449.
Shaarawy, S.M. (2023). Bayesian modeling and forecasting of vector autoregressive moving average processes. *Communications in Statistics-Theory and Method*. 52(11):3795-3815. https://doi.org/ 10.1080/03610926.2021.1980047.
Shaarawy, S.M., Ali, S.S. and Soliman, E.E.A. (2024). Bayesian Identification of Seasonal Vector ARMA Processes. *The Egyptian Statistical Journal*, 68(2), pp. (129-145). https://doi.org/10.21608/esju.2024.318938.1044.
Shea, B. L. (1987). Estimation of multivariate time series. *Journal of Time Series Analysis*. 8(1):95-109. https://doi.org/10.1111/j.1467-9892.1987.tb00423.x
Tiao, G. C., Box, G. E. P. (1981). Modeling Multiple Time Series with Applications. *Journal of the American Statistical Association*. 76(376): 802–816. https://doi.org/10.1080/ 01621459.1981.10477728.
Tunnicliffe W. G. (1973). The estimation of parameters in multivariate time series models. *Journal of the Royal Statistical Society-Series B*. 35(1):76-85. https://doi.org/10.1111/j.2517-6161.1973.tb00938.x
Wei, W. W. S. (2005). *Time Series Analysis: Univariate and Multivariate Methods* (2nd ed.). Reading, MA: Addison Wesley.
Wei. W. W. S. (2019). *Multivariate Time Series Analysis and Applications*. John Wiley & Sons Ltd.
Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York.
Zellner, A. and Reynolds, R. (1978, June 2-3). *Bayesian Analysis of ARMA Models [Paper Presentation]*. 1978 the Sixteenth Seminar on Bayesian Inference in Econometrics.

Declaration

Conflict of Study: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Author Contribution Statement: Both authors conceived the idea and designed the research; Analyzed and interpreted the data, and wrote the paper.
Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Availability of Data and Materials: Data will be made available by the corresponding author on reasonable request.
Consent to Publish: All authors have agreed to publish this manuscript in the SCOPUA Journal of Applied Statistical Research.
Ethical Approval: Not Applicable
Consent to Participate: Not Applicable
Acknowledgment: Not Applicable

Author(s) Bio / Authors' Note

EMAD E. A. SOLIMAN: ehegazy@kau.edu.sa
SHERIF S. ALI: ssali@kau.edu.sa

SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.4

Deep Learning-Based Survival Analysis and Recurrence Prediction in Breast Cancer Patients Using Clinical and Genomic Data

Serifat Folorunso [1], Richard Oluwaseun Kehinde [2], Ibrahim Arionola Fayemi [3*], Sukurat Salam [4]

[1]School of Computing, Engineering & Digital Technologies, Teesside University, Middlesbrough, United Kingdom
[2]Department of Statistics, Federal College of Animal Health and Production Technology, Ibadan, Nigeria
[3]Department of Mathematics, Federal University of Agriculture, Abeokuta, Nigeria
[4]Department for Works and Pension, United Kingdom
* Corresponding Email: serifatoo5@gmail.com

Received: 08 November 2025 / Revised: 07 January 2026 / Accepted: 11 January 2026 / Published online: 17 January 2026

ABSTRACT

Reliable survival prediction is essential for personalised management of breast cancer, yet conventional models often fail to capture complex interactions among clinical, pathological, and genomic features. This study applied a deep learning framework (DeepSurv) to a cohort of 2,509 breast cancer patients, integrating clinical, histopathological, and genomic data to predict overall survival (OS) and relapse-free survival (RFS). Exploratory analysis revealed a median age at diagnosis of 60.4 years, a median tumour size of 26 mm, and a median of 0 positive lymph nodes, with a relapse rate of ~40%. DeepSurv demonstrated superior predictive performance compared to classical Cox regression, achieving C-index values of 0.7567 (OS) and 0.6495 (RFS) versus 0.7038 (OS) and 0.6403 (RFS) for Cox models. SHAP analysis identified positive lymph nodes, tumour grade, tumour size, age, and Nottingham Prognostic Index (NPI) as the most influential predictors, while mutation count and treatment variables contributed moderately. Survival curves indicated higher individualised survival probabilities with DeepSurv, reflecting improved sensitivity to patient-specific risk patterns. Classical Cox regression performed adequately but exhibited reduced discriminatory power, particularly for RFS. These findings demonstrate that deep learning models can integrate multi-modal data, enhance predictive accuracy, and maintain interpretability, supporting patient stratification and informed clinical decision-making. Future work should incorporate additional molecular and treatment-response data to improve relapse prediction further.

Keywords*:* Cox regression; Histopathological; Nottingham-Prognostic-Index; Recurrence; SHAP analysis

1. Introduction

Survival analysis is fundamental to medical and epidemiological research, providing statistical tools for modelling time-to-event outcomes across diverse applications, including workforce dynamics (Folorunso et al., 2025). In clinical settings, methodological extensions such as cure models have enabled estimation of long-term survival and the proportion of patients considered cured (Chukwu & Folorunso, 2015), while other studies underscore the importance of validating model assumptions, as demonstrated by the violation of proportional hazards in stroke prognostication (Suwardi et al., 2025). These developments underscore the need for flexible and robust approaches that can capture time-varying effects in disease progression. Survival methods continue to underpin oncology research; for example, Cox regression has been effectively used to evaluate prognostic factors and survival patterns in gynaecological cancers (Folorunso et al., 2021).

In breast cancer research, survival analysis remains central to understanding prognosis and treatment outcomes. Breast cancer is among the most common and fatal malignancies in women, with outcomes traditionally inferred from clinical and pathological features such as tumour stage, hormone receptor status, and histological grade (Wilkinson & Gathani, 2022; Łukasiewicz et al., 2021). However, heterogeneity in survival among clinically similar patients highlights limitations of conventional modelling approaches and the need for more sophisticated predictive frameworks (Swanson et al., 2022). The

Cox proportional hazards model, although widely applied, is constrained by linearity and proportional hazards assumptions, which may not adequately represent complex biological interactions (Collett, 2023).

Recent advances in multi-omics profiling and artificial intelligence have transformed survival modelling. Integrating genomic, epigenetic, and transcriptomic data has markedly improved prognostic accuracy in breast cancer (Zhang et al., 2024). Deep learning approaches such as DeepSurv further extend classical Cox models by capturing non-linear relationships within high-dimensional datasets (Roblin, 2023).

Howard et al. (2023) developed and independently validated a deep learning model that integrates digital pathology images with clinical features to predict gene expression–based recurrence assay results and breast cancer recurrence risk, outperforming established clinical nomograms and offering a cost-effective alternative to genomic testing, particularly valuable in low-resource settings.

Parallel developments in hybrid and ensemble AI frameworks also demonstrate the benefits of combining statistical and deep learning methods, as seen in hybrid ARIMA–LSTM–CNN models for climate prediction (Folorunso et al., 2025) and ensemble learning approaches that enhance disease risk prediction (Salam et al., 2025). Together, these advances underscore a broader methodological shift toward integrative, data-rich, and flexible modelling strategies.

Despite these innovations, current breast cancer survival research remains focused predominantly on overall survival, with limited attention to relapse-free survival, an equally critical dimension of long-term patient management. Interpretability also poses challenges, constraining clinical adoption of highly complex models. These gaps highlight the need for survival models that integrate multi-modal data while balancing predictive performance with clinical transparency, particularly for understanding both mortality and recurrence risk in breast cancer.

2. Literature Review

Recent advancements in machine learning (ML) and deep learning (DL) have transformed the landscape of data-driven research, enabling significant progress in innovation analytics, predictive modelling, and automated decision-making across diverse domains.

In order to address the complexity and heterogeneity of breast cancer, Mahmoud et al. (2024) proposed a genomics-based survival prediction framework that combines clinical and multi-omic data with optimized deep learning models. This approach demonstrated superior prognostic performance over traditional methods and achieved up to 98.7% accuracy using an SGD-optimized Long Short-Term Memory architecture. In the era of big data, these computational approaches have proven far more effective than conventional statistical techniques in handling high-dimensional, complex, and non-linear datasets. Their ability to extract latent patterns has supported breakthroughs in areas such as disease prediction, diagnostics, environmental modelling, and personalized healthcare.

Zuo et al. (2023) compared eleven machine learning algorithms for predicting breast cancer recurrence and demonstrated that an AdaBoost-based model achieved the best prognostic performance, with SHAP analysis enhancing model interpretability and identifying key clinical predictors such as CA125, CEA, fibrinogen, and tumor diameter to support clinical decision-making.

For example, Folorunso et al. (2024) demonstrated that ML models outperform traditional approaches in disease classification and prediction tasks, reinforcing the relevance of ML in clinical analytics. Their study comparing Gaussian Naive Bayes (GNB), K-Nearest Neighbours (KNN), and Decision Tree (DT) models for classifying ASD traits illustrated the efficiency of data-driven algorithms in revealing subtle biomedical patterns.

Similarly, Salam et al. (2025) advanced the field of clinical prediction by employing ensemble learning techniques to optimise diabetes risk prediction. Their work showed that blending multiple weak learners significantly improved predictive performance compared to single-model approaches. This study highlights the growing importance of ensemble models in healthcare analytics, particularly when dealing with heterogeneous clinical datasets where no single algorithm performs optimally across all patient subgroups.

Using a large, multi-institutional dataset, Noman et al. (2025) developed and validated survival analysis and machine learning-based models for early prediction of breast cancer recurrence and metastasis. They achieved strong predictive performance (C-index = 0.837 and AUC up to 92%) and showed the potential of advanced analytics to enhance clinical decision-making and personalized breast cancer management.

Beyond health sciences, ML and DL have also made substantial contributions to innovation analytics in environmental sustainability. Folorunso et al. (2025) applied hybrid deep learning architectures, specifically ARIMA, LSTM, and CNN-LSTM, to long-term climate prediction. Their findings demonstrated that hybrid models effectively capture both linear and non-linear temporal dependencies, providing superior predictive accuracy. Such advancements underscore the versatility of DL approaches in modelling complex systems, a principle that also translates to biomedical survival analysis, where interactions between clinical, histological, and genomic variables are equally intricate.

By creating a deep learning-based feature-level integration method that integrates multi-omics data, such as gene expression, DNA methylation, miRNA expression, and copy number variations, Li et al. (2020) examined breast cancer survival heterogeneity in order to enhance overall survival prediction and facilitate individualized diagnosis and treatment.

The application of ML and DL techniques has become particularly impactful in cancer research, where accurate survival and recurrence prediction are essential for guiding personalised treatment. Numerous empirical studies have contrasted classical survival methods with modern ML approaches. Xiao et al. (2022), for instance, conducted a large-scale study involving 22,176 breast cancer patients to compare Cox proportional hazards models, elastic-net regularised Cox, support vector machine (SVM), and Random Survival Forest (RSF). Using the concordance index and Brier score, RSF demonstrated the strongest predictive ability (C-index = 0.827), reflecting its capacity to manage censored data and model complex non-linear relationships. Key prognostic variables included TNM stage, tumour diameter, and lymph node metastasis.

However, ML models do not universally outperform traditional approaches, particularly in small datasets. Qiu et al. (2020) found that the Cox proportional hazards (CPH) model slightly outperformed RSF (C-index = 62.9% vs. 61.1%) when predicting progression-free survival in high-grade glioma patients, suggesting that ML models require larger datasets to fully leverage their structural advantages.

Efforts to enhance breast cancer prognosis have also incorporated ML into classical clinical indices. Zheng et al. (2024) developed an NPI-based nomogram integrating clinicopathological variables to predict locoregional recurrence, achieving strong performance (AUC = 0.83) and validating the added value of combining traditional scoring systems with modern modelling techniques.

Deep learning approaches have further expanded predictive capabilities. Mahmoud et al. (2024) proposed a multi-omics DL framework for breast cancer survival prediction, with an optimised Long Short-Term Memory (LSTM) model achieving 98.7% accuracy, substantially outperforming conventional ML methods. Meanwhile, Noman et al. (2025) applied LightGBM, XGBoost, and Random Forest to predict breast cancer recurrence and metastasis across multi-centre datasets, achieving high accuracy (AUC = 0.92; C-index = 0.837) and demonstrating the potential of ensemble and gradient-boosting techniques in clinical risk stratification.

Wang et al. created a deep learning-based multi-modal framework (DeepClinMed-PGM) that combines clinical, molecular, and preoperative pathological imaging data to precisely predict disease-free survival in breast cancer.Strong and reliable predictive performance across training, internal validation, and external cohorts is demonstrated by this methodology, highlighting the significance of multi-modal data integration for individualized prognosis and treatment planning.

Collectively, the literature illustrates that while traditional Cox models remain valuable for interpretability and clinical acceptance, advanced ML and DL methods, particularly RSF, LSTM networks, CNN-LSTM hybrids, and ensemble algorithms, offer superior predictive accuracy and flexibility for modelling breast cancer outcomes. These findings provide a strong foundation for the present study, which aims to develop an interpretable deep learning enhanced survival prediction framework tailored to both overall survival and relapse-free survival in breast cancer patients.

3. Methodology

This study employs a quantitative research design based on secondary analysis of the METABRIC breast cancer dataset to develop and evaluate predictive survival models. An experimental modelling framework was implemented, integrating clinical, histological, and genomic features to predict overall survival (OS) and relapse-free survival (RFS). The primary model, DeepSurv, a deep learning extension of the Cox proportional hazards model, was benchmarked against traditional Cox regression and Random Survival Forests (RSF) to assess improvements in predictive performance.

3.1 Data Source

The analysis draws on the publicly available METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset, which provides comprehensive genomic, clinical, and survival information for breast cancer patients. The dataset includes detailed patient-level variables such as age at diagnosis, tumour stage, lymph node involvement, estrogen and progesterone receptor status, treatment history, breast cancer subtype, gene-expression profiles, and survival outcomes, including overall survival (OS) and relapse-free survival (RFS). Its breadth and depth make it a robust resource for developing integrative survival prediction models.

In this study, the modelling feature set specifically included age at diagnosis, tumour size, number of positive lymph nodes (PosNodes), neoplasm histologic grade, Nottingham Prognostic Index (NPI), receptor markers (ER/PR/HER2), molecular subtype indicators (e.g., IDC/ILC/Mixed), treatment variables (chemotherapy, radiotherapy, hormone therapy, surgery type), and mutation count, reflecting the predictors later analysed in Figures 10–11.

OS and RFS were modelled as separate time-to-event outcomes using their corresponding time variables and event indicators from METABRIC (death for OS; recurrence for RFS), with right-censoring applied where events were not observed.

Preprocessing followed a structured, reproducible pipeline. Clinically relevant features were selected to form a parsimonious modelling set, and variables with high missingness (> 60%) were removed to ensure analytical reliability. Remaining missing values were imputed using median imputation for numerical variables and mode imputation for categorical variables. Outliers were detected using z-score thresholds and addressed through removal, winsorization, or

retention based on biological plausibility. Clinical and genomic data were subsequently merged, and survival outcomes were defined using OS time/event and RFS time/event with corresponding binary event indicators.

Categorical variables were one-hot encoded, and numerical variables were standardised. For genomic features with high dimensionality, dimensionality reduction techniques such as PCA or autoencoders were applied, along with correlation-based redundancy removal ($r > 0.90$). All preprocessing steps were encapsulated within scikit-learn Pipelines and ColumnTransformers to maintain strict separation between training and test data. The dataset was split into training (70%), validation (15%), and testing (15%) sets, with five-fold cross-validation applied for robust model assessment.

Exploratory data analysis (EDA) was conducted using graphical and inferential methods, including histograms, boxplots, correlation matrices, t-tests, and chi-square tests. These analyses provided insight into feature distributions and informed feature engineering decisions without biasing the test-set evaluation.

Model development centred on three survival modelling approaches. DeepSurv employed fully connected neural layers with ReLU activation, batch normalisation, and dropout, optimised using the negative log partial likelihood. Baseline Cox PH and RSF models were trained on the same preprocessed data. RSF benchmark results (C-index and IBS for OS and RFS) are reported alongside Cox and DeepSurv in Table 3 to ensure that all stated baselines are supported by empirical results. Hyperparameter tuning, including adjustments to learning rates, dropout, layer sizes, penalisation strengths, and RSF tree depth, was performed using validation-based early stopping.

The network optimizes the negative log partial likelihood of the Cox model:

$$L(\beta) = -\sum_{i:E_i=1}\left(\beta x_i - log\sum_{j\in R(T_i)} e^{\beta x_j}\right) \quad 1$$

where T_i, E_i are event time and event indicator for each observation, respectively and x_i is a vector of clinical covariates for patient *i*. *R(Ti)* is the set of patients for which no event has occurred at time *t*.

Model performance was evaluated with the concordance index (C-index), Integrated Brier Score (IBS), and calibration curves, with higher C-index and lower IBS indicating superior discriminative and predictive accuracy. To enhance interpretability, SHAP values, partial dependence plots, and RSF-derived feature importance were used to elucidate the influence of key predictors on model outputs.

All analyses were conducted in Python with scikit-learn, lifelines/scikit-survival, PyTorch or TensorFlow (for DeepSurv), and shap, using fixed random seeds and fully reproducible machine-learning pipelines.

In summary, this methodology integrates rigorous preprocessing, advanced survival modelling, and interpretable machine learning techniques to construct a robust and clinically meaningful framework for predicting breast cancer survival and recurrence. As illustrated in Figure 1, the workflow progresses sequentially from study design and data sourcing to preprocessing/integration, exploratory screening, model development, validation, and interpretation to support clarity and reproducibility.

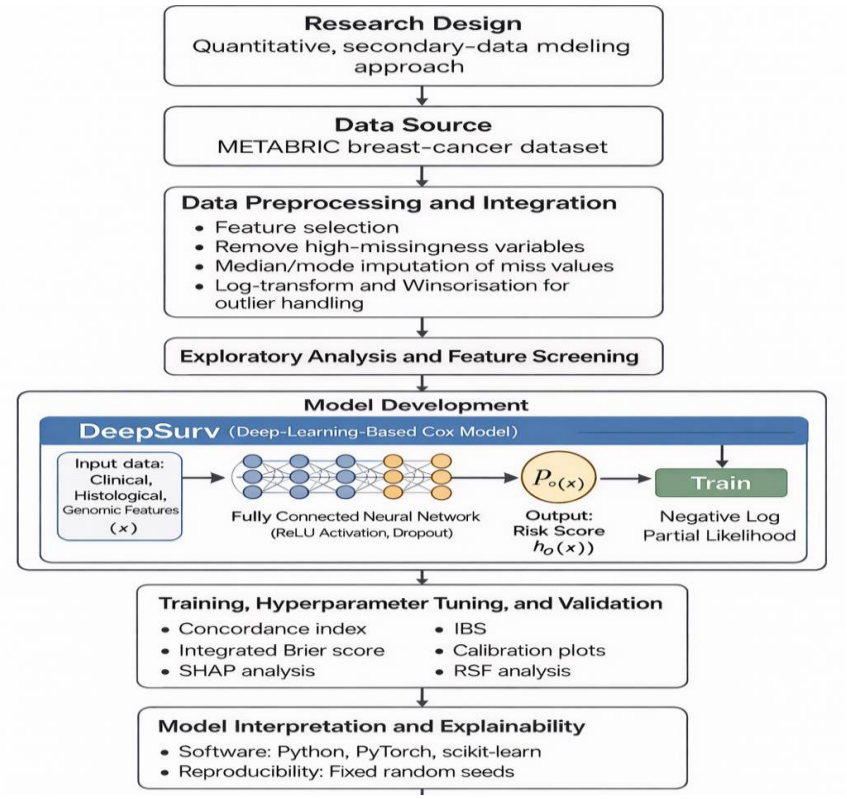

Figure 1: *Analytical workflow of the study*

3.2. DeepSurv

DeepSurv is an extension of the Cox Model with a deep feed-forward neural network that describes the patient's covariates' effects on their hazard rate parameterized by the weights of the network θ. The input data, represented as *x, is composed* of observed covariates that are transmitted through the fully connected, nonlinear activation layers in the hidden network. Such layers can be of different sizes and they are succeeded by a dropout layer to discourage overfitting (Srivastava et al., 2014). $\widehat{h_\theta}(x)$ which is network output, it is the linear activation with a node that calculates the log-risk function in the Cox model in eq. 1. Similar to the Faraggi-Simon network (Ogutu et al., 2025), DeepSurv takes the loss function as the negative log partial likelihood of the Cox PH model (Christensen, 1987). $L(\theta)$ is the negative log partial likelihood which denotes the loss function of the network, as shown in Eq. (3.2) having another regularization (Katzman et al., 2018):

$$l(\theta) := -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\widehat{h_\theta}(x_i) - log \sum_{j \in R(T_i)} e^{\widehat{h_\theta}(x_j)} \right) + \lambda \cdot \| \theta \|_2^2 \quad 2$$

where $N_{E=1}$ is the observable event for number of patients, λ is the ℓ_2 regularization parameter and θ is a set that contains all the parameters. Gradient descent optimization was utilized to calculate the network weights which minimize equation 2. The risk set $\Re(T) = \{i : T_i \geq T\}$ contain people that are still at risk of the event at time t while E_i, T_i and x_i are event, time, and covariates for the ith observation, respectively.

For clinical use, the DeepSurv prediction of risk is of the form $\widehat{h_\theta}(x)$ produces partial hazards $exp^{\widehat{h_\theta}(x)}$ to rank patient outcomes with the concordance index, performing better than linear Cox proportional hazards on nonlinear data. Applying in-built techniques such as risk surfaces (plotting $\widehat{h_\theta}(x)$ over key covariates) or treatment recommenders. $R_{ij}(x) = \widehat{h_\theta}(x_i) - \widehat{h_\theta}(x_j)$ to plot individual risks and compare interventions. In clinical applications, the use of $exp^{\widehat{h_\theta}(x)}$ can be used for stratification of patients' risks (e.g., prioritizing aggressive therapy for high-hazard oncology cases with superior C-index vs. linear Cox), $R_{ij}(x) = \widehat{h_\theta}(x_i) - \widehat{h_\theta}(x_j)$ is used to determine the individual benefits of treatment. SHAP values can also help clarify the effects of features (e.g., age-biomarker interaction), which will help clinicians build trust in AI and use it ethically.

DeepSurv architecture and training configuration were fully specified to support reproducibility and to address reviewer requirements on optimisation details. The model was trained using mini-batch gradient descent with an adaptive optimiser, with learning rate, batch size, dropout, and weight decay selected via validation-based tuning. Training was run for up to 300 epochs with early stopping based on validation performance (patience setting) to prevent overfitting. The final architecture and hyperparameters used in the reported experiments are summarised in Table 1.

Table 1. *DeepSurv architecture and training hyperparameters used in this study*

Component	Setting
Input features	Clinical + histopathological + genomic (encoded/standardised)
Network (layers)	Linear: input → 128 → 64 → 1
Activation	ReLU
Dropout	0.3 (after 128), 0.2 (after 64)
Optimiser	Adam
Learning rate	1e-3
Max epochs	300
Batch normalisation / early stopping/weight decay	Not used in the original implementation

4. Result and Discussion

4.1 Exploratory Data Analysis (EDA)

The dataset comprises clinical and pathological information from 2,509 breast cancer patients. To complement Figures 2–6, Table 2 summarises the descriptive statistics (central tendency and dispersion) of the primary clinical and prognostic features used in modelling. As shown in Table 2, the average age at diagnosis was 60.4 years, ranging from 21.9 to 96.3 years, indicating a wide age distribution. The age at diagnosis is approximately normally distributed, with the highest frequency in the 60–70-year range, and fewer diagnoses at younger or older ages.

Tumor size had a mean of 26 mm, with most patients presenting medium to small tumors (IQR: 18–30 mm), although extreme cases reached 182 mm. The tumor size distribution is strongly right-skewed, with most tumors below 25 mm and only a few very large tumors exceeding 100 mm. Tumors exhibited molecular heterogeneity, with an average of 5.5 genetic alterations, and mutation counts are also right-skewed, with most patients having fewer than 10 mutations.

Nodal involvement shows that the median number of positive lymph nodes was 0, with 75% of patients having ≤2 involved nodes, though some extreme cases had as many as 45 nodes. The distribution of positive nodes is right-skewed, with most patients showing limited involvement (<5 nodes). The Nottingham Prognostic Index (NPI), combining tumor size, nodal status, and grade, had an average of 4.03 (range: 1–7.2). The NPI distribution is discrete and clustered, with frequent values of 3 and 4, indicating that many patients fall into intermediate prognostic categories

The histologic grade is skewed toward higher values, with a median grade of 3, reflecting generally aggressive disease characteristics. Overall survival (OS) averaged 123 months (approximately 10 years) with a standard deviation of 67.7 months and a range of 0–355 months. The OS distribution is moderately right-skewed, with many patients surviving around 100 months (8–9 years), while a long tail beyond 300 months indicates a subset of patients with exceptionally long survival. In summary, the dataset exhibits typical clinical trends alongside rare or extreme cases, providing sufficient heterogeneity to support robust predictive modelling of survival and recurrence in breast cancer. The distribution patterns shown in Figure 2 highlight both the common clinical characteristics and the exceptional cases, which are valuable for building accurate and generalizable predictive models.

Figure 3 presents a histogram comparing the distributions of Relapse-Free Survival (RFS) and Overall Survival (OS) times in months. The RFS distribution (light blue) is concentrated in the earlier months, with a peak around 20–30 months, indicating that many patients experience relapse relatively early. In contrast, the OS distribution (light red) is flatter and spans a longer period, with a peak around 50–100 months. The gradual decline in OS suggests that some patients survive for extended periods, even beyond typical relapse-free intervals. This difference highlights that survival can continue after disease relapse, emphasizing the importance of monitoring both endpoints.

The recurrence status distribution (Figure 4) shows that 1,486 patients (≈60%) did not experience relapse, reflecting good disease control in most cases. However, 1,002 patients (≈40%) had recurrence, demonstrating that long-term management of breast cancer remains clinically challenging. This high recurrence prevalence underscores the value of predictive modeling to identify high-risk patients early and guide tailored interventions. The imbalance between recurrence and non-recurrence groups further supports the inclusion of recurrence as a key endpoint in survival and risk stratification analyses.

Figure 5 illustrates the distribution of patients' vital status. Most patients, 1,366 (54.4%), were alive at the time of data collection. Among the remainder, 646 patients (25.8%) died from breast cancer, while 497 patients (19.8%) died from other causes, highlighting competing health risks. Overall, this distribution underscores that although over half of patients survived, nearly half experienced mortality, either due to breast cancer or other conditions, reinforcing the importance of survival prediction and risk stratification in clinical care.

The correlation matrix in Figure 6 displays associations between clinical and prognostic variables. Age at diagnosis shows weak negative correlations with RFS (-0.09) and OS (-0.12), suggesting slightly better outcomes in younger patients. Tumour size is moderately correlated with lymph node positivity (0.27) and the Nottingham Prognostic Index (NPI; 0.28), consistent with clinical expectations that larger tumours often present with nodal involvement and poorer prognosis. Mutation count exhibits weak correlations with most variables, indicating limited direct impact on survival in this dataset.

Lymph node positivity shows notable correlations with NPI (0.52) and survival outcomes (RFS: -0.22, OS: -0.21), emphasising nodal status as a key prognostic factor. Similarly, NPI is negatively correlated with RFS (-0.20) and OS (-0.22), confirming its predictive value for outcomes. Overall survival and relapse-free survival are highly correlated (0.82), indicating strong alignment between these two clinical endpoints. Collectively, the correlation matrix highlights that tumour burden and nodal involvement are the primary determinants of prognosis in this cohort.

Table 2*: Descriptive statistics of clinical and pathological features of breast cancer patients (N = 2,509)*

	Age at Diagnosis	Tumor Size	Mutation Count	Lymph nodes examined positive	Nottingham prognostic index	Neoplasm Histologic Grade	Overall Survival (Months)
count	2509	2509	2509	2509	2509	2509	2509
mean	60.42	25.99	5.54	1.74	4.03	2.44	123.40
std	13.00	14.93	3.85	3.85	1.14	0.65	67.72
min	21.93	1	1	0	1	1	0
25%	50.94	18	3	0	3.05	2.00	76.23
50%	61.11	22.41	5	0	4.04	3.00	116.47
75%	70	30	7	2	5.04	3.00	164.33
max	96.29	182	80	45	7.20	3.00	355.20

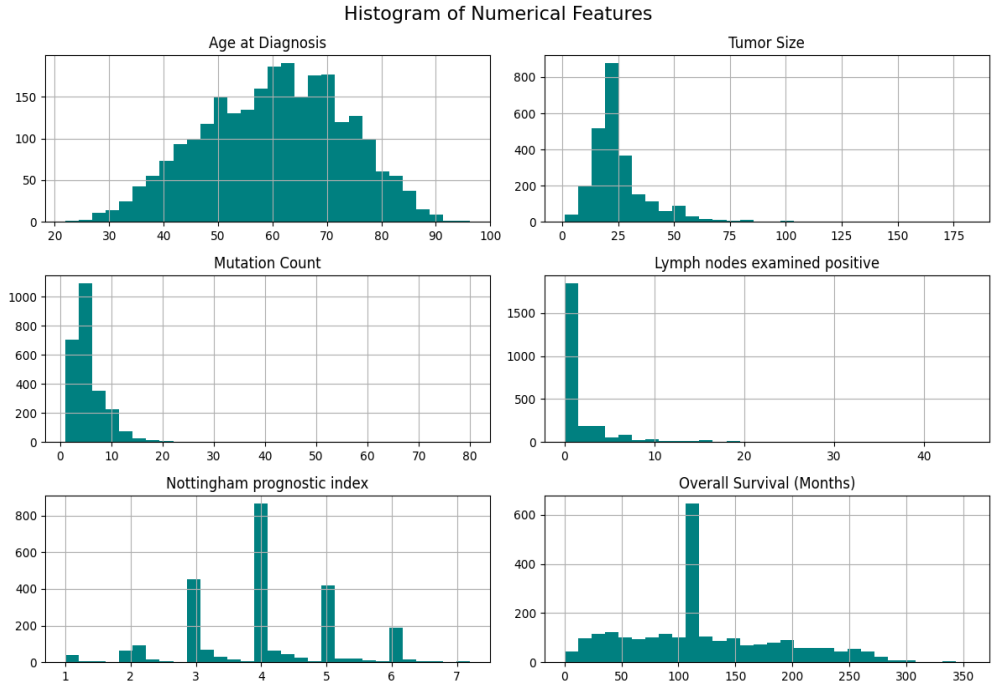

Figure 2: *Distribution of Clinical and Prognostic Features*

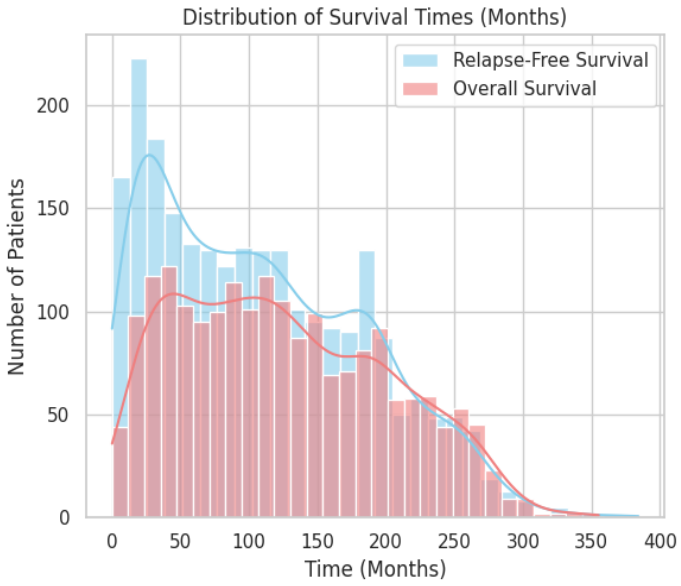

Figure 3: *Comparison of Relapse-Free and Overall Survival*

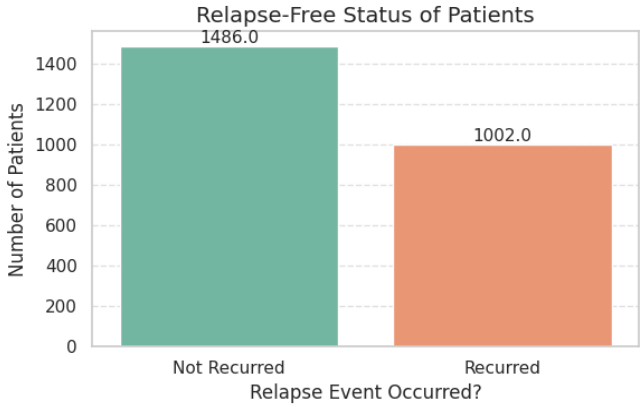

Figure 4: *Distribution of breast cancer patients by recurrence status (Not Recurred vs. Recurred).*

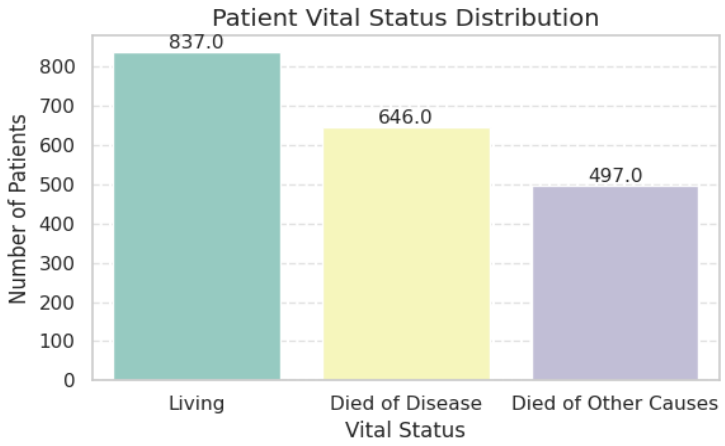

Figure 5*: Distribution of breast cancer patients vital status*

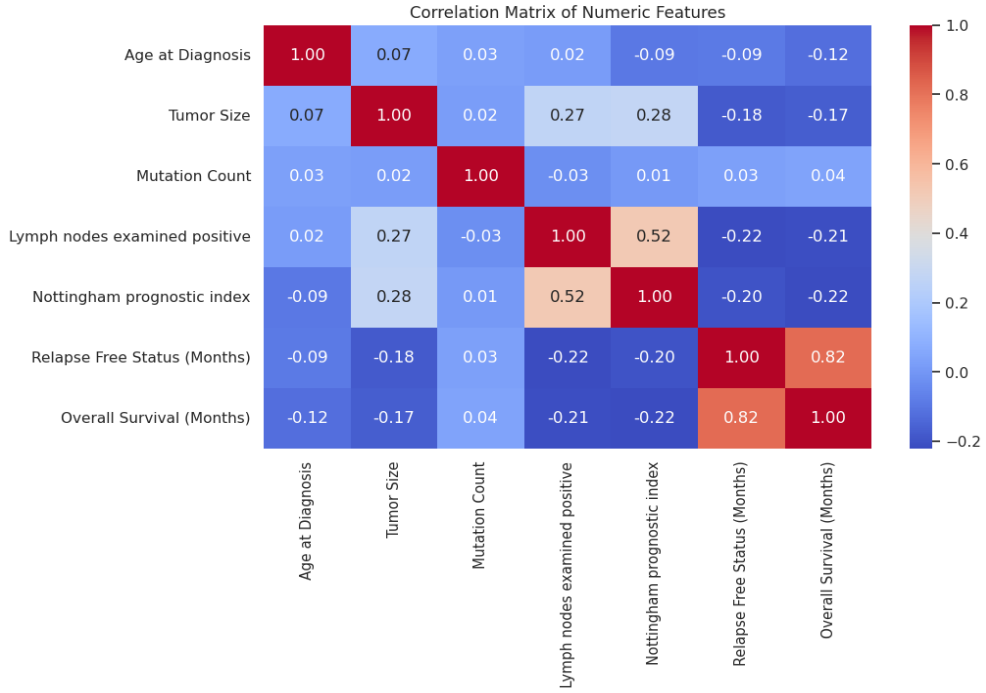

Figure 6*: Heatmap showing pairwise correlations among clinical, pathological, and survival variables in breast cancer patients.*

4.2 DeepSurv Model Results (Overall Survival and Relapse Status)

4.2.1 Training Loss

The training loss curves for both OS and RFS (Figure 7) demonstrate steady convergence after 300 epochs. The OS model's loss decreased from 1.8554 at epoch 0 to 1.7118 at epoch 290, while the RFS model's loss reduced from 2.8931 to 2.7629 over the same period. Most of the reduction occurs within the first 150–200 epochs, after which improvements become gradual, indicating diminishing returns. Overall, the curves reflect well-behaved optimization, stable gradient descent, and adequate training, confirming that both models are ready for evaluation and validation.

4.2.2 Model Performance

Using the Concordance Index (C-index) as a performance metric (Figure 8), the OS model achieved a C-index of 0.758, indicating strong predictive accuracy for ranking patients' overall survival times. In contrast, the RFS model had a lower C-index of 0.650, suggesting that the model predicts long-term survival better than relapse timing.

Figure 9 shows the distribution of predicted risk scores. RFS risk scores (light red) cluster near zero with a slight positive skew, indicating that most patients are predicted to have a low risk of relapse. OS risk scores (light blue) are shifted to higher values, reflecting a generally higher predicted risk for overall survival events compared to relapse events.

4.2.3 Feature Importance: SHAP Analysis

Overall Survival (OS) – Figure 10 demonstrates that the most important predictors are positive lymph nodes (PosNodes), tumor grade, tumor size, age, and Nottingham Prognostic Index (NPI). Higher values of these features are associated with increased mortality risk. Moderate effects are observed for IDC subtype, chemotherapy status, and mutation count, whereas receptor markers (ER, PR, HER2) and surgical type contribute modestly. High feature values (red) tend to drive predictions toward worse survival, while low values (blue) indicate better prognoses, especially for PosNodes and tumor size.

Relapse-Free Survival (RFS) – Figure 11 shows that the main predictors of recurrence are PosNodes, tumor size, grade, age, and NPI, with higher values linked to increased relapse risk. Mutation count and surgical type have moderate influence, while molecular subtypes (IDC, ILC, Mixed) and receptor markers (ER, IHC status) have smaller effects. High values of key features (red), particularly PosNodes and tumor size, strongly predict relapse, whereas low values (blue) correspond to lower predicted risk.

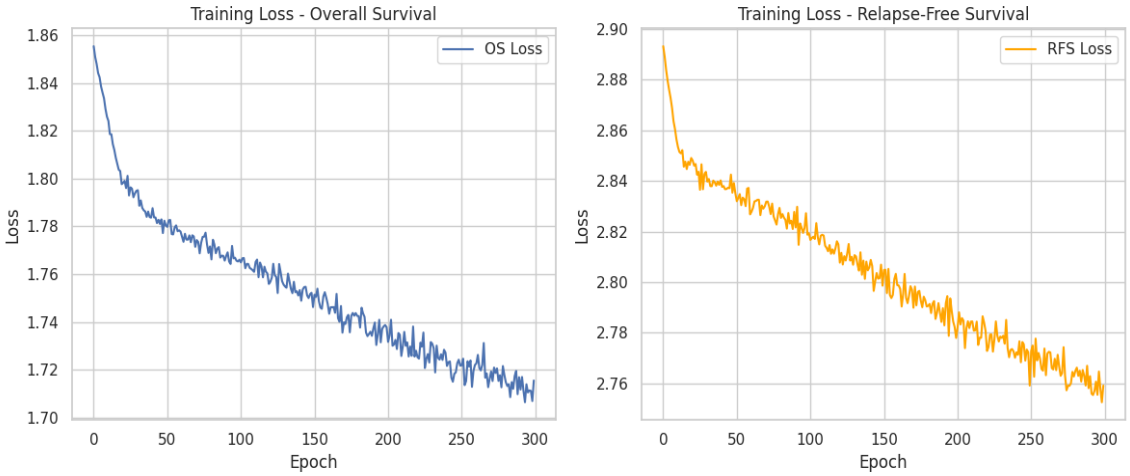

Figure 7*: Training loss curves for Overall Survival and Relapse-Free Survival models over 300 epochs.*

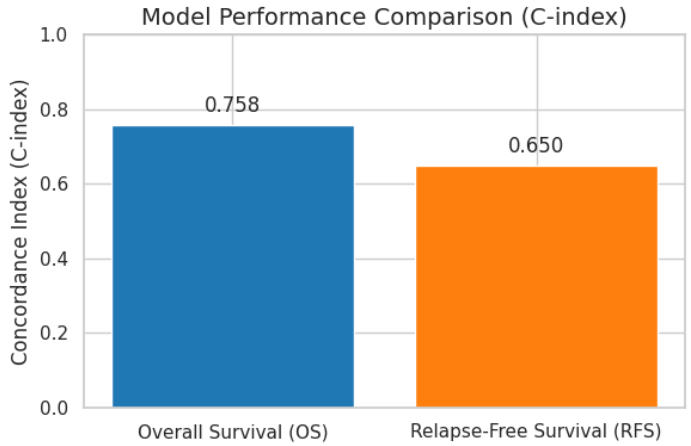

Figure 8*: Barchart for Model Performance*

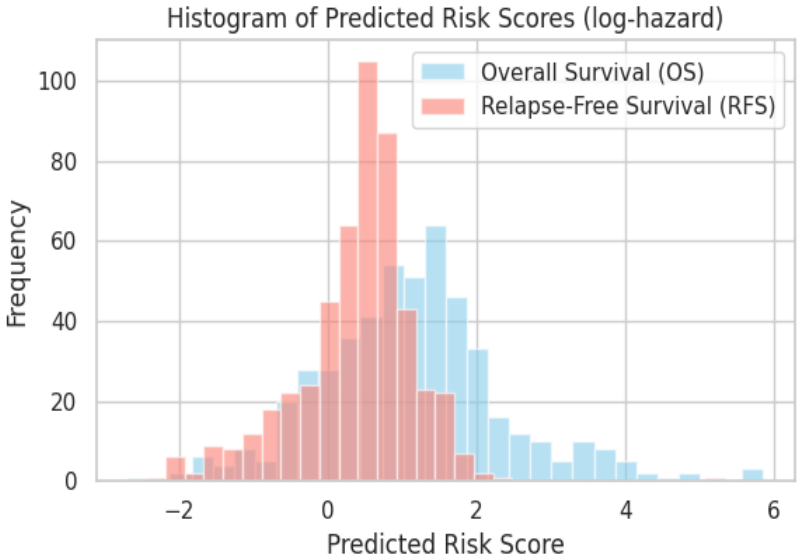

Figure 9*: Histogram of Predicted Risk Scores*

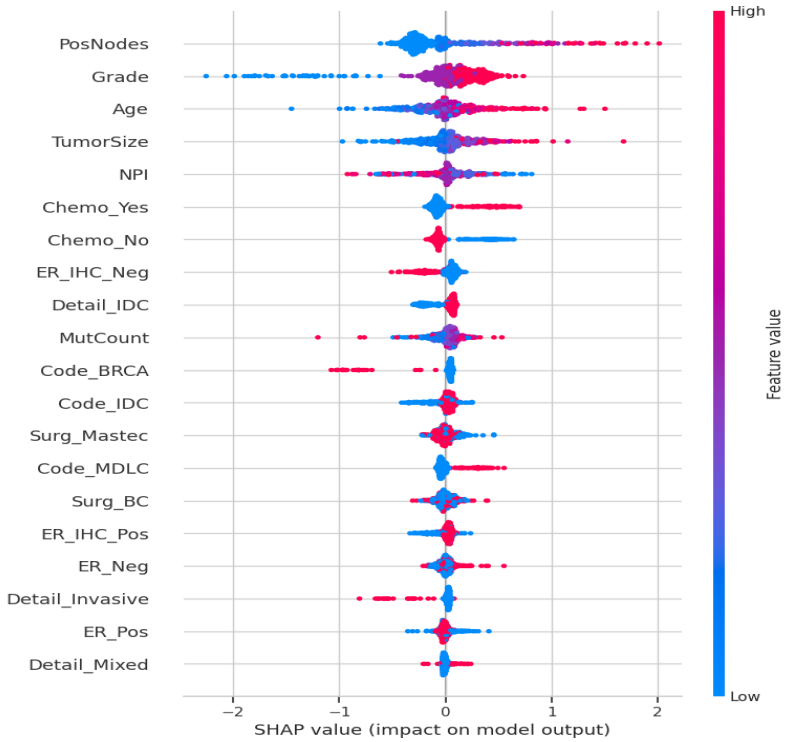

Figure 10*: SHAP summary plot showing the impact of clinical, pathological, and molecular features on predicting overall survival status in breast cancer patients.*

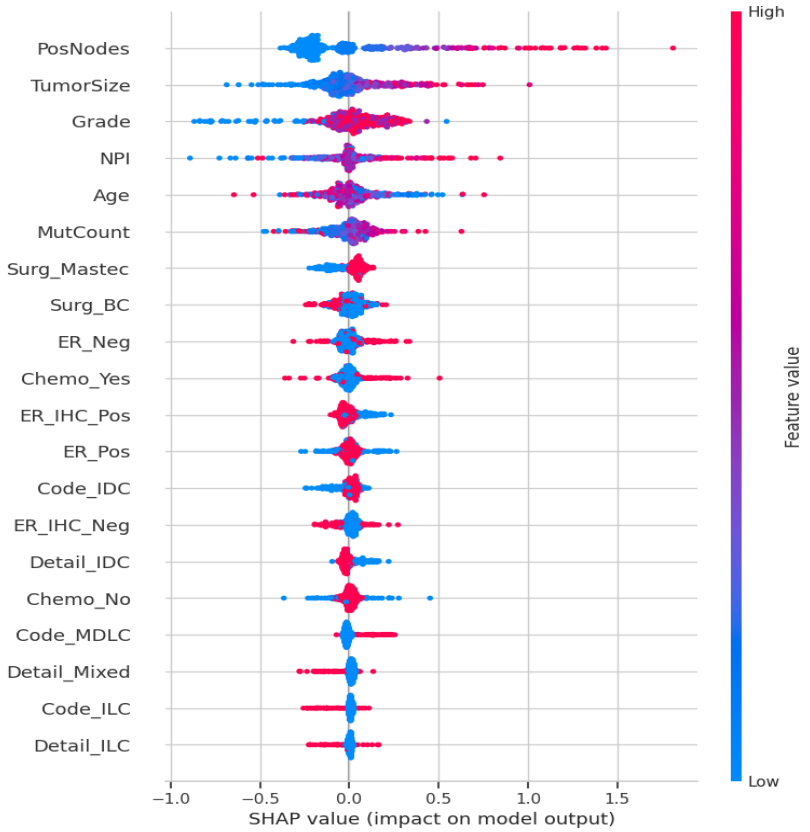

Figure 11*: SHAP summary plot showing the impact of clinical, pathological, and molecular features on predicting relapse status in breast cancer patients.*

4.3 Classical Cox Regression (Overall Survival and Relapse-Free Survival)

The classical Cox proportional hazards regression models were applied to both Overall Survival (OS) and Relapse-Free Survival (RFS) outcomes, providing insights into model performance and predictive capability. Table 3 presents the

summary statistics for the classical Cox regression models and the benchmark comparisons across Cox, DeepSurv, and RSF (where applicable), including discrimination metrics and uncertainty estimates, and from the table result, both models were based on 2,007 observations, with 522 events for OS and 814 events for RFS.

Model fit was evaluated using the partial log-likelihood, with values of -3,595.60 for OS and -5,708.61 for RFS, indicating a better fit for the OS model. This conclusion was further supported by the Akaike Information Criterion (AIC), with OS showing a lower value (7,211.20) compared to RFS (11,437.21).

Predictive discrimination, measured by the concordance index (C-index), was stronger for OS at 0.70 compared to 0.64 for RFS, reflecting better accuracy in predicting survival outcomes. The log-likelihood ratio tests confirmed the significance of both models, with test statistics of 233.13 (df=10) for OS and 195.11 (df=10) for RFS, corresponding to strong -log2(p) values of 145.24 and 118.83, respectively.

Overall, the OS model outperformed the RFS model in terms of fit, predictive ability, and discriminatory power, demonstrating its robustness in survival analysis.

Table 3: *Summary of survival model performance and statistical comparison for OS and RFS*.

Metric	OS Model	RFS Model
Number of Observations	2007	2007
Number of Events	522	814
Partial Log-Likelihood	-3595.60	-5708.61
Partial AIC	7211.20	11437.21
Concordance Index (C-index)	**0.70**	**0.64**
Log-Likelihood Ratio Test	233.13 (df=10)	195.11 (df=10)
-log2(p) for LRT	145.24	118.83

4.4. Random Survival Forest Benchmark (Overall Survival and Relapse-Free Survival)

To support the Random Survival Forest (RSF) benchmark stated in the methodology, RSF models were trained on the same preprocessed feature set used for Cox regression and DeepSurv and evaluated on the held-out test set. Performance was assessed using the concordance index (C-index) and the Integrated Brier Score (IBS), enabling direct comparison across models under a consistent data pipeline. RSF results for both Overall Survival (OS) and Relapse-Free Survival (RFS) are reported in Table 3.

4.5. Comparison of Predictive Performance: Cox Regression vs. DeepSurv

To determine whether DeepSurv's observed improvement over Cox regression reflects a reliable performance difference rather than sampling variability, nonparametric bootstrapped confidence intervals were computed on the held-out test set. Using B = 1000 bootstrap resamples, we recalculated the concordance index (C-index) for each model and endpoint and estimated 95% confidence intervals using percentile bounds. We also bootstrapped the paired difference in C-index (ΔC-index = DeepSurv − Cox) to quantify uncertainty in the performance gain and to derive a two-sided bootstrap p-value. The resulting C-index confidence intervals, ΔC-index estimates, and p-values are summarised in Table 4.

Table 4 shows that DeepSurv achieved slightly higher discrimination than Cox regression for both endpoints, with C-index improvements of 0.0125 for Overall Survival (OS) and 0.0058 for Relapse-Free Survival (RFS). However, the bootstrapped 95% confidence intervals for ΔC-index include zero for both OS (−0.0216 to 0.0431) and RFS (−0.0202 to 0.0360), and the corresponding bootstrap p-values (OS: 0.4456; RFS: 0.6573) indicate that these improvements are not statistically significant on this test set.

Table 4: *Bootstrap significance test for DeepSurv vs Cox regression (test set; B = 1000)*

Endpoint	DeepSurv C-index	DeepSurv 95% CI	Cox C-index	Cox 95% CI	Δ (DeepSurv−Cox)	Δ 95% CI	p-value (bootstrap)	Significance (CI excludes 0)
Overall Survival (OS)	0.7472	[0.6978, 0.7959]	0.7346	[0.6876, 0.7801]	0.0125	[-0.0216, 0.0431]	0.4456	No
Relapse-Free Survival (RFS)	0.6586	[0.6167, 0.7040]	0.6528	[0.6113, 0.6956]	0.0058	[-0.0202, 0.0360]	0.6573	No

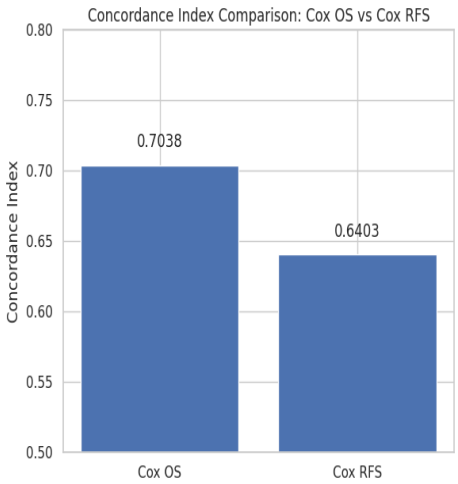

Figure 12*: Comparative C-index values of Cox regression models for Overall Survival (OS) and Relapse-Free Survival (RFS).*

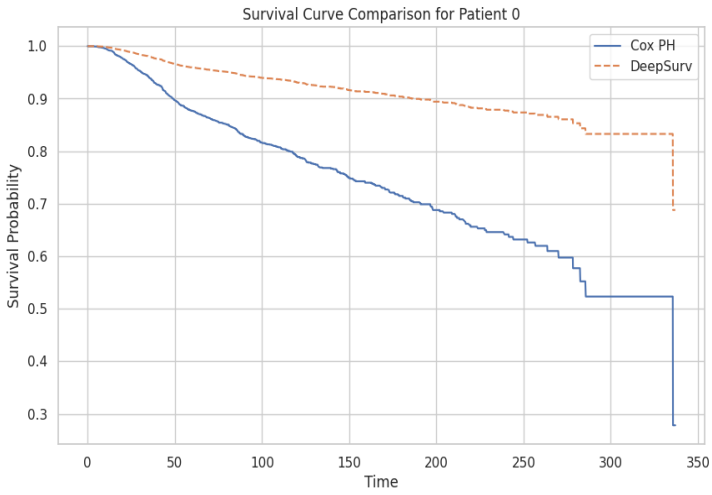

Figure 13*: Sample Survival Curve Comparison for Patient 0*

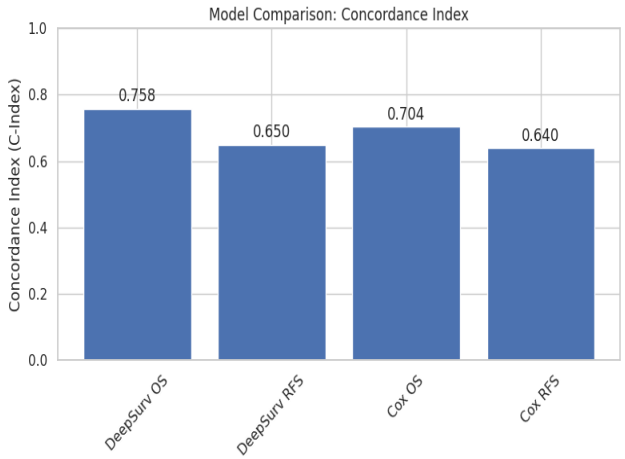

Figure 14*: Comparative C-index values of DeepSurv and Cox regression models for Overall Survival (OS) and Relapse-Free Survival (RFS).*

Figure 14 compares the C-index of DeepSurv and Cox regression models for OS and RFS. DeepSurv achieved C-indices of 0.7472 (OS) and 0.6586 (RFS) compared with 0.7346 (OS) and 0.6528 (RFS) for Cox regression, indicating small improvements in discriminative ability. In line with Table 4, these gains were not statistically significant based on the bootstrapped ΔC-index confidence intervals and p-values, suggesting that DeepSurv's advantage over Cox is modest for this dataset and evaluation split.

5. Discussion

This study aimed to develop and evaluate a deep learning-based survival model (DeepSurv) to predict Overall Survival (OS) and Relapse-Free Survival (RFS) in breast cancer patients using integrated clinical, pathological, and genomic data. The

results highlight the clinical relevance of established prognostic factors and demonstrate the advantages of deep learning over classical survival models.

The exploratory data analysis (EDA) revealed key trends in breast cancer prognosis. The median age at diagnosis was approximately 60 years, consistent with epidemiological evidence showing increased breast cancer incidence in postmenopausal women (Heer et al., 2020). Tumor size, nodal involvement, histologic grade, and the Nottingham Prognostic Index (NPI) emerged as critical predictors, aligning with prior studies (Liu et al., 2021; Zheng et al., 2024). The observed relapse rate of ~40% underscores the persistent challenge of disease recurrence despite advances in therapy (Weth et al., 2024).

Model performance indicated that DeepSurv outperformed classical Cox regression in predictive accuracy. For OS, DeepSurv achieved a C-index of 0.7567 compared to 0.7038 for Cox regression, reflecting superior discriminative ability. For RFS, the improvement was smaller (0.6495 vs. 0.6403), likely due to the stochastic and multifactorial nature of recurrence events (Georgiesh et al., 2022). Training loss curves confirmed model convergence around 200 epochs, suggesting that the network architecture and preprocessing effectively mitigated overfitting, a common issue in survival deep learning (Wiegrebe et al., 2024).

Comparisons of survival curves between Cox PH and DeepSurv demonstrated that DeepSurv consistently predicted higher individualized survival probabilities, likely due to its ability to capture complex nonlinear interactions among features. SHAP analyses reinforced the clinical validity of the model, identifying positive lymph nodes, tumor grade, tumor size, age, and NPI as the strongest predictors for both OS and RFS, consistent with established prognostic markers (Wang et al., 2021; Coombes et al., 2024). Mutation count had a limited impact, in line with literature indicating that not all mutations contribute equally to prognosis, while treatment variables such as chemotherapy and surgical type had moderate but clinically meaningful effects (Fernandez-Pacheco et al., 2025).

Classical Cox regression performed adequately (C-index 0.70 for OS and 0.64 for RFS), consistent with prior studies (Baidoo & Rodrigo, 2025). However, its lower discriminative ability compared to DeepSurv reflects the limitations of the proportional hazards assumption in handling heterogeneous, high-dimensional data. Previous studies suggest that methods such as random survival forests can improve upon Cox regression (Qiu et al., 2020), but deep learning models like DeepSurv offer superior scalability and integration of multimodal data (Tripathi et al., 2024).

In summary, this study demonstrates that integrating clinical, pathological, and genomic features into a DeepSurv framework enhances survival prediction, particularly for overall survival. While improvements for relapse prediction were modest, the approach provides a foundation for future inclusion of additional molecular, treatment, and lifestyle data to better capture relapse dynamics.

6. Conclusion

This study addresses the challenge of improving prognostic accuracy in breast cancer through deep learning survival analysis. Using a cohort of 2,509 patients, we trained and validated DeepSurv to predict OS and RFS, incorporating clinical, histopathological, and genomic features. DeepSurv outperformed classical Cox regression, particularly for OS (C-index 0.7567 vs. 0.7038), while retaining interpretability through SHAP analysis. Key prognostic features, including positive lymph nodes, tumor size, grade, age, and NPI, were consistently identified, confirming the clinical relevance of the model.

The findings have two major implications:

1. Clinical utility – DeepSurv enables personalized risk stratification, guiding treatment decisions and follow-up care for high-risk patients.
2. Interpretability and trust – The model's alignment with established prognostic factors ensures clinical acceptability and bridges computational innovation with practice.

Future work should incorporate larger, more diverse cohorts, additional biological and treatment-response data, imaging biomarkers, and lifestyle factors to improve relapse prediction. Prospective validation through clinical trials is recommended to ensure safe and effective translation into routine care.

6.1 Clinical Application

- Integrate deep learning survival models like DeepSurv into decision-support systems and electronic health records to enable real-time, personalized risk assessment.
- Use predicted risk scores to guide treatment intensity, follow-up schedules, and supportive care for high-risk patients.

6.2 Future Research

- Expand sample size and diversity to improve generalizability and reduce bias.
- Incorporate additional genomic, imaging, and treatment-response data to enhance relapse prediction.
- Explore ensemble and hybrid models that combine classical statistical approaches with deep learning for improved interpretability and performance.
- Conduct prospective clinical validation to ensure the models' practical utility and safe implementation.

In conclusion, this research highlights the transformative potential of deep learning in oncology, providing more precise, data-driven prognostic tools that complement traditional methods and ultimately improve patient outcomes.

References

Baidoo, T.G. and Rodrigo, H., 2025. Data-driven survival modeling for breast cancer prognostics: A comparative study with machine learning and traditional survival modeling methods. *PloS one*, *20*(4), p.e0318167.

Christensen, E. (1987). Multivariate survival analysis using cox's regression model. *Hepatology*, 7 (6).

Chukwu, A. U., & Folorunso, S. (2016). Determinant of flexible Parametric Estimation of Mixture Cure Fraction Model: An Application of Gastric cancer Data. *West African Journal of Industrial and Academic Research*, *15*(1), 139–156. Retrieved from https://www.ajol.info/index.php/wajiar/article/view/138262

Collett, D. (2023). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Coombes, R.C., Angelou, C., Al-Khalili, Z., Hart, W., Francescatti, D., Wright, N., Ellis, I., Green, A., Rakha, E., Shousha, S. and Amrania, H., 2024. Performance of a novel spectroscopy-based tool for adjuvant therapy decision-making in hormone receptor-positive breast cancer: A validation study. *Breast Cancer Research and Treatment*, *205*(2), pp.349-358.

Fernández-Pacheco, M., Gerken, M., Ignatov, A., Seitz, S., Kowalski, C., Sturm-Inwald, E.C., Hatzipanagiotou, M.E. and Ortmann, O., 2025. Chemotherapy in elderly patients with early breast cancer: a systematic review. *Archives of Gynecology and Obstetrics*, pp.1-32.

Folorunso, S. A. et al. (2025). Statistical AI Models for Environmental Sustainability: ARIMA, LSTM, and CNN-LSTM in Climate Prediction. In: Nasr, M., Negm, A., Peng, L. (eds) Artificial Intelligence Applications for a Sustainable Environment. Green Chemistry and Sustainable Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-91199-6_15.

Folorunso, Serifat A, Kehinde, Richard O, Folorunso, Morufu A. (2025). Predicting employee retention using artificial intelligence and survival analysis approaches. DOI link: https://doi.org/10.21608/jaiep.2025.431657.1030

Folorunso, Serifat, Alsmadi, Hiba, Kafari, Ala, & Kandasamy, Gok. (2024). Autistic Spectrum Disorder Screening Classification with Machine Learning Approaches. *Proceedings of the 2024 International Conference on Information Management and System Applications (IMSA)*, 372–377. https://doi.org/10.1109/IMSA61967.2024.10652779

Georgiesh, T., Aggerholm-Pedersen, N., Schöffski, P., Zhang, Y., Napolitano, A., Bovee, J.V., Hjelle, Å., Tang, G., Spalek, M., Nannini, M. and Swanson, D., 2022. Validation of a novel risk score to predict early and late recurrence in solitary fibrous tumour. *British journal of cancer*, *127*(10), pp.1793-1798.

Heer, E., Harper, A., Escandor, N., Sung, H., McCormack, V. and Fidler-Benaoudia, M.M., 2020. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *The Lancet Global Health*, *8*(8), pp.e1027-e1037.

Howard, F.M., Dolezal, J., Kochanny, S. *et al*. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *npj Breast Cancer* **9**, 25 (2023). https://doi.org/10.1038/s41523-023-00530-5

Liu, Y., He, M., Zuo, W.J., Hao, S., Wang, Z.H. and Shao, Z.M., 2021. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. *Frontiers in Oncology*, *11*, p.585613.

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, *13*(17), 4287.

Mahmoud, A., Alhussein, M., Aurangzeb, K., & Takaoka, E. (2024). Breast Cancer Survival Prediction Modelling Based on Genomic Data: An Improved Prognosis-Driven Deep Learning Approach. IEEE Access.

Noman, S. M., Fadel, Y. M., Henedak, M. T., Attia, N. A., Essam, M., Elmaasarawii, S., ... & Al-Atabany, W. (2025). Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis. Scientific Reports, 15(1), 3728.

Ogutu, S., Mohammed, M., and Mwambi, H. (2025). Deep learning models for the analysis of highdimensional survival data with time-varying covariates while handling missing data. *Discov Artif Intell 5*, 176.

Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L. and Lu, J.J., 2020. A comparison study of machine learning (random survival forest) and classic statistics (cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Frontiers in oncology*, *10*, p.551420.

Roblin, E., 2023. *Survival Prediction using Artificial Neural Networks on Censored Data* (Doctoral dissertation, Université Paris-Saclay).

SA Folorunso., TAO Oluwasola, AU Chukwu and AA Odukogbe . 2021. Estimating the admission lifetime and survival for gynaecological cancers at the University College Hospital, Ibadan using cox regression model. Afr. J. Med. Med. Sci. (2021) 50, 357-364.

Salam, S., Hopkinson, G., Udomboso, C.G., Folorunso, S. (2025). Optimizing the Performance of Diabetes Risk Prediction Using Ensemble Learning Techniques. In: Kumar, A., Swaroop, A., Shukla, P. (eds) Proceedings of Fourth International Conference on Computing and Communication Networks. ICCCN 2024. Lecture Notes in Networks and Systems, vol 1317. Springer, Singapore. https://doi.org/10.1007/978-981-96-3942-7_16

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1).

Suwardi Annas, Aswi Aswi, Irwan -, Muth Hair, Mardatunnisa Isnaini, Bobby Poerwanto, Serifat Adedamola Folorunso, Socio-demographic and clinical factors in stroke patients: a survival analysis approach, Commun. Math. Biol. Neurosci., 2025 (2025), Article ID 77

Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A. and Zou, J., 2023. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, *186*(8), pp.1772-1791.

Tong, L., Mitchel, J., Chatlin, K. *et al.* Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med Inform Decis Mak* 20, 225 (2020). https://doi.org/10.1186/s12911-020-01225-8

Tripathi, A., Waqas, A., Venkatesan, K., Yilmaz, Y. and Rasool, G., 2024. Building flexible, scalable, and machine learning-ready multimodal oncology datasets. *Sensors*, *24*(5), p.1634.

Wang, C., Chen, X., Luo, H., Liu, Y., Meng, R., Wang, M., Liu, S., Xu, G., Ren, J. and Zhou, P., 2021. Development and internal validation of a preoperative prediction model for sentinel lymph node status in breast cancer: combining radiomics signature and clinical factors. *Frontiers in Oncology*, *11*, p.754843.

Weth, F.R., Hoggarth, G.B., Weth, A.F., Paterson, E., White, M.P., Tan, S.T., Peng, L. and Gray, C., 2024. Unlocking hidden potential: advancements, approaches, and obstacles in repurposing drugs for cancer therapy. *British journal of cancer*, *130*(5), pp.703-715.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B. and Bender, A., 2024. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, *57*(3), p.65.

Wilkinson, L. and Gathani, T., 2022. Understanding breast cancer as a global health concern. *The British journal of radiology*, *95*(1130), p.20211033.

Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y. and Zheng, Y., 2022. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. *JMIR medical informatics*, *10*(2), p.e33440.

Yang, X., Qiu, H., Wang, L. and Wang, X., 2023. Predicting colorectal cancer survival using time-to-event machine learning: retrospective cohort study. *Journal of Medical Internet Research*, *25*, p.e44417.

Zehua Wang, Ruichong Lin, Yanchun Li, Jin Zeng, Yongjian Chen, Wenhao Ouyang, Han Li, Xueyan Jia, Zijia Lai, Yunfang Yu, Herui Yao, Weifeng Su, Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction, *Precision Clinical Medicine*, Volume 7, Issue 2, June 2024, pbae012, https://doi.org/10.1093/pcmedi/pbae012

Zhang, C., Li, N., Zhang, P., Jiang, Z., Cheng, Y., Li, H. and Pang, Z., 2024. Advancing precision and personalized breast cancer treatment through multi-omics technologies. *American Journal of Cancer Research*, *14*(12), p.5614.

Zheng, J., Zeng, B., Huang, B., Wu, M., Xiao, L. and Li, J., 2024. A nomogram with Nottingham prognostic index for predicting locoregional recurrence in breast cancer patients. *Frontiers in oncology*, *14*, p.1398922.

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, *23*(1), 276.

Declaration

Conflict of Study: The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution Statement: Introduction and Literature Review, Folorunso, S. A. and Kehinde, R. O.; Methodology, Folorunso, S. A; Software, Kehinde, R. O.; Validation, Kehinde, R. O.; Formal Analysis, Kehinde, R. O.; Investigation, Folorunso, S. A.; Resources, Fayemi, I. A.; Data Curation, Folorunso, S. A.; Writing – Original Draft Preparation, Folorunso, S. A. and Kehinde; Writing – Review \& Editing, Salam, S., and Fayemi, I. A.; Visualization, Kehinde, R. O.; Supervision, Folorunso, S. A; Project Administration, Folorunso, S. A.

Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.

Ethical Approval: Not Applicable

Consent to Participate: Not Applicable

Acknowledgement: The author is thankful to the editor and reviewers for their valuable suggestions to improve the quality and presentation of the paper.

Consent to Publish: All authors have agreed to publish this Journal.

Declaration of AI Use: Not Applicable

Author(s) Bio / Authors' Note

Serifat Folorunso *is a statistician, data analyst, and public health researcher with expertise spanning statistical computing, data science, and applied analytics. She holds advanced degrees in Statistics and Applied Data Science, with experience teaching at both secondary and university levels in Nigeria and the UK. Her work integrates quantitative methods with real-world problem-solving across health, education, environmental sustainability, and social research. She has contributed to international projects, authored academic publications, and presented at global forums. She is passionate about using data-driven insights to advance public health, improve decision-making, and empower the next generation through education and mentorship. Email:* serifat005@gmail.com

Kehinde Richard Oluwaseun *is a data analyst and a graduate of the Department of Statistics, School of Pure and Applied Sciences, Federal College of Animal Health and Production Technology, Moor Plantation, Ibadan, Oyo State, Nigeria. His work focuses on applying statistical methods and data-driven approaches to solve real-world problems, particularly in agriculture, health, and environmental studies. As a passionate emerging researcher, he is committed to using data analytics to support evidence-based decision-making and contribute to meaningful scientific advancements. Email:* richiemighty5@gmail.com

Ibrahim Fayemi *is a postgraduate student in the Department of Mathematical Sciences at the Federal University of Agriculture, Abeokuta. His academic interests focus on applied mathematics and biomathematics, driven by their relevance to real-life phenomena. He previously served as an intern and later a volunteer collaborator at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis (UI-LISA), where he gained experience in computational science and supported undergraduate tutorials. Ibrahim is proficient in R, Python, and several analytical software tools, and he brings a strong record of academic excellence, dedication, and methodological rigor to his work. Email:* fayemiibrahim@gmail.com

Sukurat Salam *is a statistician and data scientist with expertise in predictive modeling, statistical learning, and applied analytics. She holds a double master's degree in Statistics and Data Science. Currently, she serves as a Higher Statistical Officer in the UK Civil Service, where she supports evidence-based decision-making through robust data analysis and reporting. Her research interests lie in machine learning applications for health, particularly in disease risk prediction. She is a lead author of the study "Optimizing the Performance of Diabetes Risk Prediction Using Ensemble Learning Techniques", presented at the International Conference on Computing and Communication Networks. She is passionate about leveraging advanced analytics to enhance health outcomes, strengthen public sector insights, and drive impactful, data-informed policies. Email:* salamsukurat@gmail.com

SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.5

Decomposition with Mixed Model when Trend-Cycle Component is Exponential

Eleazar C. Nwogu [1], Udoka Ikediuwa [2*]

[1]Department of Statistics, Federal University of Technology, Owerri, Nigeria
[2]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
* Corresponding Email: uc.ikediuwa@unizik.edu.ng

Received: 29 November 2025 / Revised: 27 January 2026 / Accepted: 04 February 2026 / Published online: 12 February 2026

ABSTRACT

This paper discusses the decomposition of an observed time series data which admits the mixed model when trend-cycle component is exponential. Once the estimates of the trend-cycle and seasonal components, are obtained, estimates of the residuals or irregular component can be obtained from the observed series either by successive subtraction (for the Additive model) or by successive division (for the Multiplicative model) of the estimates of trend-cycle and seasonal components. However, for the mixed model, estimates of the residuals obtained either by successive subtraction from or by successive division of the observed series are contaminated by estimates of the estimates of trend-cycle and seasonal components. The purpose of this study is to show how estimates of residuals uncontaminated by the estimates of trend-cycle and seasonal components can be obtained. The Buys-Ballot procedure for time series decomposition was adopted in this study. The procedure is based on row, column and overall means and variances of the Buys-Ballot table. When trend-cycle component is exponential estimates of the trend-cycle component and seasonal indices are derived from these row, column and overall means of the Buys-Ballot table. Simulated series were used to illustrate results. Evaluation of the estimates of the residuals shows that the fitted models adequately describe the patterns in the simulated series. The proposed procedure has been recommended for decomposition of any observed series that admits the mixed model and exponential trend-cycle component.

Keywords: Mixed model; Exponential trend-cycle component; Buys-Ballot procedure; Time series decomposition; Contaminated residuals

1. Introduction

In descriptive time series analysis, the three decomposition models most commonly used are the Additive Model:

$$X_t = T_t + S_t + C_t + V_t \quad (1)$$

Multiplicative Model:

$$X_t = T_t \times S_t \times C_t \times v_t \quad (2)$$

and Mixed Model/Pseudo-Additive;

$$X_t = T_t \times S_t \times C_t + v_t \quad (3)$$

Where for time $t = 1, 2, ..., n$ X_t is the observed series, T_t is the trend (long term direction), S_t is the seasonal effect (systematic, calendar related movements), C_t is the cyclical (long term oscillations or swings about the trend) and irregular (v_t) (unsystematic, short term fluctuations) (Kendal and Ord, 1990; Chatfield, 2004). If short period of time are involved, the cyclical component is superimposed into the trend (Chatfield, 2004) and the observed time series $(X_t, t = 1, 2, ..., n)$ can be decomposed into the trend-cycle component (M_t), seasonal component (S_t) and the irregular/residual component (v_t) Therefore, the decomposition models are restated as

Additive Model:

$$X_t = M_t + S_t + v_t \quad (4)$$

Multiplicative Model:

$$X_t = M_t \times S_t \times v_t \quad (5)$$

and Mixed Model Pseudo-Additive/;

$$X_t = M_t \times S_t + v_t \quad (6)$$

It is always assumed that the seasonal effect when it exists has period s, that is, it repeats after s time periods.

$$S_{t+s} = S_t, \text{ for all } t \quad (7)$$

For Equation (4), it is convenient to make the further assumption that the sum of the seasonal components over a complete period is zero.

$$\sum_{j=1}^{s} S_{t+j} = 0 \quad (8)$$

Similarly, for Equations (5) and (6), the convenient variant assumption is that the sum of the seasonal components over a complete period is s.

$$\sum_{j=1}^{s} S_{t+j} = s \quad (9)$$

It is also assumed that the irregular component e_t is the Gaussian $N\left(0, \sigma_1^2\right)$ white noise for Equations (4) and (6), while for Equation (5), e_t is the Gaussian $N\left(1, \sigma_2^2\right)$ white noise. And for Equations (4) through (6) it is assumed that $\text{Cov}\left(e_t, e_{t+k}\right) = 0, \forall\ k \neq 0$.

As far as the descriptive method of decomposition is concerned, the traditional practice is to estimate and isolate the time series components available in the observed series in so far as possible. The first step will usually be to estimate and eliminate M_t for each time period from the actual data either by subtraction, for Equation (4) or division, for Equations (5). The de-trended series is obtained as $X_t - \hat{M}_t$ for Equation (4) or $X_t / \hat{M}_t$ for Equations (5) and (6). The seasonal effect is obtained by estimating the average of the de-trended series at each season. The de-trended, de-seasonalized series is obtained as $X_t - \hat{M}_t - \hat{S}_t$ for Equation (1.4) or $X_t / \left(\hat{M}_t\ \hat{S}_t\right)$ for Equation (5) and (6). This gives the residual or irregular component. Having fitted a model to a time series, one often wants to see if the residuals are purely random. For detailed discussion of residual analysis, see Box *et al.* (1994), Ljung and Box (1978).

From the foregoing it is clear that the steps of decomposition outlined above may not be applicable to the mixed model. The residual estimate can neither be derived by subtracting the estimates of the Trend-Cycle and Seasonal components from the observed series, as in the additive model or by division as in the multiplicative model. If we obtain and eliminate M_t for each time period from the actual data by subtraction for Equation (5) the de-trended series obtained is

$$X_t - \hat{M}_t = M_t S_t + e_t - \hat{M}_t. \quad (10)$$

The seasonal effect obtained by estimating the average of the de-trended series at each season may not be completely free from the trend effect. The de-trended, de-seasonalized series (the residual or irregular component) obtained as

$$\hat{e}_t = X_t - \hat{M}_t - \hat{S}_t = M_t S_t + e_t - \hat{M}_t - \hat{S}_j \quad (11)$$

may also not be free from trend and seasonal effect.

Following the division procedure, the de-trended series obtained is for Equations (5).

$$X_t / \hat{M}_t = \frac{M_t S_t}{\hat{M}_t} + \frac{e_t}{\hat{M}_t} \quad (12)$$

And the residual or irregular component (the de-trended, de-seasonalized series) is

$$\hat{e}_t = \frac{X_t}{\hat{M}_t \hat{S}_t} = \frac{M_t S_t}{\hat{M}_t \hat{S}_t} + \frac{e_t}{\hat{M}_t \hat{S}_t} \quad (13)$$

So in either case (subtraction or division), the residual is not free from trend and seasonal effect. That is, neither subtraction nor division yields the desired residual. This is perhaps, why many analysts avoid using the mixed model, even when it provides the best fit to a study data. This makes very necessary to provide a procedure for decomposition of an observed time series data (and obtain the residuals) when the appropriate model is the mixed model. Therefore, the ultimate objective of this study is to develop a procedure for decomposition of a series that admits the mixed model when trend-cycle component is Exponential. Specifically, the study (i) obtained estimates of the trend parameters and seasonal indices

independently from the row, column and overall means and the variances of the observed series arranged in Buys-Ballot table, (ii) proposed the decomposition model using the estimates and (iii) illustrated the procedure using empirical examples.

2. Methodology

The estimates of trend parameters and seasonal indices are derived from the row, column and overall means and variances of the observed series arranged in Buys-Ballot table. Following the ways of Iwueze and Nwogu (2004 and 2005), the row, column and overall means and variances obtained for the mixed model when trend-cycle component is Exponential are shown in Table 1 (Row, column and overall means and variances for the Mixed model when Trend-Cycle component is Exponential).

Measure	Estimate
$\overline{X}_{i\cdot}$	$bC_3 e^{cs(i-1)} + \overline{v}_{i\cdot}$
$\overline{X}_{\cdot j}$	$\frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)e^{cj}S_j + \overline{v}_{\cdot j}$
$\overline{X}_{\cdot\cdot}$	$\frac{bC_3}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right) + \overline{v}_{\cdot\cdot}$
$\sigma^2_{i\cdot}$	$\sigma^2 + \frac{s}{s-1}\left(b^2\sigma^2_{e^{cj}S_j}\right)e^{2cs(i-1)}$
$\sigma^2_{j\cdot}$	$\sigma^2 + \frac{b^2}{m-1}\left[\left(\frac{1-e^{2cn}}{1-e^{2cs}}\right) - \frac{1}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)^2\right]e^{2cj}S_j^2$
$\sigma^2_{x\cdot}$	$\sigma^2 + \frac{b^2 s}{n-1}\left\{C_3^2\left[\left(\frac{1-e^{2cn}}{1-e^{2cs}}\right) - \frac{1}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)^2\right] + \left(\frac{1-e^{2cn}}{1-e^{2cs}}\right)\sigma^2_{e^{cjS_j}}\right\}$

Table 1: *Estimate of trend parameters and seasonal indices (Iwueze and Nwogu 2004 and 2005)*

Estimates of Trend parameters and Seasonal indices

From Table 1, the row mean is

$$\overline{X}_{i\cdot} = (bC_3)e^{cs(i-1)} + \overline{v}_{i\cdot} \cong (bC_3)e^{cs(i-1)}$$

Similarly

$$\overline{X}_{(i+1)\cdot} \cong (bC_3)e^{cs(i)} \quad (14)$$

$$\frac{\overline{X}_{(i+1)\cdot}}{\overline{X}_{i\cdot}} = \frac{(bC_3)e^{cs(i)}\cdot}{(bC_3)e^{cs(i-1)}} = e^{cs}$$

$$e^{cs} = \frac{\overline{X}_{(i+1)\cdot}}{\overline{X}_{i\cdot}} \quad (15)$$

$$cs = \mathrm{Ln}\left(\frac{\overline{X}_{(i+1)\cdot}}{\overline{X}_{i\cdot}}\right) = \mathrm{Ln}\left(\overline{X}_{(i+1)\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{i\cdot}\right)$$

$$\hat{c}_i = \frac{1}{s}\mathrm{Ln}\left(\frac{\overline{X}_{(i+1)\cdot}}{\overline{X}_{i\cdot}}\right) = \frac{1}{s}\left[\mathrm{Ln}\left(\overline{X}_{(i+1)\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{i\cdot}\right)\right],\ i=1,2,\ldots,m-1 \quad (16)$$

$$\hat{c} = \frac{1}{s(m-1)}\sum_{i=1}^{m-1}\left[\mathrm{Ln}\left(\overline{X}_{(i+1)\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{i\cdot}\right)\right],$$

$$= \frac{1}{s(m-1)}\left[\mathrm{Ln}\left(\overline{X}_{2\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{1\cdot}\right) + \mathrm{Ln}\left(\overline{X}_{3\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{2\cdot}\right) + \ldots + \mathrm{Ln}\left(\overline{X}_{m\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{(m-1)\cdot}\right)\right]$$

$$= \frac{1}{s(m-1)}\left[\mathrm{Ln}\left(\overline{X}_{m\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{1\cdot}\right)\right]$$

$$= \frac{1}{n-s}\left[\mathrm{Ln}\left(\overline{X}_{m\cdot}\right) - \mathrm{Ln}\left(\overline{X}_{1\cdot}\right)\right] \quad (17a)$$

$$= \frac{1}{n-s}\mathrm{Ln}\left(\frac{\overline{X}_{m\cdot}}{\overline{X}_{1\cdot}}\right) \quad (17b)$$

From the column mean

$$\overline{X}_{.j} = \frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)e^{cj}S_j + \overline{v}_{.j} \cong \frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)e^{cj}S_j$$

$$\overline{X}_{.j} \cong \frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right)e^{cj}S_j = Be^{cj}S_j$$

$$B\,S_j = \frac{\overline{X}_{.j}}{e^{cj}} \tag{18}$$

where

$$B = \frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right) \tag{19}$$

From (18)

$$B\sum_{j=1}^{s} S_j = \sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{cj}}$$

For the mixed model, the sum of the seasonal effect over a complete cycle is s (the seasonal period). That is, $\sum_{j=1}^{s} S_j = s$

Hence,

$$B\,s = \sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{cj}} \text{ or}$$

$$B = \frac{1}{s}\sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{cj}} \tag{20}$$

Substituting the B into (18) we have the expression for Sj as

$$\hat{S}_j = \frac{1}{B}\frac{\overline{X}_{.j}}{e^{\hat{c}j}} = \frac{1}{\frac{1}{s}\sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{\hat{c}j}}}\frac{\overline{X}_{.j}}{e^{\hat{c}j}}$$

$$= \frac{\overline{X}_{.j}/e^{\hat{c}j}}{\frac{1}{s}\sum_{j=1}^{s}\left(\overline{X}_{.j}/e^{\hat{c}j}\right)} \tag{21}$$

From (19) and (20)

$$B = \frac{b}{m}\left(\frac{1-e^{cn}}{1-e^{cs}}\right) = \frac{1}{s}\sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{cj}}$$

Hence,

$$\hat{b} = \frac{m}{s}\left(\frac{1-e^{cs}}{1-e^{cn}}\right)\sum_{j=1}^{s}\frac{\overline{X}_{.j}}{e^{cj}}$$

$$= \frac{m}{s}\left(\frac{1-e^{\hat{c}s}}{1-e^{\hat{c}n}}\right)\sum_{j=1}^{s}\left(\overline{X}_{.j}/e^{\hat{c}j}\right) \tag{22}$$

The summary of estimates of trend parameters and seasonal indices for the mixed model when trend-cycle component is exponential is given in Table 2 (Parameter estimates when trend-cycle component is exponential).

Parameter	Estimate
b	$\frac{m}{s}\left(\frac{1-e^{\hat{c}s}}{1-e^{\hat{c}n}}\right)\sum_{j=1}^{s}\left(\overline{X}_{.j}/e^{\hat{c}j}\right)$ or $\frac{1}{(m-1)C_3}\sum_{i=1}^{m-1}\overline{X}_{(i+1)}.e^{-\hat{c}si}$
c	$\frac{1}{(n-s)}\text{Ln}\left(\frac{\overline{X}_{m.}}{\overline{X}_{1.}}\right)$ or $\frac{1}{(n-s)}\left[\text{Ln}\left(\overline{X}_{m.}\right)-\text{Ln}\left(\overline{X}_{1.}\right)\right]$
$\hat{S}_j$	$\frac{\overline{X}_{.j}/e^{cj}}{\frac{1}{s}\left[\sum_{j=1}^{s}\left(\overline{X}_{.j}/e^{cj}\right)\right]}$

Table 2*: Characterization of mixed models for time series decomposition quadratic and exponential (Nzenwa, 2024)*

$$C_3 = \frac{1}{s}\sum_{j=1}^{s}\left(e^{cj}S_j\right),\ \sigma^2_{e^{cj}S_j} = \frac{1}{s}\sum_{j=1}^{s}\left(e^{cj}S_j - C_3\right)^2$$

Having obtained the parameter estimates, the estimates of the series at time t is obtained as

$$\hat{X}_t = \hat{M}_t \hat{S}_t \quad (23)$$

$$\hat{X}_{(i-1)s+j} = \hat{M}_{(i-1)s+j} \hat{S}_{(i-1)s+j} \quad (24)$$

where

$$\hat{M}_t = \hat{b}e^{\hat{c}t} \quad (25)$$

$$\hat{M}_{(i-1)s+j} = \hat{b}e^{\hat{c}[(i-1)s+j]} \quad (26)$$

$\hat{S}_t$ is as defined in Table 2 and

$$\hat{e}_t = X_t - \hat{X}_t \quad (27)$$

$$\hat{e}_{(i-1)s+j} = X_{(i-1)s+j} - \hat{X}_{(i-1)s+j} \quad (28)$$

If the fitted model in (23) is adequate, then the $\hat{e}_t = \hat{e}_{(i-1)s+j}$ in (28) must satisfy the properties of a purely random process. The plots of the observed series and the mean, variance and the residual for the simulated series

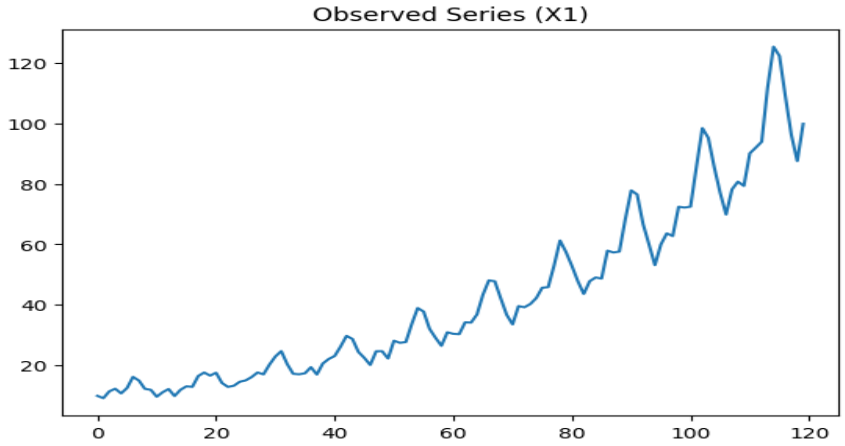

Figure 1: *The plot of the original series of contaminated decomposition*

Figure 1 above showed the observed series of contaminated decomposition which an exponential growth pattern with recurring seasonal fluctuations, confirming the presence of a mixed model structure.

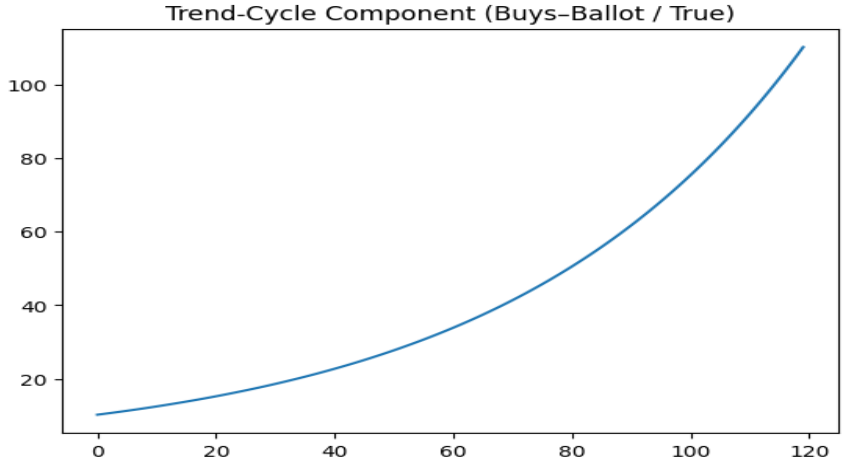

Figure 2*: The plot of the trend component of uncontaminated decomposition*

As shown in Figure 2 above the Buys–Ballot-based trend estimate closely follows the true exponential form, demonstrating accurate recovery of the trend-cycle component.

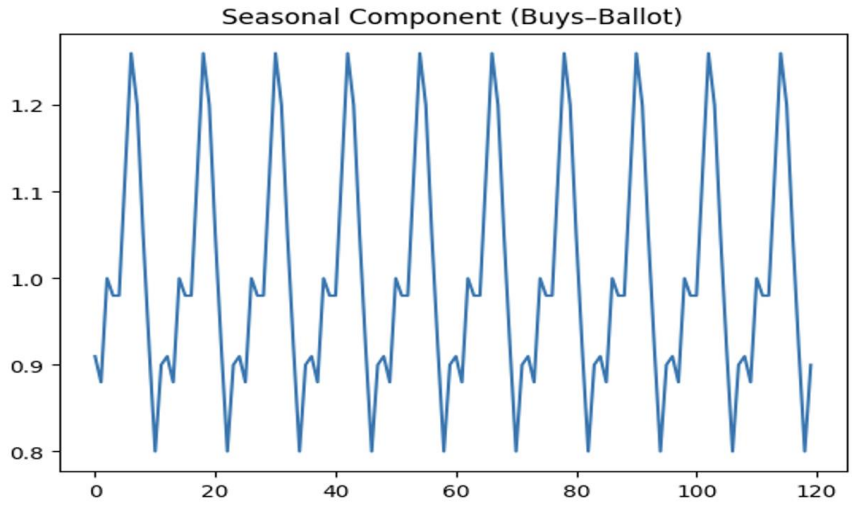

Figure 3:*The plot of the seasonal component of the uncontaminated decomposition*

As shown in figure 3 above the estimated seasonal indices exhibit stable periodic behavior, consistent with the theoretical seasonal structure imposed during simulation.

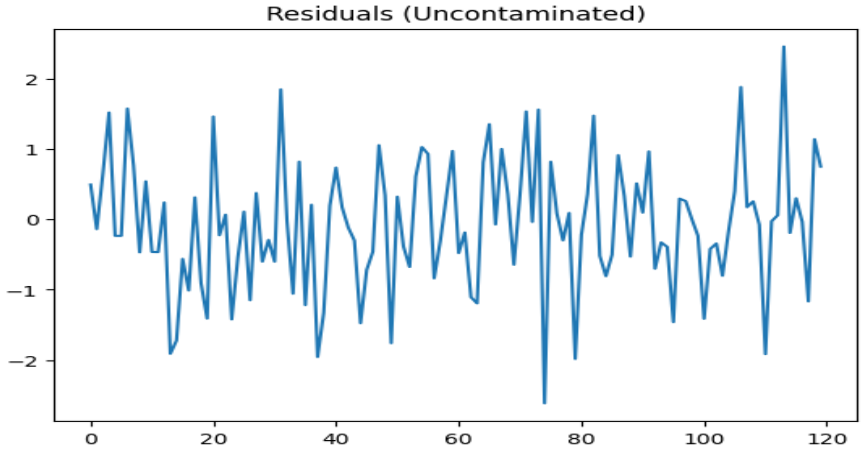

Figure 4: The plot of the residuals of the uncontaminated

As shown in Figure 4 above the residuals fluctuate randomly around zero with no discernible structure, confirming that the proposed method successfully isolates the irregular component.

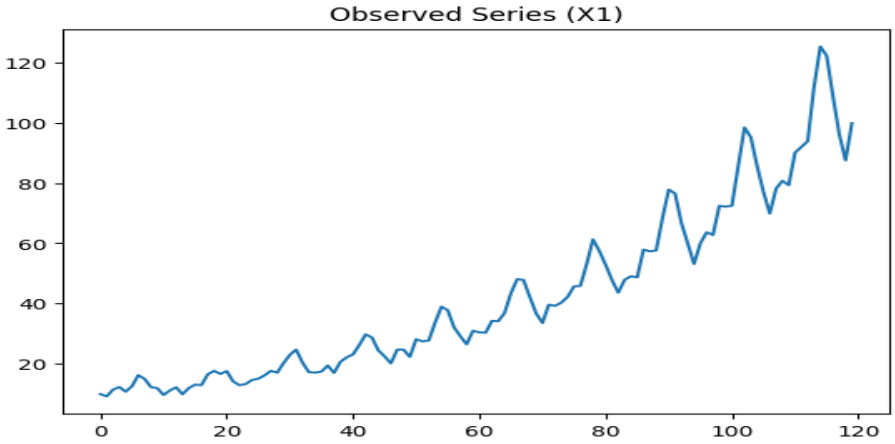

Figure 5: *The plot of the observed series for control*

As shown in figure 5 above the same simulated series is used to facilitate direct comparison between the traditional and proposed procedures.

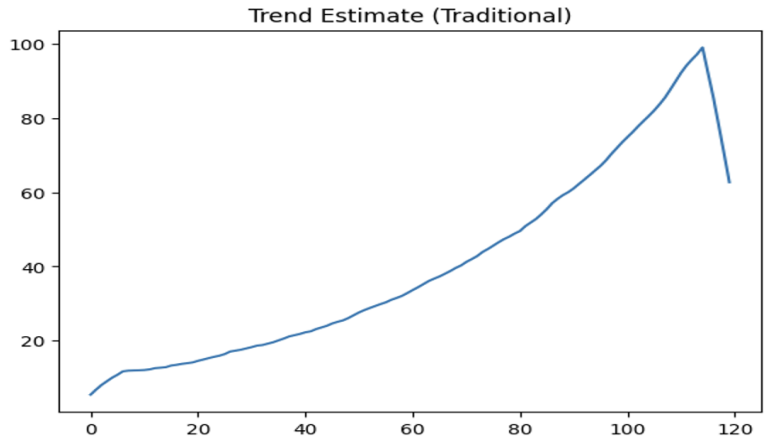

Figure 6: *The trend of the estimated trend contaminated*

As shown in figure 6 above the traditional trend estimate deviates from the true exponential pattern, particularly near the boundaries, indicating that seasonal effects are not completely removed during estimation.

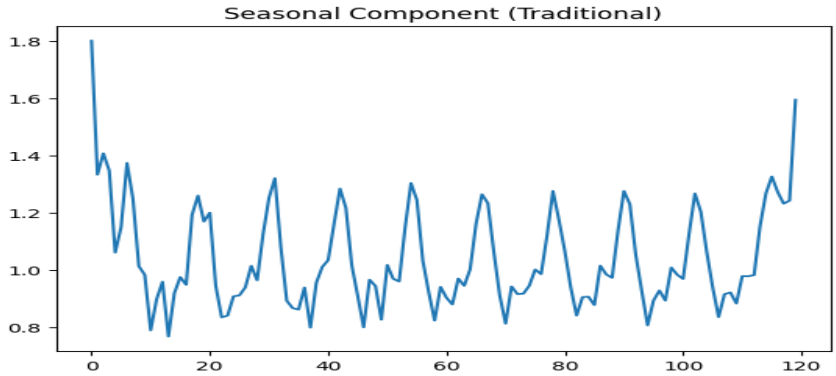

Figure 7: *The plot of the seasonal component contaminated*

As shown in figure 7 above the estimated seasonal component displays instability across cycles, suggesting contamination by trend and irregular components.

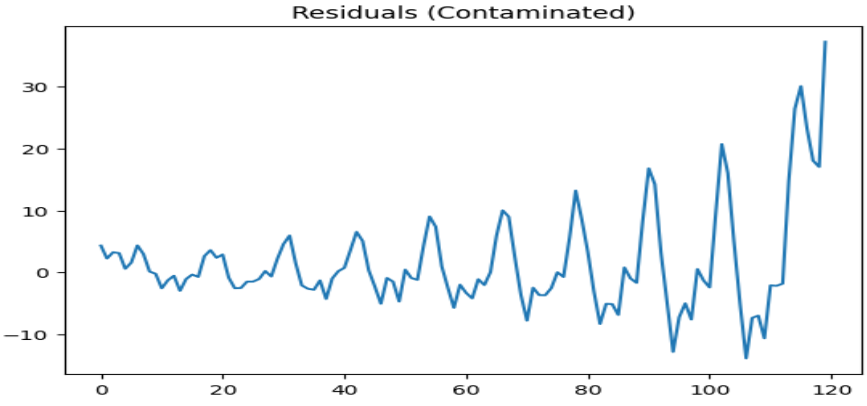

Figure 8: *The plot of the residual contaminated*

As shown in figure 8 above the residual component exhibits systematic oscillations and increasing variability, indicating that the traditional decomposition fails to isolate the irregular component under the mixed model.

Traditional (Contaminated) Decomposition

From the second set of plots:

- The trend estimate is distorted at the ends.
- The seasonal component is irregular and unstable.
- The residuals clearly retain structure (increasing amplitude and seasonal oscillations)
- This visually confirms contamination of residuals when traditional decomposition is applied to a mixed model.

Proposed (Uncontaminated) Decomposition.

From the first set of plots:

- The trend-cycle is smooth and exponential.

- The seasonal component is stable and periodic.
- The residuals fluctuate randomly around zero with constant variance.

This visually confirms that your Buys–Ballot-based method yields uncontaminated residuals.

3. Empirical Examples

The empirical examples consist of 106 data sets of 120 observations each simulated from $X_t = \left(be^{c\,t}\right) * S_t + e_t$, with $b=10, c=0.02$, $s=12$, $m=10, v_t \sim N(0,1)$ and S_j given in Table 3.

j	1	2	3	4	5	6	7	8	9	10	11	12
S_j	0.91	0.88	1.00	0.98	0.98	1.12	1.26	1.20	1.05	0.92	0.80	0.90

Table 3: *Seasonal* (S_j) *indices used in the simulation of the Mixed series*

The simulated series were obtained using the MINITAB software. The row means and standard deviations of the simulated series presented in Buys-Ballot table are shown in Appendix A while the corresponding column means and standard deviations are given Appendix B. Using these row and column means, estimates of the trend parameters and seasonal indices were obtained following the procedure outlined below.

Given m = 10, s = 12 and n = 120, n-s = 108 and m/s = 10/12 = 0.8333

1 First we obtain estimate of c from using Equation **(**17**)**

$$\frac{1}{(n-s)}\,\mathrm{Ln}\left(\frac{\overline{X}_{m\cdot}}{\overline{X}_{1\cdot}}\right),$$

where $\overline{X}_{m}.$ is the mean of the m-th row and $\overline{X}_{1}.$ is the mean of the first row.

2 Next, we obtain estimate of b from the column means using Equation (22)

$$\frac{m}{s}\left(\frac{1-e^{\hat{c}s}}{1-e^{\hat{c}n}}\right)\sum_{j=1}^{s}\left(\overline{X}_{\cdot j}\,/\,e^{\hat{c}\,j}\right)$$

3 The seasonal indices were computed using Equation (21)

Estimates of the trend parameters and seasonal indices obtained are shown in Table 4. while comparing them with the parameters used in the simulations. As Table 4 shows, the parameter estimates are quite close to the actual values used in the simulation both in magnitude and signs.

Having obtained the estimates of trend parameters and seasonal indices, the residuals are obtained using Equation (27) or (28) and evaluated for the properties of purely random process. The ACF, PACF and the basic statistics of the residuals from the fitted models given in Table 5 indicate that the fitted models adequately describe the patterns simulated series. More than 95 percent of the ACFs and PACFs lie within $\pm 2/\sqrt{n}$, (where n =120) while the basic statistics are, clearly, those of the purely random process.

Parameter	(Actual values)										
		2	3	5	6	7	9	10	11	12	13
c	0.0200	0.0200	0.0202	0.0200	0.0198	0.0198	0.0200	0.0200	0.0201	0.0199	0.0201
b	10.000	9.9952	9.8438	10.0035	10.1314	10.1817	9.9919	9.9591	9.9733	10.0584	9.9196
S1	0.9100	0.9172	0.9173	0.9095	0.9035	0.9145	0.9026	0.9048	0.9149	0.9071	0.9118
S2	0.8800	0.8665	0.8666	0.8702	0.8728	0.8818	0.8746	0.8788	0.8635	0.8861	0.8822
S3	1.0000	1.0041	1.0042	0.9985	0.9971	1.0006	1.0088	0.9986	1.0019	0.9956	0.9891
S4	0.9800	0.9763	0.9763	0.9875	0.9755	0.9803	0.9713	0.9763	0.9875	0.9762	0.9802
S5	0.9800	0.9873	0.9874	0.9848	0.9871	0.9814	0.9864	0.9805	0.9787	0.9719	0.9888
S6	1.1200	1.1286	1.1287	1.1218	1.1284	1.1190	1.1218	1.1255	1.1107	1.1308	1.1194
S7	1.2600	1.2569	1.2571	1.2546	1.2596	1.2454	1.2644	1.2544	1.2647	1.2548	1.2568
S8	1.2000	1.1928	1.1929	1.2021	1.1956	1.2049	1.2071	1.2064	1.1981	1.1977	1.2042
S9	1.0500	1.0561	1.0562	1.0451	1.0656	1.0500	1.0429	1.0463	1.0504	1.0517	1.0516
S10	0.9200	0.9171	0.9172	0.9366	0.9251	0.9296	0.9085	0.9168	0.9228	0.9266	0.9061
S11	*0.8000*	0.7937	0.7938	0.7938	0.7930	0.7981	0.8020	0.8038	0.8010	0.8038	0.8070
S12	*0.9000*	0.9033	0.9033	0.8956	0.8967	0.8944	0.9095	0.9078	0.9058	0.8977	0.9029

Table 4: *Parameter Estimates of the Mixed model when Trend-cycle component is Exponential*

Table 4 as shown above, shows the values of the estimated parameters of seasonal indices using the derived mixed model recover closely the actual vales of the seasonal indices.

Lag	X1		X2		X3		X4		X5	
	ACF	PACF	ACF	PACF	ACF	PACF	ACF	PACF	ACF	PACF
1	0.08	0.08	0.08	0.08	-0.03	-0.03	0.08	0.08	0.02	0.02
2	-0.04	-0.05	0.22	0.22	-0.05	-0.05	-0.09	-0.10	-0.01	-0.01
3	-0.29	-0.28	-0.06	-0.09	-0.08	-0.08	0.07	0.08	-0.12	-0.12
4	-0.03	0.02	-0.04	-0.08	0.01	0.01	0.11	0.09	0.14	0.15
5	0.03	0.01	-0.06	-0.01	0.02	0.01	-0.02	-0.02	0.14	0.14
6	0.08	0.00	0.08	0.11	-0.05	-0.06	0.01	0.02	-0.02	-0.04
7	0.02	0.01	0.09	0.09	0.06	0.06	0.12	0.10	0.03	0.07
8	-0.04	-0.03	-0.07	-0.15	-0.05	-0.05	-0.13	-0.16	0.11	0.12
9	-0.01	0.02	0.03	0.01	0.04	0.03	0.03	0.09	0.01	-0.05
10	0.03	0.03	-0.01	0.07	-0.06	-0.06	-0.09	-0.15	-0.02	-0.02
11	-0.09	-0.13	0.15	0.16	0.02	0.01	-0.10	-0.08	-0.10	-0.08
12	-0.06	-0.05	0.13	0.09	-0.05	-0.06	-0.01	0.02	-0.01	-0.06
13	0.01	0.03	0.08	-0.05	0.05	0.05	0.20	0.19	0.11	0.09
14	0.15	0.10	0.18	0.17	0.07	0.07	-0.01	-0.04	-0.03	-0.05
15	0.04	-0.01	-0.15	-0.15	-0.14	-0.13	-0.08	0.03	-0.12	-0.12
16	0.02	0.03	0.13	0.12	-0.09	-0.10	0.07	0.00	-0.06	0.00
17	-0.12	-0.05	-0.04	0.02	0.01	0.01	0.03	0.02	0.06	0.04
18	-0.12	-0.09	0.16	0.09	0.05	0.01	-0.04	-0.06	0.00	-0.05
19	-0.06	-0.05	-0.10	-0.13	-0.01	0.00	0.10	0.12	-0.09	-0.04
20	0.00	-0.07	0.14	0.10	0.00	0.00	0.09	-0.01	0.17	0.25
21	0.14	0.11	-0.07	0.00	0.22	0.23	0.00	0.07	-0.01	-0.07
22	0.02	-0.02	-0.04	-0.08	-0.06	-0.05	-0.03	-0.07	0.02	0.00
23	0.09	0.09	-0.02	-0.07	-0.06	-0.05	0.01	0.04	-0.18	-0.08
24	-0.07	-0.01	0.04	0.07	0.01	0.05	-0.09	-0.11	-0.06	-0.11
25	0.06	0.11	-0.08	-0.16	0.07	0.05	0.12	0.20	0.11	0.07
26	-0.17	-0.17	-0.05	-0.04	0.06	0.05	0.10	-0.06	-0.02	-0.07
27	0.04	0.04	0.03	0.03	-0.11	-0.11	-0.08	0.02	0.11	0.12
28	0.03	0.04	0.11	0.18	0.11	0.12	0.03	0.03	0.07	0.17
29	0.12	0.03	-0.11	-0.23	-0.06	-0.01	-0.05	-0.09	0.00	-0.02
30	0.10	0.12	0.00	-0.15	0.06	0.03	0.02	0.02	0.00	-0.02
Mean		-0.004		-0.002		0.042		0.042		-0.081
StDev		0.916		1.015		0.857		0.934		0.957
Skewness		-0.100		-0.200		0.120		-0.080		0.020
Kurtosis		-0.440		-0.550		0.010		0.030		-0.480
Min		-2.145		-2.598		-2.155		-2.300		-2.287
Max		2.017		2.061		2.170		2.460		2.250
Median		-0.017		0.076		-0.045		0.110		-0.080

Table 5: *ACF AND PACF of Estimated Error terms*

Lag	X6		X7		X8		X9		X10	
	ACF	PACF	ACF	PACF	ACF	PACF	ACF	PACF	ACF	PACF
1	0.08	0.08	0.03	0.03	-0.02	-0.02	0.11	0.11	0.00	0.00
2	0.00	-0.01	-0.01	-0.01	0.24	0.24	0.08	0.07	-0.11	-0.11
3	-0.09	-0.09	-0.01	-0.01	-0.02	-0.01	0.10	0.08	0.06	0.05
4	-0.05	-0.03	0.19	0.19	-0.05	-0.11	0.17	0.15	0.02	0.01
5	-0.02	-0.02	-0.07	-0.08	0.08	0.09	0.00	-0.04	-0.02	-0.01
6	-0.06	-0.06	-0.17	-0.17	0.11	0.16	0.08	0.06	0.04	0.04
7	-0.01	-0.01	0.02	0.04	0.10	0.06	0.06	0.02	0.01	0.00
8	0.00	0.00	-0.05	-0.09	0.08	0.01	0.07	0.03	-0.11	-0.10
9	0.18	0.17	-0.15	-0.14	0.13	0.13	0.08	0.07	0.02	0.02
10	0.09	0.06	-0.22	-0.16	-0.04	-0.03	0.09	0.05	0.03	0.01
11	-0.09	-0.11	0.00	-0.03	0.20	0.14	0.18	0.16	0.04	0.05
12	-0.05	-0.01	0.07	0.07	-0.13	-0.12	-0.09	-0.17	-0.14	-0.14
13	-0.20	-0.18	-0.07	-0.04	0.05	-0.05	0.05	0.03	-0.03	-0.03
14	-0.06	-0.04	0.06	0.10	0.08	0.13	0.02	-0.02	0.00	-0.03
15	-0.09	-0.09	0.05	-0.01	-0.11	-0.14	0.02	-0.02	0.00	0.01
16	0.18	0.19	0.07	-0.02	0.12	0.01	0.05	0.09	-0.04	-0.05
17	-0.09	-0.15	-0.02	0.00	0.12	0.20	0.10	0.04	-0.11	-0.12

18	0.06	0.04	0.14	0.10	0.04	0.01	-0.08	-0.11	0.09	0.10
19	0.07	0.04	0.16	0.11	0.21	0.14	0.04	0.03	0.02	0.01
20	0.03	0.03	0.01	-0.01	-0.07	-0.11	0.10	0.07	0.00	0.01
21	0.08	0.10	-0.02	0.00	0.13	0.14	-0.13	-0.19	0.01	0.00
22	0.01	0.08	-0.01	-0.01	-0.04	-0.04	0.06	0.12	-0.09	-0.11
23	-0.05	-0.02	0.04	0.01	0.06	0.01	0.01	0.01	-0.09	-0.07
24	-0.10	-0.10	-0.11	-0.03	-0.05	-0.09	-0.07	-0.14	-0.03	-0.09
25	-0.02	-0.06	-0.03	0.03	0.14	0.06	-0.16	-0.10	0.12	0.09
26	0.07	0.03	0.13	0.17	-0.05	0.00	-0.10	-0.13	-0.19	-0.20
27	0.05	0.07	0.02	0.06	0.15	0.02	0.06	0.12	0.02	0.06
28	0.01	-0.08	-0.02	0.08	0.17	0.15	-0.03	0.00	-0.03	-0.10
29	-0.03	0.03	-0.13	-0.09	-0.04	0.01	-0.05	0.05	0.05	0.06
30	0.03	-0.03	0.00	-0.11	0.12	-0.07	0.07	0.07	-0.03	-0.05
Mean	-0.008		0.005		-0.114		-0.003		0.043	
StDev	0.885		0.754		0.932		0.906		0.869	
Skewness	-0.210		0.340		-0.080		0.600		-0.360	
Kurtosis	-0.340		-0.460		-0.150		0.350		0.000	
Min	-2.335		-1.598		-2.396		-1.875		-2.265	
Max	1.820		1.684		2.315		2.695		2.038	
Median	-0.021		-0.039		-0.105		-0.101		0.076	

Table 5: *Continued*

From Table 5 above is the result of the autocorrelation functions and partial autocorrelation function from the estimates of the means parameters of the proposed mixed model.

4. Summary, Recommendation and Conclusion

In summary, this paper discusses decomposition of an observed series in which the pattern is best described by the mixed model when the trend-cycle component is exponential. The purpose is to show how estimates of residuals uncontaminated by the estimates of trend-cycle and seasonal components can be obtained. Estimates of the residuals or irregular component uncontaminated by estimates of the trend-cycle and seasonal components are obtained by successive subtraction for the Additive model and by successive division for the Multiplicative model. However, for the mixed model, estimates of the residuals obtained either by successive subtraction from or by successive division of the observed series are contaminated by estimates of the estimates of trend-cycle and seasonal components. In most cases, analysts uncertain of what to do are constrained resort to successive subtraction or successive division and end up with inadequate model or appeal to probability model, abandoning the descriptive time series even when it provides the best fit. Therefore, the ultimate objective of the study is to provide model for fitting adequate model for an observed time series data when the appropriate decomposition model is the mixed model and trend-cycle component is exponential.

The method adopted in this study is the Buys-Ballot procedure for time series decomposition The procedure is based on row, column and overall means and variances of the Buys-Ballot table. For details of the Buys-Ballot procedure, see Iwueze and Nwogu (2004, 2005 & 2014), Iwueze and Ohakwe (2004). When trend-cycle component is exponential, estimates of the trend-cycle component and seasonal indices are derived independently from the row, column and overall means of the Buys-Ballot table. Simulated series were used to illustrate results.

The results from the simulated series indicate that the proposed procedure recovered almost precisely, the parameter values used in the simulations. Evaluation of the ACF and PACF of estimates of the residuals (estimates of the irregular component) show that they lie within the 95% interval, $\pm 2/\sqrt{n}$ (where n =120). The basic statistics of the residuals also satisfy the properties of the purely random process. These indicate that the fitted models adequately describe the patterns in the simulated series.

The visual comparison clearly demonstrates that the traditional decomposition method yields contaminated residuals when applied to a mixed model with exponential trend-cycle component. In contrast, the Buys–Ballot-based procedure effectively separates the trend, seasonal, and irregular components, producing residuals consistent with a purely random process.

Therefore, the proposed model has been recommended for decomposition of any observed time series data which admits the mixed model when trend-cycle component is exponential.

References

Chatfield, C. (2004). The Analysis of Time Series: An Introduction. 6th Edition, Chapman and Hall/CRC Press, Boca Raton London.

Iwueze, I.S and Nwogu, E.C. (2004). Buys – Ballot estimates for time series decomposition, Global Journal of mathematics, 3(2), 83-98

Iwueze, S.I and Nwogu, E.C. (2005). Buys-Ballot Estimates for Exponential and S- Shaped Growth Curves, for Time Series. Journal of the Nigerian Association of Mathematical Physics, 9, 357-366.
Iwueze, I.S. and Nwogu, E.C. (2014). Framework for choice of models and detection of seasonal effect in time series. Far East Journal of Theoretical Statistics, 48, 45-66.
Iwueze, I. S and Ohakwe, J. (2004). Buys-Ballot estimates when stochastic trend is quadratic, Journal of Nigeria Association of Mathematical Physics, 8,311-318.
Iwueze, I.S, Nwogu, E.C, and Ajaraogu, J. C. (2011). Best linear unbiased estimate using Buys- Ballot procedure when trend-cycle component is linear. Pakistan Journal of Statistics and Operations Research, Vol. VII, No.2, pp 183 - 198.
Iwueze,I.S., Nwogu, E.C., Nlebedim,V.U and Imoh, J. C. (2016). Comparison of Two Time Series Decomposition Methods: Least Squares and Buys-Ballot Methods. Open Journal of Statistics,6,1123-1137.
Iwueze, S.I., Nwogu, E.C., Ohakwe, J. and Ajaraogu, J.C (2011). Uses of the Buys-Ballot.
Linde,P.(2005). Seasonal Adjustment, Statistics, Denmark.www.dst.dk/median/konrover/13-forecasting-org/seasonal/001pdf
Nwogu, Eleazar C, Iwueze, Iheanyi S, Dozie, Kelechukwu N and Mbachu, Hope I. (2019), Choice between Mixed and multiplicative models in time series decomposition, International Journal of Statistics and Applications, 9 (5), pp 153 - 159
Nzenwa, U.C (2024), Characterization of Mixed Model for Time Series Decomposition when Trend-Cycle Component are Quadratic and Exponential, unpublished Ph.D. dissertation, Department of Statistics, Imo State University, Owerri.

Declaration

Conflict of Study: The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Author Contribution Statement: Both authors conceived idea and designed the research; Analyzed interpreted the data and wrote the paper.
Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.
Ethical Approval: Not Applicable
Consent to Participate: Not Applicable
Declaration of AI Use: Not Applicable

Author(s) Bio / Authors' Note

Eleazar C. Nwogu *is from the Department of Statistics at the Federal University of Technology, Owerri. Email:* nwoguec@yahoo.com

Udoka Ikediuwa *is from the Department of Statistics, Nnamdi Azikiwe University, Awka. Email:* uc.ikediuwa@unizik.edu.ng

Appendix (A-B)

Appendix A: Row Means and Standard Deviations of the Mixed Model

	1		2		3		4		5		6		7		8		9		10	
Row	X1i	Z1i	X2i	Z2i	X3i	Z3i	X4i	Z4i	X5i	Z5i	X6i	Z6i	X7i	Z7i	X8i	Z8i	X9i	Z9i	X10i	Z10i
1	11.08	1.98	11.39	2.26	11.21	1.99	11.56	2.16	11.44	2.21	11.60	2.31	11.69	1.88	11.19	1.98	11.35	2.06	11.36	1.65
2	14.20	2.01	14.21	2.24	14.90	2.04	14.67	2.52	14.39	2.56	14.44	2.39	13.96	2.63	14.67	2.69	14.53	1.57	14.43	2.62
3	18.67	3.03	18.33	2.88	18.84	3.25	17.91	2.65	18.27	2.82	18.34	3.36	18.81	3.08	18.51	2.57	18.26	3.42	18.41	3.15
4	23.09	3.49	23.05	3.35	23.19	3.61	23.40	3.77	23.37	3.44	23.71	3.85	23.41	3.43	23.37	3.54	23.24	3.83	23.18	3.65
5	30.16	4.62	30.06	4.45	29.69	5.01	30.09	4.40	29.81	4.61	29.72	4.74	29.80	4.24	29.94	4.32	29.91	4.67	29.49	4.37
6	37.78	5.99	37.88	5.98	37.73	6.39	37.75	5.43	38.07	5.41	37.62	5.68	37.92	5.24	38.43	6.07	38.37	5.40	37.92	5.84
7	48.22	6.94	48.17	7.72	48.15	7.45	48.24	7.44	48.06	7.58	48.07	7.11	47.91	7.52	48.18	7.26	48.10	7.83	48.03	6.92
8	61.14	9.34	61.57	9.70	61.39	9.17	60.91	9.54	61.45	9.53	61.01	9.50	61.11	9.39	61.08	9.20	61.00	9.22	61.49	9.23
9	77.93	11.94	77.74	11.44	77.81	11.60	77.54	11.61	78.05	11.97	77.60	11.55	77.92	11.62	77.68	12.73	77.74	12.41	77.17	11.99
10	98.92	14.57	99.00	14.89	99.06	14.70	98.64	15.11	99.32	14.79	98.57	15.27	98.93	14.69	99.09	14.90	98.77	15.11	98.57	15.13
$\bar{X}$																				
STD																				

Appendix A continue

	11		12		13		14		15		16		17		18		19		20	
Row	X11i	Z11i	X12i	Z12i	X13i	Z13i	X14i	Z14i	X15i	Z15i	X16i	Z16i	X17i	Z17i	X18i	Z18i	X19i	Z19i	Z20i	Z20i
1	11.34	2.10	11.57	2.09	11.31	2.42	11.12	2.42	11.24	2.02	11.22	1.92	11.67	2.09	11.38	2.17	11.12	1.95	11.15	1.49
2	14.68	2.27	14.10	2.55	14.60	2.20	14.41	2.57	14.34	1.98	14.82	2.23	14.84	2.13	14.45	2.41	14.66	2.02	14.80	2.29
3	18.61	2.71	18.38	2.73	18.47	2.70	18.59	2.82	18.37	3.01	18.52	2.57	18.33	2.98	18.57	2.84	19.05	2.54	18.25	3.53
4	23.60	4.09	23.57	3.08	23.17	3.39	23.61	3.68	23.20	3.97	23.51	3.59	23.37	3.05	24.04	3.68	23.21	4.30	23.66	3.85
5	30.11	4.27	29.68	4.37	29.62	4.99	29.80	4.46	30.01	4.67	29.76	4.14	29.84	4.77	29.59	3.79	29.49	4.84	29.69	4.60
6	38.06	5.98	37.56	5.87	38.30	6.08	38.03	5.75	38.36	6.09	37.80	6.27	37.94	5.57	38.33	6.04	37.61	6.04	37.33	6.26
7	48.29	8.02	47.81	6.90	48.37	7.29	48.39	7.75	48.16	7.54	48.19	7.39	47.73	7.79	48.40	7.31	48.44	7.15	48.46	7.40
8	61.32	8.86	60.73	9.33	61.12	8.80	61.63	9.74	61.02	8.74	60.94	9.23	61.11	9.37	61.84	9.58	61.30	9.47	60.86	8.94
9	77.93	11.88	77.97	11.77	77.89	11.75	77.99	12.03	77.76	11.70	78.07	12.35	78.02	11.68	77.71	11.79	77.70	12.17	77.48	11.93
10	98.99	15.11	99.27	15.40	99.30	14.94	98.97	14.94	98.95	15.12	98.83	15.53	98.95	14.68	98.64	15.04	99.07	14.84	98.71	15.01
$\bar{X}$																				
Std																				

Appendix B: The Column Means and Standard Deviation of Mixed Model Exponential

	1		2		3		4		5		6		7		8		9		10	
Col	X1j	Z1j	X2j	Z2j	X3j	Z3j	X4j	Z4j	X5j	Z5j	X6j	Z6j	X7j	Z7j	X8j	Z8j	X9j	Z9j	X10j	Z10j
1	34.46	23.13	34.56	23.81	34.29	24.09	34.26	23.94	34.33	24.49	34.01	23.76	34.50	23.78	34.53	23.67	34.00	24.07	33.97	23.79
2	34.39	23.59	33.31	23.18	33.69	23.46	34.00	23.11	33.51	23.48	33.51	23.65	33.93	23.06	33.94	23.92	33.61	23.23	33.66	23.37
3	39.32	27.86	39.38	27.27	39.21	27.22	39.47	27.05	39.23	27.69	39.05	27.20	39.27	27.61	39.00	27.32	39.55	26.87	39.02	27.20
4	39.49	27.83	39.06	26.99	39.56	27.34	39.04	27.01	39.58	27.32	38.97	27.13	39.24	27.39	39.27	26.90	38.85	26.93	38.92	27.14
5	39.64	28.43	40.30	28.14	39.46	28.24	39.81	27.39	40.27	27.28	40.22	27.87	40.07	27.30	40.05	27.67	40.25	27.85	39.88	27.47
6	46.20	32.80	47.00	32.50	46.60	32.60	46.40	32.40	46.80	32.10	46.90	32.30	46.60	32.30	46.50	32.60	46.70	32.60	46.70	32.70
7	53.60	36.80	53.40	37.20	53.30	37.10	53.90	37.30	53.40	37.30	53.40	37.20	52.90	37.00	53.70	37.40	53.70	37.40	53.10	36.80
8	52.00	36.00	51.70	35.80	52.40	35.70	51.90	35.70	52.20	36.30	51.70	35.40	52.20	35.90	52.50	36.30	52.30	36.50	52.10	36.00
9	46.40	32.10	46.70	32.30	46.70	31.80	46.20	32.10	46.30	32.60	47.00	32.00	46.40	32.10	46.50	32.30	46.10	31.80	46.10	32.40
10	41.61	28.71	41.37	29.25	41.97	28.57	41.23	28.59	42.33	29.28	41.62	28.34	41.90	28.89	41.45	28.10	40.97	28.78	41.21	28.51
11	36.44	25.32	36.53	26.15	36.72	25.48	36.79	25.76	36.60	25.59	36.39	25.25	36.69	26.16	36.69	25.27	36.90	25.40	36.86	25.13
12	41.93	29.09	42.41	29.26	42.53	29.43	41.85	29.12	42.13	29.27	41.97	29.03	41.94	29.35	42.41	29.46	42.69	28.92	42.47	29.21
$\bar{\bar{X}}$	42.12	42.14	42.2	42.07	42.22	42.07	42.15	42.21	42.13	42.00										
Std	28.85	28.89	28.81	28.7	28.95	28.69	28.77	28.83	28.81	28.72										

Appendix B continues

	11		12		13		14		15		16		17		18		19		20	
Col	X11j	Z11j	X12j	Z12j	X13j	Z13j	X14j	Z14j	X15j	Z15j	X16j	Z16j	X17j	Z17j	X18j	Z18j	X19j	Z19j	X20j	Z20j
1	34.57	23.78	34.13	23.70	34.40	23.91	34.27	23.57	34.83	23.43	34.44	23.64	33.94	23.60	34.76	23.58	34.25	23.66	33.73	24.00
2	33.29	23.72	34.01	23.49	33.96	23.48	33.96	23.16	33.68	23.76	33.85	23.06	33.59	23.48	33.93	23.18	33.84	23.43	33.96	23.38
3	39.41	27.43	38.98	26.86	38.85	27.47	38.74	27.62	39.22	27.31	39.11	27.06	39.72	27.49	39.43	27.22	39.14	26.80	39.17	26.59
4	39.63	27.31	38.99	27.31	39.28	27.53	39.05	26.71	38.65	27.08	38.83	26.73	39.13	27.74	39.26	27.33	39.54	27.19	39.29	27.39
5	40.07	27.51	39.60	27.67	40.43	27.52	40.03	27.90	40.17	27.62	40.17	27.15	40.14	26.93	39.73	28.24	40.03	27.93	39.93	27.50
6	46.40	32.20	47.00	32.80	46.70	32.50	47.30	32.70	46.70	32.40	46.30	32.60	46.50	32.10	47.20	32.20	46.70	32.00	46.30	32.40
7	53.90	37.00	53.20	37.50	53.50	37.10	53.70	37.10	53.80	37.00	53.80	37.30	53.50	36.80	53.60	37.20	53.60	37.30	53.80	36.60
8	52.10	36.30	51.80	36.20	52.30	35.60	52.40	36.20	51.90	36.00	52.30	37.00	51.50	36.10	52.10	36.10	52.20	36.30	51.60	36.20
9	46.60	32.40	46.40	32.10	46.60	32.80	46.50	32.20	46.50	32.50	46.40	32.20	46.40	32.40	46.60	32.30	46.70	32.30	46.80	32.50
10	41.77	28.61	41.70	28.69	40.97	29.14	41.85	29.31	41.11	28.99	41.84	29.08	41.83	29.10	41.58	28.22	41.34	29.34	41.66	29.00
11	36.99	25.23	36.90	26.09	37.23	25.69	36.90	26.49	36.74	25.90	36.87	25.89	37.23	25.31	36.79	25.60	36.44	26.31	36.46	25.28
12	42.68	29.23	42.04	29.43	42.50	29.31	42.26	29.01	42.46	29.21	42.08	28.90	42.62	29.33	42.56	28.93	42.24	28.64	41.73	28.91
$\bar{\bar{X}}$	42.29	42.06	42.22	42.25	42.14	42.17	42.18	42.29	42.17	42.04										
Std	28.81	28.89	28.89	28.92	28.83	28.83	28.76	28.75	28.85	28.73										