Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/140104

Odd Half Logistic Chen Distribution for Analyzing Air Quality Data in Kathmandu, Nepal

Ramesh Prasad Tharu[1,*], *Govinda Prasad Dhungana*[2] *and Ramesh Kumar Joshi*[3]

[1]Department of Statistics, Tribhuvan University, Mahendra Multiple Campus, Nepalgunj, Nepal
[2]Department of Statistics, Tribhuvan University; Birendra Multiple Campus, Bharatpur Chitwan, Nepal
[3]Department of Statistics, Tribhuvan University, Trichandra Multiple Campus Saraswoti Sadan, Kathmandu, Nepal

Received: 29 Jul. 2023, Revised: 2 Oct. 2024, Accepted: 9 Oct. 2024
Published online: 1 Jan. 2025

Abstract: This research paper presents a novel statistical model, namely the Odd Half Logistic Chen Distribution, which combines the continuous Chen distribution with the half logistic-G family of distribution. Where,α and λ are scale parameters and β is the shape parameter. The model exhibits unimodal characteristics with a negative skewness, while the hazard rate function displays an increasing J-shaped pattern. The study derives explicit expressions for various important statistical functions, including the reliability/survival function, hazard rate function, revised hazard rate function, cumulative hazard rate function, quantile function, mode function, and order statistics. The maximum likelihood estimation method is employed to estimate the model parameters and a simulation study confirms the efficiency of the approach as the sample size increases, indicated by the decreasing mean squared errors of the individual parameters. Furthermore, the proposed model is applied to a real-world air quality data analysis in Kathmandu Valley. The finding reveals that residents of the area experience only seven days of fresh air per month, highlighting the severity of the air pollution problem. The model is validated through graphical techniques such as P-P plot, Q-Q plot, estimated cumulative distribution function (CDF) with empirical distribution and numerical tests including the Kolmogorov-Smirnov (KS) test, Anderson Darling test and Cramér–von Mises test. The parameter estimation, model validation and statistical analysis are performed using R programming. Therefore, the proposed model emerges as a promising alternative for predicting air quality data and performing reliability analysis in various domains.

Keywords: Chen distribution, Maximum likelihood Estimation, Nepal, Order statistics, PM 2.5 air quality

1 Introduction

The development of novel methods for broadening existing probability distributions has been a significant scholarly concern in recent years. The theory of distributions has witnessed rapid expansion, facilitated by various approaches such as compounding of distribution, mixing of distribution, utilizing any distribution as a generator, power transformation techniques and inverse transformation techniques. The primary objective behind the generalization of new distributions is to capture the diverse characteristics exhibited by different types of data that cannot be fully explored by conventional probability distributions [1]. As a result, flexible probability distributions with additional shape parameters have been explored and proposed in the literature [2].

For instance, the type II half-logistic exponential distribution is formed by employing the type II half logistic distribution as a generator within the generalized gamma family of distributions [3]. Type I generalized half logistic distribution is derived through exponential random variable transformation, and a theorem relating the generalized distribution to the Pareto distribution has been established [4]. This half-logistic distribution has been applied in reliability analysis. A novel continuous distribution for modelling positive real-life data, achieved through the transformation of a half logistic random variable, resulting in the development of the generalized half logistic and power half logistic distributions. The power half logistic distribution, in particular, exhibits greater convexity, concavity, and heavy-tailedness, depending on the values of its parameters [5]. Furthermore, the half-logistic Lomax distribution, combining parameters from the half-logistic and Lomax distributions, has been proposed as a new lifetime model,

* Corresponding author e-mail: rameshpt02@gmail.com

commonly utilized in reliability, engineering, and survival analysis. The proposed distribution is heavily tail, decreasing upside-down bathtub (unimodal) shaped hazard rate function [6]. The half logistic exponential extension model is developed by compounding the type I half logistic-G family with the exponential extension distribution, resulting in a flexible distribution with negative-skewed, positive-skewed and symmetrical properties. The hazard rate function of the proposed distribution is also flexible due to its various shapes, such as monotonically decreasing, increasing and constant [7]. Hence, numerous other distributions have also been developed to provide flexibility and effectively represent the diverse characteristics of data.

In terms of parameter estimation, various techniques have been applied in different distribution models. For example, the odd exponentiated half-logistic exponential distribution employs eight techniques, including maximum likelihood estimation, least squares, Anderson-Darling, the maximum product of spacing, weighted least squares, Cramér-von Mises, percentiles, and right-tail Anderson-Darling, to estimate its parameters[8]. Likewise, The Type-II Quasi Lambert-G family of probability distributions has been employed, utilizing various estimation methods including maximum likelihood estimation, maximum product spacing estimation, least squares estimation, weighted least squares estimation, Anderson-Darling estimation, and Cramer-von Mises estimation [9]. Similarly, various estimation techniques have been employed in the available literature [10,11]. Consequently, this study was motivated by the development of a new probability distribution, using different estimation methods for modelling of air quality data.

In the context of introducing a new distribution and estimating its parameters using real data, we focus on air quality data such as the PM2.5 concentration in the Kathmandu Valley, Nepal. PM2.5 refers to fine particulate matter with a diameter of 2.5 micrometers or less, and it is widely recognized as a crucial indicator of air quality. The World Health Organization [12] has defined PM2.5 levels below $10\,\mu g/m^3$ as safe for human health. However, with the impact of global warming in Nepal, there is a concerning upward trend in PM2.5 levels. Exceeding the threshold of $10\,\mu g/m^3$ for PM2.5 poses a significant risk of poisoning and adverse health effects for individuals. To address this issue, a new three-parameter flexible distribution known as the Odd Half Logistic Chen (OHLC) distribution has been proposed as a means to predict air quality in Kathmandu, Nepal.

2 Material and Methods

2.1 Model Analysis

Let $\phi(t)$ be the probability density function of random variable $T \in$[a,b],defined on interval $-\infty < a \leq b < \infty$. $\omega(F(x))$ be any function of the cumulative distribution function $F(x)$ of any random variable X, which satisfied the following properties:

$\omega(F(x)) \in [a,b]$;

$\omega(F(x))$is differentiable and monotonic nondecreasing function; and

$\omega(F(x)) \to a$ as $x \to -\infty$ and$\omega(F(x)) \to$ b as $x \to \infty$.

Let, $\omega(F(x))$ is an odd function and $F(x) = 1 - e^{\lambda\left(1-\exp\left(x^{\beta}\right)\right)}$ is a cumulative distribution function of Chen distribution which satisfied the above properties, and odd function can be written as;$\omega(F(x)) = \exp\left\{-\lambda\left(1-\exp\left(x^{\beta}\right)\right)\right\} - 1$; where,$\beta$ is shape parameter and $\lambda$is the scale parameter. Similarly, the T-X family of distribution is an extended form of beta generated distribution, where random variable T serves as a generator instead of a beta random [13] defined as;

$$F(x) = \int_a^{\omega(F(x))} \phi(t)dt; \tag{1}$$

where,$\phi(t)$ as a generator which has been used for probability density function of the half logistic distribution. The half logistic distribution has widespread application of reliability and survival data modeling [6]. The pdf of $\phi(t)$ is

$$\phi(t) = \frac{2\alpha e^{-\alpha t}}{(1+e^{-\alpha t})^2}\ ;\mathrm{t} > 0, \alpha > 0. \tag{2}$$

Applying the equation (2) in equation (1) and integrating equation (1) using $\omega(F(x))$=$\exp\left\{-\lambda\left(1-\exp\left(x^{\beta}\right)\right)\right\} - 1$; it yields the Cumulative Distribution Function (CDF) of the Odd Half Logistic Chen (OHLC) distribution,

$$F(x) = \int_0^{\exp\left\{-\lambda\left(1-\exp\left(x^{\beta}\right)\right)\right\}-1} \frac{2\alpha e^{-\alpha t}}{(1+e^{-\alpha t})^2}\mathrm{dt}$$

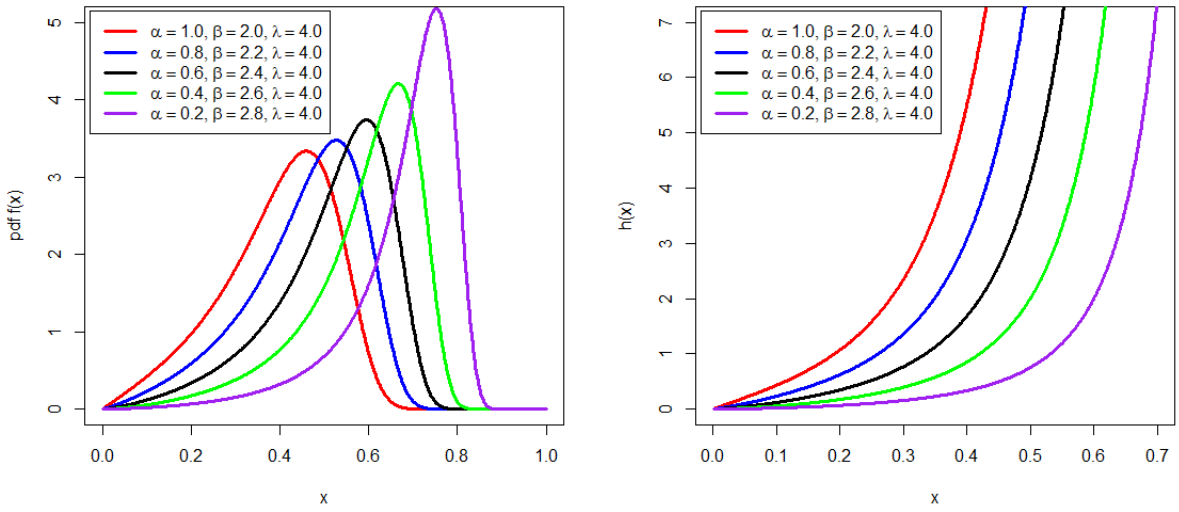

Fig. 1: Plot of density function (left panel) and hazard rate function (right panel) of odd half logistic Chen distribution.

$$= \frac{1-\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}{1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}; x \geq 0, \alpha > 0, \beta > 0, \lambda > 0. \tag{3}$$

The corresponding Probability Density Function (PDF) of proposed distribution is

$$f(x) = \frac{2\alpha\beta\lambda x^{\beta-1}e^{x^\beta}\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}{\left[1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]\right]^2}; x \geq 0, \alpha > 0, \beta > 0, \lambda > 0. \tag{4}$$

The hazard rate function is the conditional density given that the event has not occurred before time x,

$$h(x) = \frac{\alpha\beta\lambda x^{\beta-1}e^{x^\beta}\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}}{1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}; x \geq 0, \alpha > 0, \beta > 0, \lambda > 0. \tag{5}$$

The fig. 1 suggests that the probability density function plot exhibits a negative skew, while the hazard rate function shows a monotonically increasing and J-shaped form. The survival function of the proposed distribution is

$$R(x) = \frac{2\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}{1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}. \tag{6}$$

Likewise, the reverse hazard rate function is also significant in reliability and survival data analysis. It is defined as the ratio of the density function to the distribution function. The reverse hazard rate function of the proposed model is

$$r(x) = \frac{2\alpha\beta\lambda x^{\beta-1}e^{x^\beta}\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]\left[1-\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]\right]}{1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]}. \tag{7}$$

The cumulative hazard rate function is defined as the integral of the hazard rate function from time 0 up to a specific time x. It represents the cumulative risk or accumulated failure probability up to that point in time. Mathematically, $H(x) = -\ln[R(x)]$.

$$H(x) = -\ln(2) - \alpha\left[1 - \left\{\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right] + \ln\left[1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]\right]. \tag{8}$$

2.2 *Statistical Properties*

In this section some properties of the OHLC distribution have been derived. The proposed distribution derived from the generalized binomial and exponential series [14]. For,$|z|<1$, $n>0$; we have,

$(1+z)^{-n} = \sum_{i=0}^{\infty}(-1)^i \binom{n+i-1}{i} z^i; (1-z)^n = \sum_{j=0}^{\infty}(-1)^j \binom{n}{j} z^j$;and $e^{-ax} = \sum_{k=0}^{\infty}(-1)^k \frac{(ax)^k}{k!}$

The PDF of proposed distribution (4) derived by using the generalized binomial and exponential series as,

$$f(x) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} \kappa_{ijklm} x^{\beta-1} e^{(m+1)x^\beta}; \tag{9}$$

$$\text{where,} \kappa_{ijklm} = \frac{2\alpha\beta\lambda(-1)^{i+j+k+l+m}(i+1)^{K+1}\{\lambda(j+1)\}^l \binom{k}{j}\binom{l}{m}}{k!l!}.$$

Similarly, the CDF of proposed distribution (3) derived by using the generalized binomial and exponential series as,

$$F(x) = 1 - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty} \xi_{ijklm} e^{mx^\beta} \tag{10}$$

where,

$$\xi_{ijklm} = \frac{(-1)^{i+j+k+l+m}\{(1-i)\alpha\}^j\{\lambda k\}^l \binom{j}{k}\binom{l}{m}}{j!l!}.$$

2.2.1 **Quantile Function and Median**

The quantile functions are utilized in the theoretical aspects of probability theory. They serve as an alternative to the probability density function and cumulative distribution function and are used to calculate statistical measures such as the median, skewness, and kurtosis. The quantile function is defined as follows: $Q(\rho) = F^{-1}(x)$. Consequently, the corresponding quantile function of the proposed distribution is

$$F^{-1}(x) = Q(\rho) = \left[\ln\left[1+\frac{1}{\lambda}\ln\left\{1-\frac{1}{\alpha}\ln\left(\frac{1-\rho}{1+\rho}\right)\right\}\right]\right]^{\frac{1}{\beta}}; 0<\rho<1. \tag{11}$$

Whereas, in particular, the median is derived by setting $\rho = \frac{1}{2}$ in quantile function (11), we get;

$$Median = \left[\ln\left[1+\frac{1}{\lambda}\ln\left\{1-\frac{1}{\alpha}\ln\left(\frac{1}{3}\right)\right\}\right]\right]^{\frac{1}{\beta}}.$$

2.2.2 **Mode**

To calculate the mode of the proposed distribution, we need to find the maximum recurring value. This can be achieved by differentiating equation (4) with respect to the variable x or by taking the logarithm of equation (4), which is an equivalent approach.

$$\begin{aligned}\ln f(x) = {} & \ln(2\alpha\beta\lambda) + (\beta-1)\ln(x) + x^\beta - \lambda\left(1-e^{x^\beta}\right) + \alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\} \\ & - 2\ln\left[1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^\beta}\right)\right\}\right\}\right]\right]\end{aligned} \tag{12}$$

The equation (12) is differentiating with respect to x and apply the condition $f(x) \neq 0$ and $f'(x) = 0$, the mode of proposed distribution is

$$\beta\left[1+x^{\beta}\left\{1+\lambda e^{x^{\beta}}\left\{1-\alpha\exp\left\{-\lambda\left(1-e^{x^{\beta}}\right)\right\}\right\}\right\}\right]+\frac{f(x)}{x}\left[1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x^{\beta}}\right)\right\}\right\}\right]\right]=1$$

The above equation is nonlinear equation which is solved by numerical methods.

2.2.3 **Order Statistics**

Order statistics play a crucial role in various fields of statistics, including reliability analysis and life testing. Order statistics involve arranging a set of observations in ascending or descending order and studying the properties of specific positions within the ordered sequence. They provide valuable insights into the distribution, variability and extreme values of a dataset, which are essential in reliability analysis and life testing applications. Order statistics have been widely used in many fields of statistics, including reliability and life testing. Let, $X_1, X_2, ..., X_n$ be random sample from (4) and$X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ are the corresponding order statistics. The probability density function of r^{th} order statistics say$X_{r:n}; 1 \leq r \leq n$ [15] is

$$f_{r:n}(x) = \frac{n!}{(r-1)!(n-r)!} f(x)\left[F(x)\right]^{r-1}\left[1-F(x)\right]^{n-r}. \tag{13}$$

We apply the equation (9) and (10) in equation (13) then it becomes the equation

$$\begin{aligned} f_{r:n}(x) = {} & \tfrac{n!}{(r-1)!(n-r)!}\textstyle\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\kappa_{ijklm}{x_{(r)}}^{\beta-1}e^{(m+1)x_{(r)}{}^{\beta}} \\ & \left[1-\textstyle\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\xi_{ijklm}e^{mx_{(r)}{}^{\beta}}\right]^{r-1}\left[\textstyle\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\xi_{ijklm}e^{mx_{(r)}{}^{\beta}}\right]^{n-r} \end{aligned} \tag{14}$$

When, $r = n$ then from equation (14), the pdf of the largest order statistics $X_{n:n}$is

$$f_{n:n}(x) = n\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\kappa_{ijklm}{x_{(n)}}^{\beta-1}e^{(m+1)x_{(n)}{}^{\beta}}\left[1-\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\xi_{ijklm}e^{mx_{(n)}{}^{\beta}}\right]^{n-1};x_{(n)}>0$$

Similarly, $r = 1$, then from equation (14), the pdf of smallest order statistics $x_{1:1}$ is

$$f_{1:n}(x) = n\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\kappa_{ijklm}{x_{(1)}}^{\beta-1}e^{(m+1)x_{(1)}{}^{\beta}}\left[\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\xi_{ijklm}e^{mx_{(1)}{}^{\beta}}\right]^{n-1};x_{(1)}>0.$$

2.3 Maximum Likelihood Estimation

We have to estimate the unknown parameters of proposed model by maximum likelihood estimation. Let, $x_1, x_2, ..., x_n$are random sample drawn from OHLC distribution with parameters$(\alpha,\ \beta,$and $\lambda)$, then likelihood function of proposed distribution is product of n^{th} time of sample of proposed distribution. Mathematically, $\ell(x;\zeta) = \prod_{i=1}^{n} f(x_i;\zeta)$where,$\zeta$ is the parameters pace belong to$(\alpha,\ \beta,$and $\lambda)$. The likelihood function is equivalent to log likelihood function. Therefore, log likelihood function of proposed distribution is

$$\begin{aligned} \ell(x;\zeta) = {} & n\ln(2\alpha\beta\lambda)+(\beta-1)\textstyle\sum_{i=1}^{n}\ln(x_i)+\sum_{i=1}^{n}x_i{}^{\beta}-\lambda\sum_{i=0}^{n}\left(1-e^{x_i{}^{\beta}}\right) \\ & +\alpha\textstyle\sum_{i=0}^{n}\left\{1-\exp\left\{-\lambda\left(1-e^{x_i{}^{\beta}}\right)\right\}\right\}-2\sum_{i=0}^{n}\ln\left[1+\exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x_i{}^{\beta}}\right)\right\}\right\}\right]\right]. \end{aligned} \tag{15}$$

The parameters are obtained by partial differentiating in equation (15) with respect to $(\alpha,\ \beta,$and $\lambda)$. Let,$\xi_i = \exp\left\{-\lambda\left(1-e^{x_i{}^{\beta}}\right)\right\}$ and $\upsilon_i = \exp\left[\alpha\left\{1-\exp\left\{-\lambda\left(1-e^{x_i{}^{\beta}}\right)\right\}\right\}\right]$

we have

$$\frac{\partial\ell(x;\zeta)}{\partial\alpha} = \frac{n}{\alpha}+\sum_{i=1}^{n}(1-\xi_i)-2\sum_{i=1}^{n}\left(\frac{\upsilon_i(1-\xi_i)}{1+\upsilon_i}\right). \tag{16}$$

$$\frac{\partial \ell(x;\zeta)}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n} \ln(x_i) + \sum_{i=1}^{n} {x_i}^{\beta} \ln(x_i) + \lambda \sum_{i=1}^{n} {x_i}^{\beta} \ln(x_i) e^{{x_i}^{\beta}} - \alpha\lambda \sum_{i=1}^{n} \xi_i {x_i}^{\beta} \ln(x_i) e^{{x_i}^{\beta}} + 2\alpha\lambda \sum_{i=1}^{n} \frac{\xi_i \upsilon_i}{1+\upsilon_i} {x_i}^{\beta} \ln(x_i) e^{{x_i}^{\beta}} . \quad (17)$$

$$\frac{\partial \ell(x;\zeta)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} \left(1 - e^{{x_i}^{\beta}}\right) + \alpha \sum_{i=1}^{n} \xi_i \left(1 - e^{{x_i}^{\beta}}\right) - 2\alpha \sum_{i=1}^{n} \frac{\upsilon_i \xi_i}{1+\upsilon_i} \left(1 - e^{{x_i}^{\beta}}\right). \quad (18)$$

Finally, solving and estimating non-linear equations $\frac{\partial \ln(\ell)}{\partial \alpha} = 0, \frac{\partial \ln(\ell)}{\partial \beta} = 0, \frac{\partial \ln(\ell)}{\partial \lambda} = 0$ and estimate ($\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$) for parameters (α, β, and λ). Furthermore, the asymptotic normality of MLEs, approximate $100(1-\gamma)\%$ confidence intervals of α β, and λcan be constructed as: $\hat{\alpha} \pm z_{\gamma/2} SE(\hat{\alpha})$, $\hat{\beta} \pm z_{\gamma/2} SE(\hat{\beta})$,$\hat{\lambda} \pm z_{\gamma/2} SE(\hat{\lambda})$ and $\hat{\delta} \pm z_{\gamma/2} SE(\hat{\delta})$;$z_{\gamma/2}$is the upper percentile of standard normal variate.

3 Results and Discussion

Data Analysis
Data analysis is a crucial technique used to draw valid conclusions based on facts and information. In order to assess the suitability of the proposed model for a given dataset, two different techniques have been employed: simulation study and real data analysis. In a simulation study, the proposed model is applied to simulated data that generating data according to the proposed model and comparing the results with known properties or characteristics. On the other hand, real data analysis involves applying the proposed model to actual observed data. Hence, using both methods, we can evaluate the performance and adequacy of the model.

3.1 ***Simulation Study***

In order to validate the theoretical performance of the maximum likelihood estimators for the new probability distribution, a Monte Carlo simulation was conducted. The goal of this simulation was to examine the performance of estimation methods, primarily in terms of mean square errors (MSEs), across different sample sizes. For the purpose of estimation, 10,000 random samples of size 50, 100, 200, and 500 were generated from R programming. These samples were drawn from the proposed probability distribution to closely on real-world scenarios.

1.Compute the MLEs for 10000 samples, say $\left(\hat{\alpha}_i, \hat{\beta}_i, \hat{\lambda}_i\right)$;for *i*=1, 2, …, 10000.

2.Compute the mean square error (MSEs) of different sample size as $MSE_{\underset{\sim}{\delta}}(n) = \frac{1}{10000} \sum_{i=1}^{10000} \left(\underset{\sim}{\hat{\delta}}_i - \underset{\sim}{\delta}\right)^2$,where $\delta = (\alpha, \beta, \lambda)$.

3.To calculate the MSEs, the initial parameter values of the proposed distribution;($\alpha = 1.0, \beta = 1.5$ and $\lambda = 2.0$),and($\alpha = 1.5, \beta = 2.0$ and $\lambda = 2.5$)has been set.

For the given sample sizes of 100, 200, 300, 400 and 500, the maximum likelihood estimation method is applied to estimate the parameters. The MSEs for individual parameters are then computed. Upon analyzing the results, it is observed that as the sample size increases, the MSEs for the individual parameters decrease. This finding emphasizes the effectiveness of the maximum likelihood estimation method in accurately estimating the parameters of the proposed distribution. The decreasing MSEs indicate that with larger sample sizes, the estimators tend to converge towards the true parameter values, resulting in more precise and reliable estimations (Fig 2.).

3.2 ***Real Data Analysis***

In the Kathmandu valley, there are seven air quality monitoring stations: Ratnapark, Sankhapark, Bhaisepati, Pulchowk, Bhaktapur, Kirtipur, and Dhulikhel. Among the seven places, we randomly selected four location; Bhaktapur, Bhaisepati, Kirtipur, and Ratnapark as sample sites. The researchers obtained data on particulate matter (PM 1, PM 2.5, PM 10) and total suspended particulates (TSP) for the period from January 1, 2021, to December 31, 2021, from the Department of

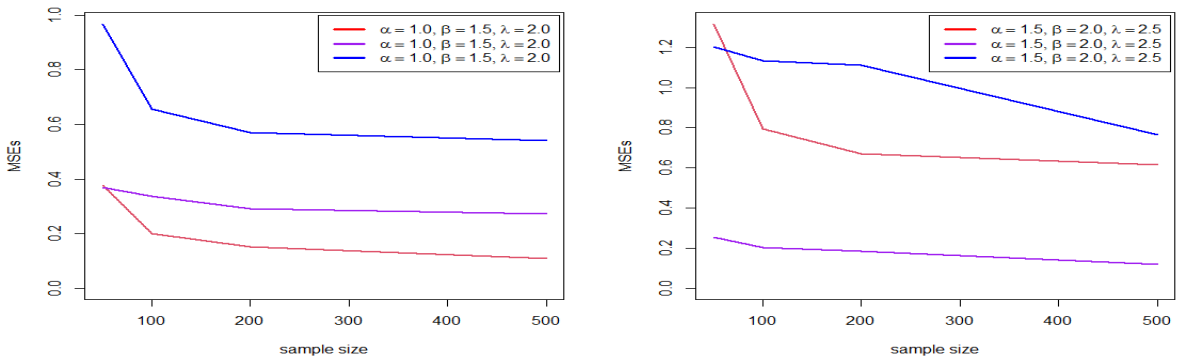

Fig. 2: Plot of mean square error of different values with different sample size.

Environment, Kathmandu, Nepal [16]. Specifically, the researchers focused on the $PM2.5(\mu/m^3)$ dataset for proposed distribution because PM2.5 is the most important air pollutant among PM1, PM10, and TSP due to its small size, deep lung penetration, and association with various serious health problems, including heart disease, stroke, lung cancer, asthma, chronic bronchitis, and premature death. Monitoring and reducing PM 2.5 levels is crucial for protecting public health issue. Also, World Health Organization (WHO) recommended that the air quality levels of PM 2.5 exceeded $10\mu/m^3$, which is considered dangerous for human health [12].

3.2.1 **Exploratory Data Analysis**

The descriptive statistics of the data reveal a significantly skewed distribution and notably high levels of PM 2.5 in the Kathmandu Valley. This observation strongly suggests that the air quality in the Kathmandu Valley exceeded the threshold outlined by the WHO, indicating a severe deterioration in air quality conditions (Table 1, Fig. 3 and 4).

Table 1: Descriptive statistic of PM 2.5 level in different station in Kathmandu Valley

Location	Minimum	Q_1	Mean	Median	Q_3	Maximum
Bhakatpur	6.33	13.93	54.69	55.90	78.68	226.19
Bhaisipati	4.90	13.30	54.11	46.85	78.25	238.30
Kritipur	5.30	13.15	42.08	23.08	61.83	210.80
Ratnapark	8.88	16.82	53.92	48.36	80.95	206.03

Similarly, we have presented the data using box plots to explain the variations in different locations during distinct seasons. The seasons has been defined as Nepalese contest: Winter spans from November 17 to March 15, Spring occurs from March 16 to May 14, Summer extends from May 15 to September 17, and Autumn encompasses the period from September 18 to November 16. The findings revealed that the air quality levels in the Kathmandu Valley were problematic during the Spring and Winter seasons. This implies that there are significant challenges related to air quality during these particular seasons. However, it was contradiction finding on Kirtipur and Bhaktapur in spring, autumn and winter season. (Fig. 5 and 6).

The TTT plot is a useful tool to understand the behavior of data's Hazard Rate Function (HRF). A diagonal line suggests a constant HRF, while a concave TTT plot indicates an increasing HRF and a convex TTT plot suggests a decreasing HRF. The shape can also be a combination of concave and convex, indicating a unimodal or bathtub hazard rate. In Figure 7 and 8, the scaled TTT plot has a concave shape as well as convex, indicating that the air quality data has a unimodal or bathtub hazard rate. This confirms that the proposed distribution is suitable for modelling the data (Fig. 7, 8).

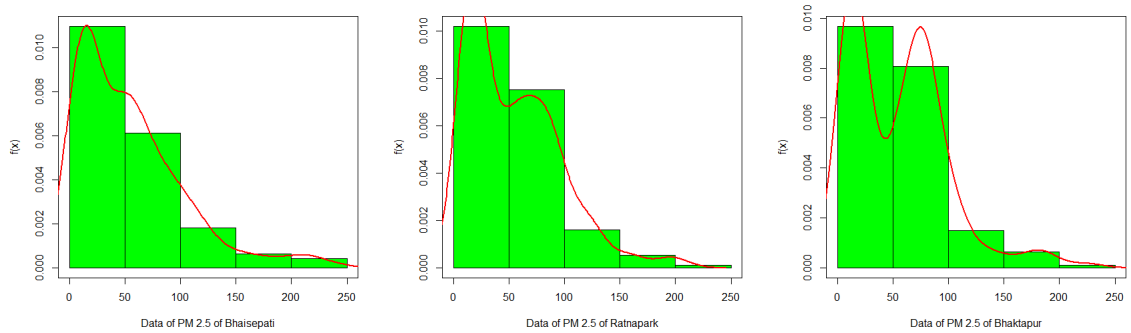

Fig. 3: Histogram and density plot of Bhaisipati, Ratnapark and Bhakatpur along with the data points.

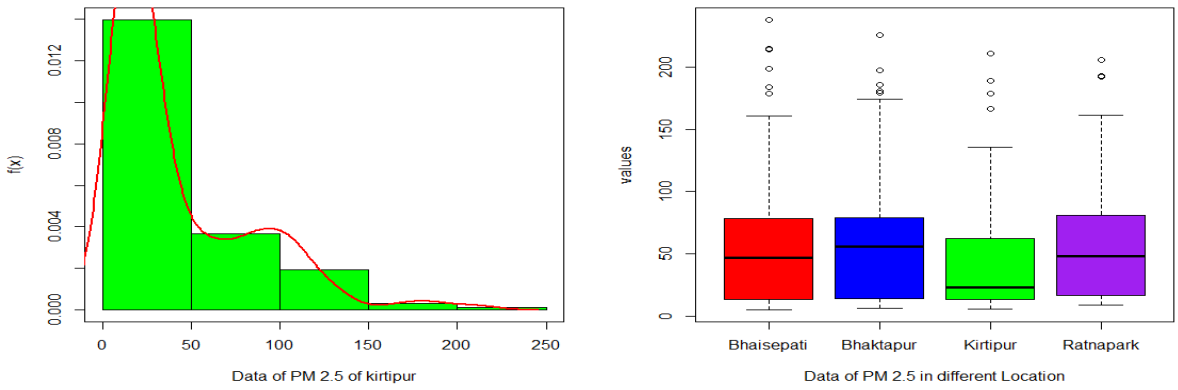

Fig. 4: Histogram and density plot (left panel) of Kirtipur and boxplot (right panel) along with the data points of different location like as spring, summer, autumn and winter.

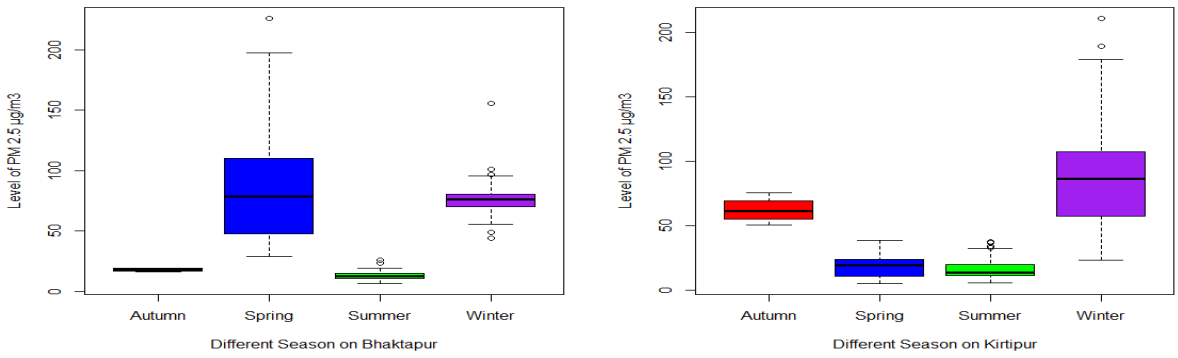

Fig. 5: Boxplot along with the data points of Bhaktapur (left panel) and Kirtipur (right panel) according to spring, summer, autumn and winter.

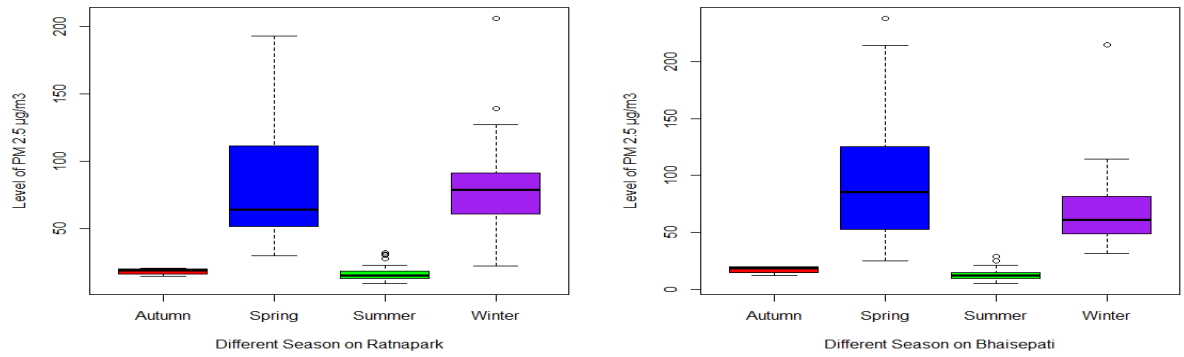

Fig. 6: Boxplot along with the data points of Ratnapark (left panel) and Bhaisepati (right panel) according to spring, summer, autumn and winter.

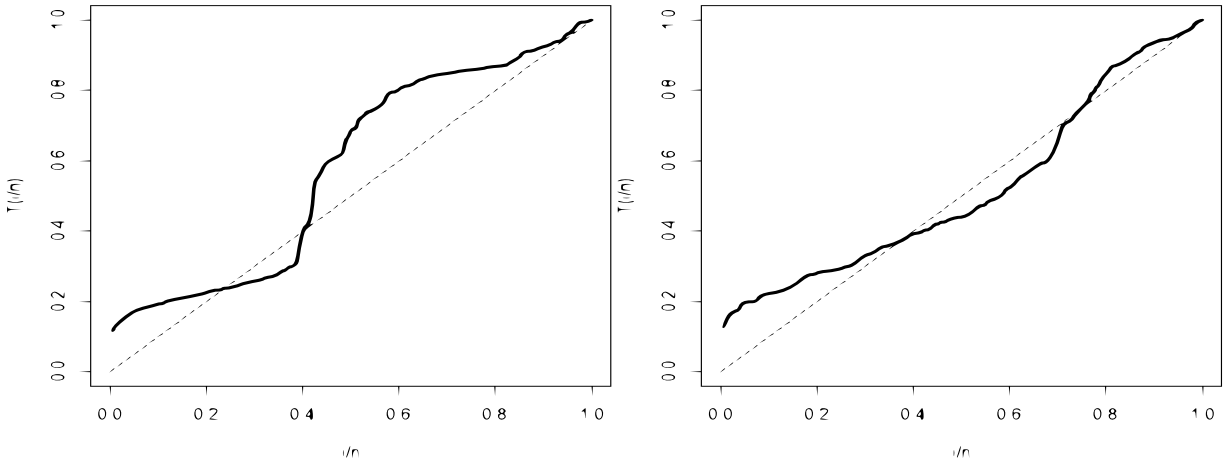

Fig. 7: TTT plots (Bhakatpur) and (Kritipur) along with the data points.

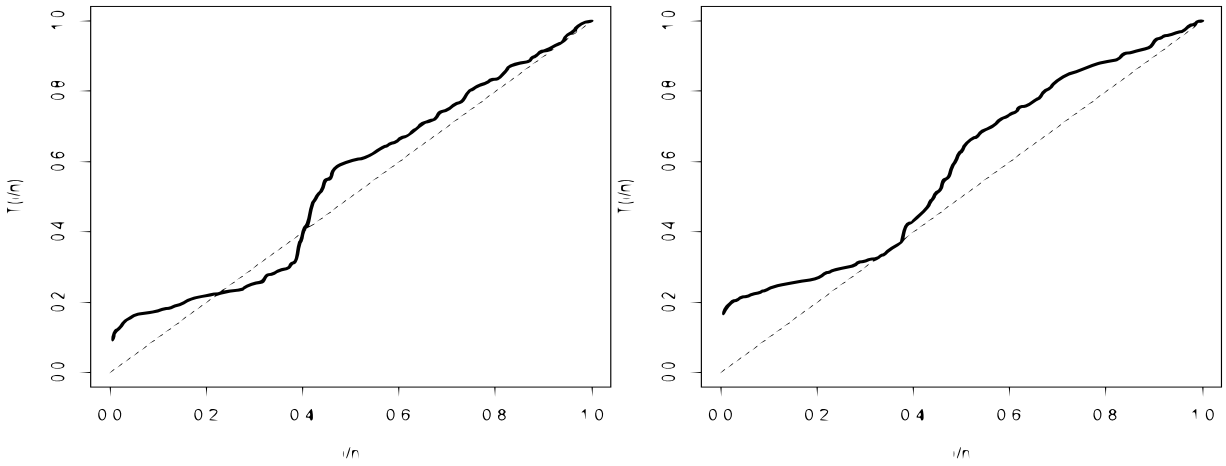

Fig. 8: TTT plots (Bhaisepati) and (Ratnapark) along with the data points.

3.2.2 **Parameter Estimation**

The estimated parameters and their corresponding standard errors were obtained using the maximum likelihood estimation method with the "CG" method in the R software [17,18]. The parameters were estimated and presented separately for different seasons. Table 2 summarizes the estimated parameter values and their corresponding standard errors. The findings revealed that during the Autumn season, there was no significant variation observed in the estimates of all parameters. However, in the Summer, Spring, and Winter seasons, significant variations were observed in the shape parameters, indicating differences in the spread or variability of the data across these seasons. The finding revel that shape parameter ($\beta < 1$) indicates that the failure rate decreases over time. In the context of air quality data, this suggests that air quality is more likely to worse at early season. On the other hand, no significant variation was observed in the scale parameters (α and λ) , suggesting that the data did not differ significantly across these seasons (Table 2).

Table 2: Estimated value of MLE with SE of four different seasons

Seasons	$\hat{\alpha}$	t-value	P-value	$\hat{\beta}$	t-value	P-value	$\hat{\lambda}$	t-value	P-value
Summer	0.01281	1.419	0.156	0.19558	11.037	<0.001	1.02793	4.039	<0.001
	(0.00903)			(0.01772)			(0.25451)		
Winter	0.003925	1.238	0.216	0.14816	12.478	<0.001	0.99875	3.975	<0.001
	(0.003171)			(0.0118)			(0.25126)		
Spring	0.01357	0.865	0.3873	0.12017	5.932	<0.001	1.04404	2.331	0.0197
	(0.01569)			(0.02026)			(0.44785)		
Autumn	0.01901	0.278	0.7811	0.13633	1.801	0.0718	1.08364	0.745	0.4562
	(0.06839)			(0.07571)			(1.45430)		

Parentheses indicate the standard error

3.2.3 **Prediction of PM 2.5 Air Quality in Kathmandu Valley**

Based on the estimated parameter values, PM2.5 air quality levels of $< 10\,\mu g/m^3$ are considered safe for human health, whereas $\geq 10\,\mu g/m^3$ pose serious threats of poisoning for human health [12]. Here, air quality predictions for four different seasons reveal a significant prevalence of poor air quality in Kathmandu. During summer, 69.74% of the observed air samples exceeded the safe threshold for PM2.5 levels of $10\,\mu g/m^3$ or less. This percentage increased to 78.53% in autumn, 88.88% in spring, and 95.94% in winter. Only 7 days per month exhibited acceptable air quality, while the remaining 23 days per month posed a substantial risk of poisoning due to elevated PM2.5 levels. These findings emphasize a significant public health burden, contributing to an increased occurrence of respiratory morbidities such as chronic obstructive pulmonary disease and acute respiratory infections, including pneumonia, within the Kathmandu Valley (Table 3).

Table 3: Prediction of air quality in four different seasons by proposed model.

Seasons		(0-5) μ/m^3	(5-10) μ/m^3	($\geq$10) μ/m^3
Summer	Predicted probability	0.1238	0.1787	0.6974
	Expected days/month	04	06	20
Autumn	Predicted probability	0.1285	0.0861	0.7853
	Expected days/month	04	03	23
Spring	Predicted probability	0.0732	0.0379	0.8888
	Expected days/month	03	02	25
Winter	Predicted probability	0.0232	0.1733	0.9593
	Expected days/month	01	01	28

3.2.4 **Model Validation**

The development of a new model necessitates the validation process, which includes obtaining sample data from different seasons. However, for this particular case, only winter and spring season has been presented due to the higher likelihood (95.93%) and (88.88%) of experiencing foul weather during these seasons. To validate the winter data using the proposed model, various goodness-of-fit criteria were employed. The Anderson-Darling test (A_n=2.0488, p-value = 0.08714), Kolmogorov-Smirnov test (D=0.25107, p-value = 0.1123) and the Cramér–von Mises test (Ω_2=0.40481, p-value 0.069) respectively. These criteria indicated that the proposed distribution is valid. Furthermore, to assess the validity of the model fit, a Q-Q plot is constructed, representing the relationship between the ordered values derived from the dataset and the corresponding quantiles obtained from the theoretical distribution. Additionally, a P-P plot is generated, illustrating the comparison between the cumulative probabilities derived from the dataset and those originating from the proposed distribution. These plots collectively demonstrate that the model has been fitted appropriately, thus indicating the validity of the proposed distribution for the given dataset (Fig 9). Similarly, to validate the Spring season data using the proposed model, various goodness-of-fit criteria were applied likes

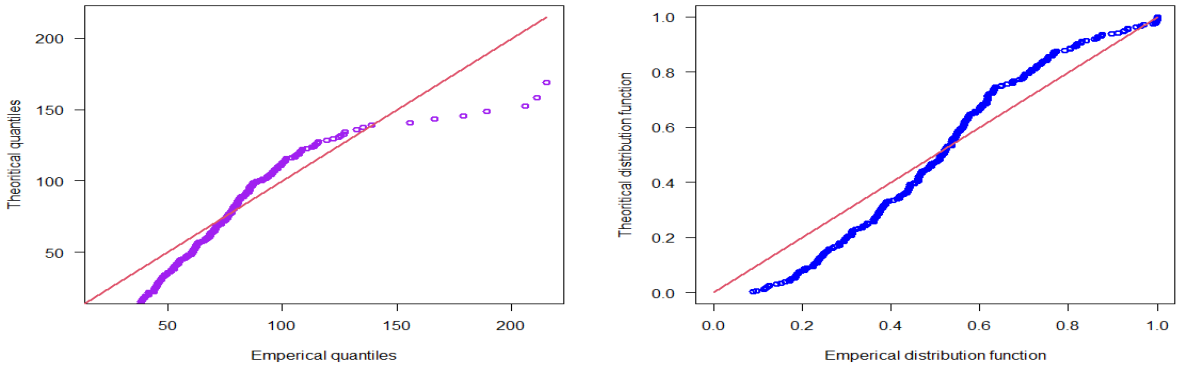

Fig. 9: Q-Q plot (left panel) and P-P plot (right panel) along with Winter seasons.

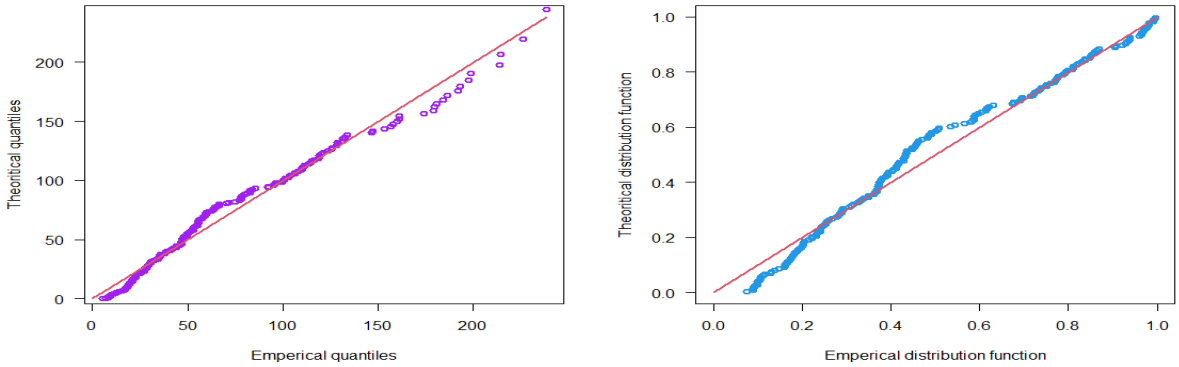

Fig. 10: Q-Q plot (left panel) and P-P plot (right panel) along with Spring seasons.

Kolmogorov-Smirnov test (D=0.082978, p-value = 0.07108) and Anderson Darling test (A_n=2.3995, p-value = 0.05602) and Cramér–von Mises test (Ω_2=0.32656, p-value 0.1139) respectively. These results indicate that the proposed distribution remains valid for the Spring season. This is supported by Q-Q plot and the P-P plot of the proposed distribution (Fig. 10).

A plot is created to compare the empirical cumulative distribution function with the theoretical cumulative distribution function. The empirical cumulative distribution represents the observed cumulative probability of the dataset, while the CDF corresponds to the cumulative probability based on the theoretical distribution. This plot allows for an evaluation of the agreement between the observed data and the theoretical distribution. Evidenced by the empirical cumulative distribution function plotted against the theoretical cumulative distribution function proposed model, it is observed that the proposed model is satisfactory and valid on both seasons (Fig. 11).

4 Discussion:

The proposed model is based on the odd function of the Chen distribution with a half logistic distribution as a generator, resulting in a unimodal distribution with a monotonically increasing and J-shaped hazard rate function. This distribution has been extensively applied in various domains, including its utilization for studying PM2.5 levels in the Kathmandu Valley to assess air quality. Additionally, researchers have successfully employed different mathematical models such as linear programming, skewed data capturing models, higher-order decision rules, and data envelopment analysis for

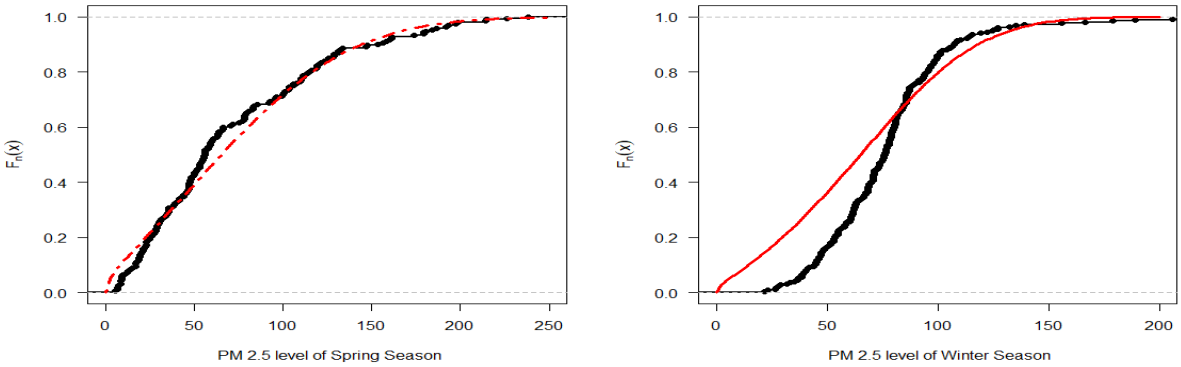

Fig. 11: Empirical cumulative distribution function verses theoretical cumulative distribution function of Spring seasons (left panel) and winter season (right panel)

studying air pollution [19]. Furthermore, fitting air pollution data has been accomplished through the use of the generalized λ distribution [20]. Likewise, a discrete-time Markov chain model has been employed to determine the quality of air in the environment [21]. Given the diverse models employed by various researchers, the development of a new model is utilized for predicting air quality in the Kathmandu Valley, Nepal, as employed by the authors. The selection of air quality data for analysis is motivated by the significance of air pollution as a critical environmental concern in the Kathmandu Valley, which encompasses the capital city of Nepal [22].

A study conducted in Kathmandu revealed that adverse weather conditions significantly contribute to the air pollution issue in the city. Out of the 30 days in a month, only seven days were deemed suitable for fresh air quality, while the remaining 23 days exhibited poor air quality. Long-term exposure to PM2.5 has the ability to penetrate deep into the lungs, corrode the alveolar wall, and impair lung functions, thereby posing a significant threat to public health [23]. Furthermore, research conducted by Lopez-Feldman et al. [24] found a positive relationship between PM2.5 air pollution and the probability of mortality after contracting COVID-19. Moreover, a nationwide investigation in China by Chen et al. [25] provided robust evidence of the associations between short-term exposure to PM2.5 and increased mortality from various cardiopulmonary diseases. The study revealed that even a $10\,\mu g/m^3$ increase in daily PM2.5 concentrations was significantly associated with a percentage increase in mortality from non-accidental causes, cardiovascular diseases, hypertension, coronary heart diseases, stroke, respiratory diseases, and chronic obstructive pulmonary disease. Similarly, Liu et al. [26] conducted a comprehensive global analysis across more than 600 cities, reinforcing the link between mortality and PM concentration. Their findings demonstrated independent associations between short-term exposure to PM2.5 and daily all-cause, cardiovascular, and respiratory mortality. Hence, the higher the exposure to PM2.5 air pollution, the greater the increase in mortality rates from various cardiopulmonary diseases. Thus, we focused on minimizing PM2.5 concentrations in order to mitigate the negative health effects associated with air pollution.

5 Conclusion

This study focuses on the development of a new distribution called the Odd Half Logistic Chen distribution. This distribution is obtained by compounding the half logistic-G family with the odd function of the Chen distribution, resulting in a negatively skewed distribution. The researchers derive important properties of this distribution, including quintile and median, mode, and order statistic. The parameters of the distribution are estimated using maximum likelihood estimation methods, and the applicability of this estimation technique is tested through simulation studies and real data analysis.

To understand the characteristics of the proposed distribution and predict air quality, the researchers analyze PM2.5 air quality data from stations in Bhaktapur, Bhaisepati, Kirtipur, and Ratnapark of Kathmandu Valley. Since there is no variation in location, the analysis is conducted based on seasonal variations. In Kathmandu, fresh air suitable for human health is available only for seven days per month. The researchers utilize the sample data to assess the efficiency of the proposed model, and different criteria are employed to validate its performance. The results indicate that the proposed model provides reasonably better predictions. Therefore, Odd Half Logistic Chen distribution offers an alternative model for predicting air quality and other environmental data.

Acknowledgement

We would like to express our deep appreciation to Research Management Cell, Birendra Multiple Campus, Tribhuvan University for providing financial support through the Mini Research Grant program. Likewise, we would also like to extend our heartfelt thanks to the Department of Environment for providing the data.
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

Author Contributions

Ramesh Prasad Tharu played a role in conceptualizing the research and developing the methodology. Govinda Prasad Dhungana utilized computer applications for data analysis and also contributed to writing the manuscript. Professor Ramesh Kumar Joshi was involved in collecting the data and editing the entire manuscript.

Conflict of interest

The author declares no conflict of interest

References

[1] Dhungana, G. P., and Kumar, V. Modified Half Logistic Weibull Distribution with Statistical Properties and Applications. International Journal of Statistics and Reliability Engineering, **8**, 29–39 (2021).
[2] Hassan, A. S., Elgarhy, M., ul Haq, M. A., and Alrajhi, S. On Type II Half Logistic Weibull Distribution with Applications. Mathematical Theory and Modeling, **9**(1)(2019). https://doi.org/10.7176/MTM/9-1-05
[3] Elgarhy, M., ul Haq, M. A., and Perveen, I. Type II Half Logistic Exponential Distribution with 37 Applications. Annals of Data Science, **6** (2), 245–257(2019). https://doi.org/10.1007/s40745-018-38 0175-y
[4] Olapade, A. K. The Type I Generalized Half Logistic Distribution. Journal of the Iranian Statistical Society, **13**(1), 69–82 (2014).
[5] Krishnarani, S. D. On a Power Transformation of Half-Logistic Distribution. Journal of Probability and Statistics, **2016** 1–10 (2016). https://doi.org/10.1155/2016/2084236
[6] Anwar, M., and Zahoor, J. The Half-Logistic Lomax Distribution for Lifetime Modeling. Journal of Probability and Statistics, **2018**, 1–12 (2018). https://doi.org/10.1155/2018/3152807
[7] Chaudhary, A. K., and Kumar, V. Half Logistic Exponential Extension Distribution with Properties and Applications. International Journal of Recent Technology and Engineering, **9** (3), 506–512(2020). https://doi.org/10.35940/ijrte.C4625.099320
[8] Aldahlan, M. A. D., and Afify, A. Z. The Odd Exponentiated Half-Logistic Exponential Distribution: Estimation Methods and Application to Engineering Data. Mathematics, **8**(10), 1684(2020). https://doi.org/10.3390/math8101684
[9] Hamedani, G. G., Korkmaz, M. C., Butt, N. S., and Yousof, H. M. The Type II Quasi Lambert Family. Pakistan Journal of Statistics and Operation Research, 963–983(2022). https://doi.org/10.18187/pjsor.v18i4.3907
[10] Dhungana, G. P., and Kumar, V. Exponentiated Odd Lomax Exponential distribution with application to COVID-19 death cases of Nepal. PloS one, **17**(6), e0269450 (2022).
[11] Chaudhary, A. K., Telee, L. B. S., Karki, M., and Kumar, V. Statistical analysis of air quality dataset of Kathmandu, Nepal, with a New Extended Kumaraswamy Exponential Distribution. Environmental Science and Pollution Research, **31**(14), 21073-21088(2024). https://doi.org/10.1007/s11356-024-32129-z
[12] WHO. WHO air quality guidelines global update 2005, Report on a Working Group meeting, Bonn, Germany(2005).
[13] Alzaatreh, A., Lee, C., and Famoye, F. A new method for generating families of continuous distributions. METRON, **71**(1), 63–79(2013). https://doi.org/10.1007/s40300-013-0007-y
[14] Lawless, J. F. Statistical models and methods for lifetime data. John Wiley and Sons.(2011).
[15] Dey, S., Kumar, D., Ramos, P. L., and Louzada, F. Exponentiated Chen distribution: Properties and estimation. Communications in Statistics - Simulation and Computation, **46**(10), 8118–8139 (2017). https://doi.org/10.1080/03610918.2016.1267752
[16] Department of Environment, Kathmandu, Nepal (GoN/MoFE,2023)
[17] Henningsen, A., and Toomet, O. maxLik: A package for maximum likelihood estimation in R. Computational Statistics, **26**(3), 443-458(2011).
[18] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023). ¡https://www.R-project.org/¿.
[19] Cooper, W. W., Hemphill, H., Huang, Z., Li, S., Lelas, V., Sullivan, D. W. Survey of mathematical programming models in air pollution management. European Journal of Operational Research, **96**(1), 1–35(1997). https://doi.org/10.1016/S0377-2217(97)86747-1

[20] Okur, M. C. On fitting the generalized λ-distribution to air pollution data. Atmospheric Environment (1967), **22**(11), 2569-2572(1988).
[21] Alyousifi, Y., Masseran, N., and Ibrahim, K. Modeling the stochastic dependence of air pollution index data. Stochastic Environmental Research and Risk Assessment, **32**(6), 1603–1611(2018). https://doi.org/10.1007/s00477-017-1443-7
[22] Zhong, M., Saikawa, E., Avramov, A., Chen, C., Sun, B., Ye, W., ... and Panday, A. K. Nepal Ambient Monitoring and Source Testing Experiment (NAMaSTE): emissions of particulate matter and sulfur dioxide from vehicles and brick kilns and their impacts on air quality in the Kathmandu Valley, Nepal. Atmospheric Chemistry and Physics, **19**(12), 8209-8228(2019)..
[23] Dhakal, S., Gautam, Y., and Bhattarai, A. Exploring a deep LSTM neural network to forecast daily PM2.5 concentration using meteorological parameters in Kathmandu Valley, Nepal. Air Quality, Atmosphere and Health, 14(1), 83–96(2021). https://doi.org/10.1007/s11869-020-00915-6
[24] López-Feldman, A., Heres, D., and Marquez-Padilla, F. Air pollution exposure and COVID-19: A look at mortality in Mexico City using individual-level data. Science of The Total Environment, 756, 143929(2021). https://doi.org/10.1016/j.scitotenv.2020.143929
[25] Chen, R., Yin, P., Meng, X., Liu, C., Wang, L., Xu, X., Ross, J. A., Tse, L. A., Zhao, Z., Kan, H., and Zhou, M. Fine Particulate Air Pollution and Daily Mortality. A Nationwide Analysis in 272 Chinese Cities. American Journal of Respiratory and Critical Care Medicine, **196**(1), 73–81(2017). https://doi.org/10.1164/rccm.201609-1862OC
[26] Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., Coelho, M. S. Z. S., Saldiva, P. H. N., Lavigne, E., Matus, P., Valdes Ortega, N., Osorio Garcia, S., Pascal, M., Stafoggia, M., Scortichini, M., Hashizume, M., Honda, Y., Hurtado-Díaz, M., Cruz, J., ... Kan, H. Ambient Particulate Air Pollution and Daily Mortality in 652 Cities. New England Journal of Medicine, 381(8), 705–715(2019). https://doi.org/10.1056/NEJMoa1817364

Ramesh Kumar Joshi completed his doctorate degree in "The analysis of some statistical models using Bayesian study" at Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India in 2018.Mr Joshi has published more than 25 papers in national and International referred journals.Mr. Joshi is Editor in chief of Nepal University Teachers' Association (NUTA) Journal since 2020. Mr joshi is serving Tribhuvan University Nepal since last 27 years under the Department of Statistics, Trichandra Multiple Campus, Ghantaghar, Kathmandu. Mr Joshi have received "National education award" by Ministry of education, science and technology, Government of Nepal in 2009, "Nepal Bidhya bhusan" award by Right honorable president of Nepal on the occasion of National education day in 2019 and "Deergh sewa padak" by Tribhuvan University on the occasion of Annual Day of Tribhuvan University in 2023.Main research area of Mr Joshi are analysis of statistical models by Bayesian study using MCMC approach and application and properties of statistical distributions.

Govinda Prasad Dhungana completed his doctorate degree in "Statistical Modeling of Survival Data" at Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, Uttar Pradesh, India in 2023. Dr. Dhungana has published more than 40 papers in national and international referred journals.Dr. Dhungana is serving Tribhuvan University Nepal since last 17 years under the Department of Statistics, Birendra Multiple Campus, Chitwan. Dr. Dhungana has received "Nepal Bidhya Bhusan Kha" award by Government of Nepal on the occasion of National education day in 2024. The main research areas of Dr. Dhungana are the development of new probability models and their use in predicting real-life phenomena, along with the development and validation of Bayesian models, and data analysis in public health, business, and social sciences.

Ramesh Prasad Tharu completed his M.Sc. Statistics in 1999, and M.B.S. degree in 2012 at Tribhuvan University, Nepal. Mr. Tharu is serving Tribhuvan University Nepal since last 22 years under the Department of Statistics, Mahendra Multiple Campus, Nepalgunj. Mr. Tharu has published more than 7 papers in national and international referred journals in the fields of applied statistics, public health, social science, statistical modelling, and probability distribution. His research interests are probability distribution, Bayesian statistics, statistical modelling, applied statistics, statistical inference, and regression models.