Nonparametric Techniques for Estimating Electric Load Probability Density

Begad Elsaid[1] and Samuel Feng (Faculty Advisor)[2]

[1]Khalifa University of Science, Technology and Research
[2]Department of Mathematics, Khalifa University

May 5, 2020

Abstract

Probability Density Estimation of stochastic electric load is of most importance nowadays in power system operations and urban planning. This is due to the continuous demand to integrate intermittent renewable energy resources that introduce uncertainties in the operating state of power systems which in turn requires accurate and reliable methods to estimate load. This paper is the first to employ a nonparametric techniques called Root Transform Local Linear Regression for estimating electric load. This robust model proposed estimates electric load data more accurately than parametric models used in current literature. The performance of the root transform local linear regression model is compared with two kernel density estimation models and two parametric models (Gaussian and Gamma distributions) and is assessed using the Kolmogorov-Smirnov goodness-of-fit test, Coefficient of determination and four error metrics. Results confirm the accuracy of the nonparametric models over the parametric models with the root transform model performing best across all error metrics and K-S test, followed by the kernel density estimation model. An interactive web application is developed to perform the same analysis presented in this paper on any type of univariate data.

INTRODUCTION

Electric Load is irregular in nature (Fig. 1) and there exists no system as of yet of storing energy on a wide scale that is cheap, sustainable, efficient and environmentally-friendly. Statistical models can be used to predict electrical power demand in a certain time frame using electrical data of domestic consumers and enterprises. Employing electric load density estimation will enable us to predict the electrical power required for the day, which we will then need to store less energy for future use and in turn, diminish the energy lost in the process of storage. Further applications of electric load density estimation include planning studies for optimal allocation of renewable distributed generation to minimize annual energy loss [1], evaluating the reliability and accuracy of power systems with PV power generation [2], integrating of batteries with photovoltaic plants on a large scale [3].

In literature, electric load probability is usually modeled using parametric models, such as the Gaussian distribution [4] [5] and the beta distribution [3]. However, through careful inspection of electric load, it would not be accurate to model electric load using a Gaussian distribution nor a Beta distribution due to its bimodal nature [6] (Fig. 5). In this paper, two nonparametric approaches, Root Transform Local Linear Regression (RTLLR) and Kernel Density Estimation (KDE), are proposed to provide an accurate model of electric load probability density functions. The former turns the probability estimation problem into a regression problem through binning and variance stabilizing transformation of electric load data. The latter assigns each data observation a certain weight where more populated intervals end up having higher probability density function values. The performances of RTLLR and two KDE models with Rule-of-Thumb bandwidth selectors are compared to two parametric models, Gaussian and Gamma Distributions, and are evaluated through the Kolmogorov-Smirnov goodness-of-fit test, Coefficient of Determination (R^2) and four error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Biased Error (MBE).

The major contributions of this work can be summarized as follows:

1. For the first time, RTLLR is proposed for estimating electric load data. It is based on converting the probability density estimation problem into a regression problem that is solved by local linear regression. The resulting procedure is easy to use and computationally efficient.
2. We show that the RTLLR model avoids the boundary bias present in KDE models and is less sensitive to outliers in the data. The RTLLR model visually follows the shape of electric load distribution and has the lowest error metric values and highest R^2 values when compared to the KDE models and the Gaussian and Gamma models.
3. An interactive web application has been built to equip users with all the tools to replicate the analysis presented in this paper on any type of univariate data.

The remainder of the paper is organized as follows. A detailed description of the problem and electric load data is presented in Section 2. Then, the statistical model, assessment methods, and developed software are described in Section 3. Results are discussed in Section 4 and the paper then closes with a conclusion and acknowledgments.

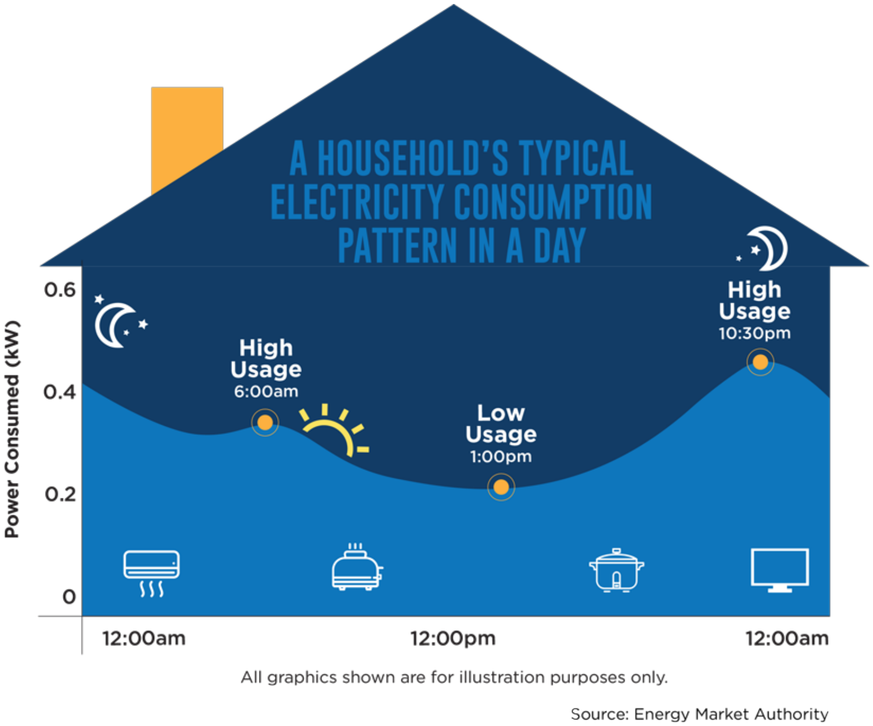


Figure 1: Irregularity of Electricity Consumption, or Electrical Load, throughout the day [7].

2. DETAILED DESCRIPTION OF PROBLEM

2.1 Electric Load Data

Data provided by the Customer-Led Network Revolution [8] contains the power consumption over 30-minute intervals for the years 2011 to 2013 for 908 enterprise and 1538 residential locations in the UK. The enterprises are divided into 4 main sectors which are Public Sector & Other (PSO), Commercial Offices (CO), Industrial (IND) and Agriculture, Hunting, Farming & Fishing Enterprises (AHFF). Those sectors are then divided further depending on the number of employees, the tariff rate and whether the enterprise is a single-site or multi-site. Residential Locations are also divided into sectors called Mosaic classes, devised by Experian [9], which takes into account income, age, location and other characteristics of residents. The divisions of enterprise and residential locations are visualized in Fig. 2 and Fig. 3.

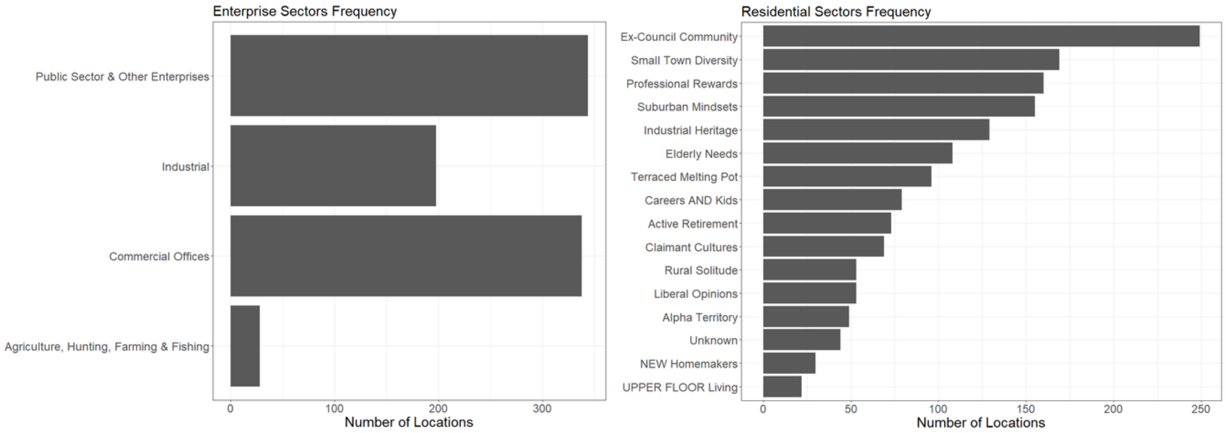


Figure 2: Left Panel: Number of enterprises analyzed based on their sector. Right Panel: Number of Residential locations analyzed based on their mosaic class.

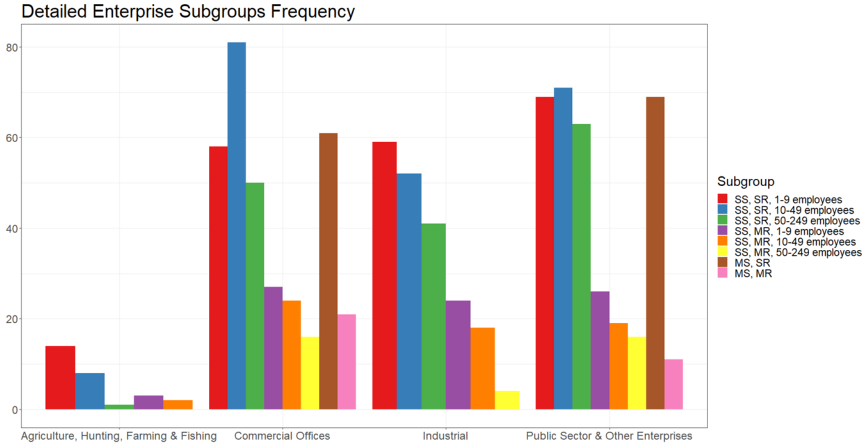


Figure 3: Detailed Number of Enterprises analyzed based on enterprises' sector subgroups. Subgroups are divided based on number of employees (1-9, 10-49, or 50-249 employees), tariff rate (Single-Rate (SR) or Multi-Rate (MR)) and size of site (Single-Site (SS) or Multi-Site (MS)).

2.2 The Statistical Problem

Electric load data is collected as a time series with n observations $X_1, X_2, \ldots, X_n$. In the case of the electric load data in subsection 2.1, a yearly collection of electric load data yields 17520 electric load observations, denoted as $X_1, X_2, \ldots, X_{17520}$. Power system planning and optimization require accurate probability estimation at random points in the future during the life of a power network. In other words, let X denote a random variable whose value is the electric load at some point in the future. The goal is to use the data to estimate the value of $\mathbb{P}(a < X < b)$ for various values of a and b [10].

The goal of estimating $\mathbb{P}(a < X < b)$ is a well-known problem in the statistics field [11] [12] which supposes that X, the electric load data at some point in the future, has a probability density function f_X, and that the data$X_1, X_2, \ldots, X_n$ are n independent, identically distributed observations drawn from the same distribution f_X. The probability density function f_X of electric load is unknown and our goal is to find or build a function $\hat{f_X}$ from the data in order to estimate the unknown f_X. After finding a good enough estimate $\hat{f_X}$, $\mathbb{P}(a < X < b)$ reduces to a numerical integration problem of $\hat{f_X}$ over the interval (a, b).

3. IMPLEMENTATION AND METHODS

In this section, we define and describe five statistical models used to estimate electric load PDF in this paper, explore methods to assess models' performances as well as introduce an interactive web application for users to perform the analysis done in this paper.

3.1 Statistical Models

Here we describe the 3 types of statistical models we consider: Parametric Estimation, Kernel Density Estimation, Root Transform Local Linear Regression. Parametric Models are used in previous work [4] [5]. KDE is commonly used in related studies on renewable energy [13]. RTLLR proving more recent method show to be successful on renewables.

3.1.1 Parametric Estimation

Parametric models are widely used in estimating probability density functions for their ease of use. They require an investigator to assume that the data comes from a certain distribution defined by a finite number of parameters and that it follows a specific shape. Therefore, they are considered to have high bias since the investigator has to make that assumption which may not be true. In Fig. 5, it can be seen that electric load may be bimodal in nature and thus the Gaussian distribution used in literature would not be a good fit. In this report, we consider two parametric distributions, Gaussian and Gamma, and compare their performance with nonparametric techniques.

The PDF of the Gaussian Distribution is:

$$f_{\text{Gaussian}} = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where $x \in (0, \infty)$, μ is the mean and σ is the standard deviation. The PDF of the Gamma distribution is:

$$f_{\text{Gamma}} = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x}$$

where Γ is the Gamma function, $x \in (0, \infty)$ and α, β are the shape and rate parameter, respectively.

These models are fit using Maximum Likelihood Estimation which is the most common method used to estimate parameters for parametric models.

3.1.2 Kernel Density Estimation

KDE is one of the most popular nonparametric estimation methods. The purpose of this technique is to estimate the unknown probability density directly from the data without making any assumptions on the shape of the distribution that parametric distributions make. The estimate $\hat{f}_{\text{KDE}}$ of the unknown density f_X is constructed from n observed data points as follows:

$$\hat{f}_{\text{KDE}} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - X}{h}\right)$$

where $X_1, \ldots, X_n$ are the n observed data points, $h \in (0, \infty)$ is the bandwidth parameter, and the kernel function K is a nonnegative function with $\int K = 1$

In order to build the estimate $\hat{f}_{\text{KDE}}$, an investigator must decide on the kernel function and the bandwidth parameter to be used. Research shows that different choices of kernel function have no significant difference on the fit of the data while the selection of bandwidth is of great importance in the build of $\hat{f}_{\text{KDE}}$ [12] [14]. In this paper, two common rule-of-thumb formulas are used to calculate the bandwidths for the KDE models:

$$h_{ROT1} = 1.059 \times \hat{\sigma} \times n^{-\frac{1}{5}}$$
$$h_{ROT2} = \hat{\sigma} \times n^{-\frac{1}{6}}$$

A limitation of KDE is boundary bias where it underestimates data that is close to the boundaries. This may pose significant problems when fitting KDEs to electric load data since the bulk of the data do not lie around the center of the range but rather near the boundaries.

3.1.3 Root Transform Local Linear Regression

Root Transform Local Linear Regression (RTLLR) is a nonparametric technique that aims to turn probability density estimation into a nonparametric regression problem. The original method proposed in statistics literature [15] aims to decrease the bias from choosing parametric models and applies a transformation in order to stabilize the variance of the data. RTLLR also avoids the issue with boundary bias that comes with KDE estimation [11]. In this subsection, we introduce the motivation behind the RTLLR model, build the foundation and reasons why the model works and present the implementation of the model.

MOTIVATION

The motivation behind RTLLR is improving on the histogram's estimate of the pdf. Let $X_1, X_2, ...X_n$ be the univariate data with pdf f_X. Without loss of generality, we assume that this data has been normalized to $[0, 1]$. Let T be a positive integer such that $T \approx \frac{n}{10}$ [11]. Bin the data into T equal length intervals on the unit interval and let Q_i be the number of observations that fall in each subinterval $I_i = \left[\frac{i-1}{T}, \frac{i}{T}\right)$. Then, the joint distribution of the Q_i's is multinomial $Multi\,(n, p_1, ..., p_T)$ where p_i is equal to the probability that some data point X_k will fall in the interval I_i ($p_i = \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx$). In order to understand the relationship between the Q_i's and the pdf f, we will need to know the marginal distributions of the Q_i's. We present two arguments for why $Q_i \sim Poisson(np_i)$.

MATHEMATICAL FOUNDATIONS OF RTLLR

1. **Large Sample Size.** The marginal distributions of the Q_i's are $Binomial\,(n, p_i)$. However, as sample size n goes to infinity, the interval of I_i converges to a single point and thus p_i is driven to 0. We will show that as $n \to \infty, Q_i \overset{D}{\to} Poisson(10f(x_i))$.

Theorem 1 *Let n be the data sample size and $Q_i \sim Binomial\,(n, p_i)$. Then, as $n \to \infty, Q_i \overset{D}{\to} Poisson(10f(x_i))$ where $x_i = \frac{i-1}{T}$ is the left endpoint of the interval I_i.*

Proof. Assume that f is continuous. We first show that as $n \to \infty$, $\lambda_i = np_i$ converges to $10f(x_i)$.

$$\begin{aligned}
\lim_{n\to\infty} \lambda_i &= \lim_{n\to\infty} np_i \\
&= \lim_{n\to\infty} n \int_{\frac{i-1}{T}}^{\frac{i}{T}} f(x)dx \\
&= \lim_{T\to\infty} 10T \left(F\left(\frac{i}{T}\right) - F\left(\frac{i-1}{T}\right)\right) \\
&= \lim_{T\to\infty} 10 \frac{F\left(\frac{i}{T}\right) - F\left(\frac{i-1}{T}\right)}{\frac{1}{T}} \\
&= \lim_{\frac{1}{T}\to 0} 10 \frac{F(x_i + \frac{1}{T}) - F(x_i)}{\frac{1}{T}} \\
&= 10f(x_i)
\end{aligned}$$

We now show that as $n \to \infty, Q_i \overset{D}{\to} Poisson(10f(x_i))$.

$$\begin{aligned}
\lim_{n\to\infty} P(Q_i = k) &= \lim_{n\to\infty} \binom{n}{k} p_i^k (1-p_i)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda_i}{n}\right)^k \left(1 - \frac{\lambda_i}{n}\right)^{n-k} \\
&= \left(\lim_{n\to\infty} \frac{\lambda_i}{k!}\right) \left(\lim_{n\to\infty} \frac{n!}{n^k(n-k)!}\right) \left(\lim_{n\to\infty} \left(1-\frac{\lambda_i}{n}\right)^{-k}\right) \left(\lim_{n\to\infty} \left(1-\frac{\lambda_i}{n}\right)^{n}\right) \\
&= \left(\frac{10f(x_i)^k}{k!}\right)(1)(1)\left(e^{-10f(x_i)}\right) \\
&= \frac{10f(x_i)^k}{k!} e^{-10f(x_i)}
\end{aligned}$$

Thus, as $n \to \infty, Q_i \overset{D}{\to} Poisson(10f(x_i))$.

From the above argument, we can deduce that for a large sample size n, Q_i can approximated by $Poisson(np_i)$.

2. **Poissonization**. Assume that the sample size N is random and Poissoned (i.e $N \sim Poisson\,(n)$ and N is independent of X_i). Then, the marginal distributions of the Q_i's end up being Poisson by the following theorem.

Theorem 2 *Let $N \sim Poisson(n)$ and $Q_1, Q_2, ...Q_T \sim Multi(N, p_1, p_2, ..., p_T)$. Then, the marginal distributions of the Q_i's are $Poisson(np_i)$ and are independent [16].*

By the two arguments presented above, $Q_i \sim Poisson(np_i)$ where $np_i = n\int_{\frac{i-1}{T}}^{\frac{i}{T}} f\,(x)\,dx \;\approx\; \frac{n}{T} f\left(\frac{2i-1}{2T}\right) = 10f\,(\hat{x_i})$ and $\hat{x_i}$ is the center of the interval I_i.

The Q_i's estimate the underlying distribution f_X. Therefore, it is possible to build f_X through performing nonparametric regression between the Q_i's and the center points of the intervals I_i. However, an issue arises because the variances of the Q_i's are not equal ($Var\left(Q_i\right) = np_i$). This would violate the major assumption of homoscedasticity required by many nonparametric regression techniques [17] [11]. In order to combat the heteroscedasticity of the Q_i's, a transformation $g\left(Q_i\right)$ is applied to Q_i that would fulfill two major objectives:

1. **Minimize Bias.** The expected value of $g\left(Q_i\right)$ should be deterministically related to the expected value of Q_i. This is essential since $g\left(Q_i\right)$ would enable the estimation of mean of Q_i which in turn leads to an estimate of the unknown pdf, as shown above. In other words, we need to find a transformation that would minimize the bias of the estimate of f_X.

2. **Approximate Homoscedasticity.** All $g\left(Q_i\right)$ random variables should have the same constant variance effectively turning the problem into a homoscedastic regression problem.

Brown [15] found that the transformation $\sqrt{Q_i + c}$ achieves both of those just goals. We present the result here for completeness:

Theorem 3 *Let $X \sim Poisson(\lambda)$ with $\lambda > 0$ and let $c \geq 0$ be a constant. Then,*

$$E(\sqrt{X+c}) = \lambda^{1/2} + \frac{4c-1}{8}\lambda^{-1/2} - \frac{16c^2-24c+7}{128}\lambda^{-3/2} + O(\lambda^{-5/2})$$

$$Var(\sqrt{X+c}) = \frac{1}{4} + \frac{3-8c}{32}\lambda^{-1} - \frac{32c^2-52c+17}{128}\lambda^{-2} + O(\lambda^{-3})$$

Proof. We first apply the Taylor series expansion of the function $g(X) = \sqrt{X+c}$ around the constant $\lambda - c$.

$$g(X) = g(\lambda-c) + g'(\lambda-c)(X-(\lambda-c)) + \frac{g''(\lambda-c)}{2!}(X-(\lambda-c))^2 + \frac{g''(\lambda-c)}{3!}(X-(\lambda-c))^3 + \dots$$

$$\sqrt{X+c} = \lambda^{1/2} + \frac{1}{2}\lambda^{-1/2}(X-\lambda+c) - \frac{1}{8}\lambda^{-3/2}(X-\lambda+c)^2 + \frac{1}{16}\lambda^{-5/2}(X-\lambda+c)^3 - \frac{15}{128}\lambda^{-7/2}(X-\lambda+c)^4 + \dots$$

$E(\sqrt{X+c})$ can then be found by applying expectation on the previous equation and using
1. **Binomial expansion**: $(X-\lambda+c)^n = \sum_{k=0}^{n}\binom{n}{k}(X-\lambda)^{n-k}c^k$
2. **Central moment recursive formula for the Poisson Distribution**: $\mu_{r+1} = \lambda\left(\frac{d\mu_r}{d\lambda} + r\mu_{r-1}\right)$ where $\mu_r = E\left((X-\lambda)^r\right)$ is the r^{th} central moment of X [18].

$Var(\sqrt{X+c})$ can then be found by using the formula $Var(\sqrt{X+c}) = E(X+c) - E(\sqrt{X+c})^2$.

By Theorem 3, the expectation of $\sqrt{Q_i + c}$ has approximately the square root of the expectation of Q_i which establishes the first objective. Furthermore, the variance of all the $\sqrt{Q_i + c}$'s is approximately $\frac{1}{4}$ establishing the second objective. Therefore, the final goal is to choose the constant c so that both approximations are reduced. The best candidates for the constant c are $c = \frac{1}{4}$ to minimize the first order bias or $c = \frac{3}{8}$ which would minimize the difference in variance between $g\left(Q_i\right)$'s. It is at this point that there is a payoff between minimizing the bias and stabilizing the variance. However, Brown [15] shows, through visual plots of bias and variance vs. lambda, that minimizing the bias at $c = \frac{1}{4}$ outweighs the loss in stability in variance. Therefore, the transformation $\sqrt{Q_i + \frac{1}{4}}$ is chosen.

IMPLEMENTATION

After building a transformation, we can now go to the RTLLR implementation. Building the estimate $\hat{f}_{\mathrm{RTLLR}}$ of the unknown density f_X can be summarized into five steps [11] [19]:

1. **Binning**. Electric load data is divided into $T \approx n/10$ bins, where n is the number of data observations. Let $Q_1, \ldots, Q_T$ denote the positive integer corresponding to the number of observations in each bin, and $x_1,\ \ldots, x_T$ represent the centers of each of the T bins.
2. **Variance Stabilizing Root Transform**. Calculate $y_i = \sqrt{\frac{1}{10}} \cdot \sqrt{Q_i + \frac{1}{4}}$, thus yielding a new paired data set with T observations: $(x_1,\ y_1),\ \ldots, (x_T,\ y_T)$.
3. **Nonparametric Regression**. Any nonparametric regression can then be used on the new paired data $(x_1,\ y_1)$, $\ldots, (x_T,\ y_T)$. We elect to use *local linear regression*. This will build a regression function $\hat{r}(x)$ where $\hat{r}(x)^2$ is an estimate of the PDF f_X. We have used local linear regression because of its efficiency and accuracy in regression modelling.
4. **Unroot**. Reverse the root transform by squaring the function to obtain $\hat{f}_u(x) = \hat{r}(x)^2$
5. **Normalize**. To ensure the estimator is a PDF (i.e. the estimator integrates to 1), we normalize $\hat{f}_{\mathrm{RTLLR}}$ so that $\hat{f}_{\mathrm{RTLLR}} = \frac{\hat{f}_u(x)}{\int_0^1 \hat{f}_u(x)\mathrm{dx}}$.

3.2 Methods for Assessing Model Performance

In this subsection, performance assessment methods are explored in order to evaluate the models presented in subsection 3.1.

3.2.1 Error Metrics

In renewable energy research, some commonly used error metrics to assess the performance of PDF models are the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Biased Error (MBE) [20] [21]. Each of these errors measures different characteristics of fit, but all serve the same purpose. The error metrics range from $[0, \infty)$, with the exception of MBE which can be any real number, with 0 signifying that the model is a perfect fit for the data. The error metrics formulas are presented down below.

$$RMSE = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(p_i - \hat{p}_i)^2}$$

$$MAE = \frac{1}{t}\sum_{i=1}^{t}|p_i - \hat{p}_i|$$

$$MAPE = \frac{1}{t}\sum_{i=1}^{t}\left|\frac{p_i - \hat{p}_i}{y_i}\right|$$

$$MBE = \frac{1}{t}\sum_{i=1}^{t}(p_i - \hat{p}_i)$$

where t is the number of bins of data chosen using the Freedman-Diaconis Rule [22], p_i is the probability of electric load being within bin i calculated from the data set, and $\hat{p}_i$ is the probability within the same bin calculated from the estimated data set which found by integrating the model within bin i.

3.2.2 Coefficient of Determination (R^2)

Coefficient of determination R^2 is a statistic that calculates the proportion of variance of the data explained by a model. Ideally, a model with a perfect fit of the data completely captures the variance of the data and would have an R^2 of 1. Another interpretation of R^2 is how well the model does relative to a constant model with the value of the data mean. A negative value of R^2 signifies that the model studies is worse than a model a constant model with the value of the data mean through out the whole domain. The formula for R^2 is presented below.

$$R^2 = 1 - \frac{\sum_{i=1}^{t} (p_i - \hat{p_i})}{\sum_{i=1}^{t} (p_i - \bar{p_i})} \ , \ \bar{p_i} = \frac{1}{t} \sum_{i=1}^{t} p_i$$

3.2.3 Kolmogorov Smirnov Test

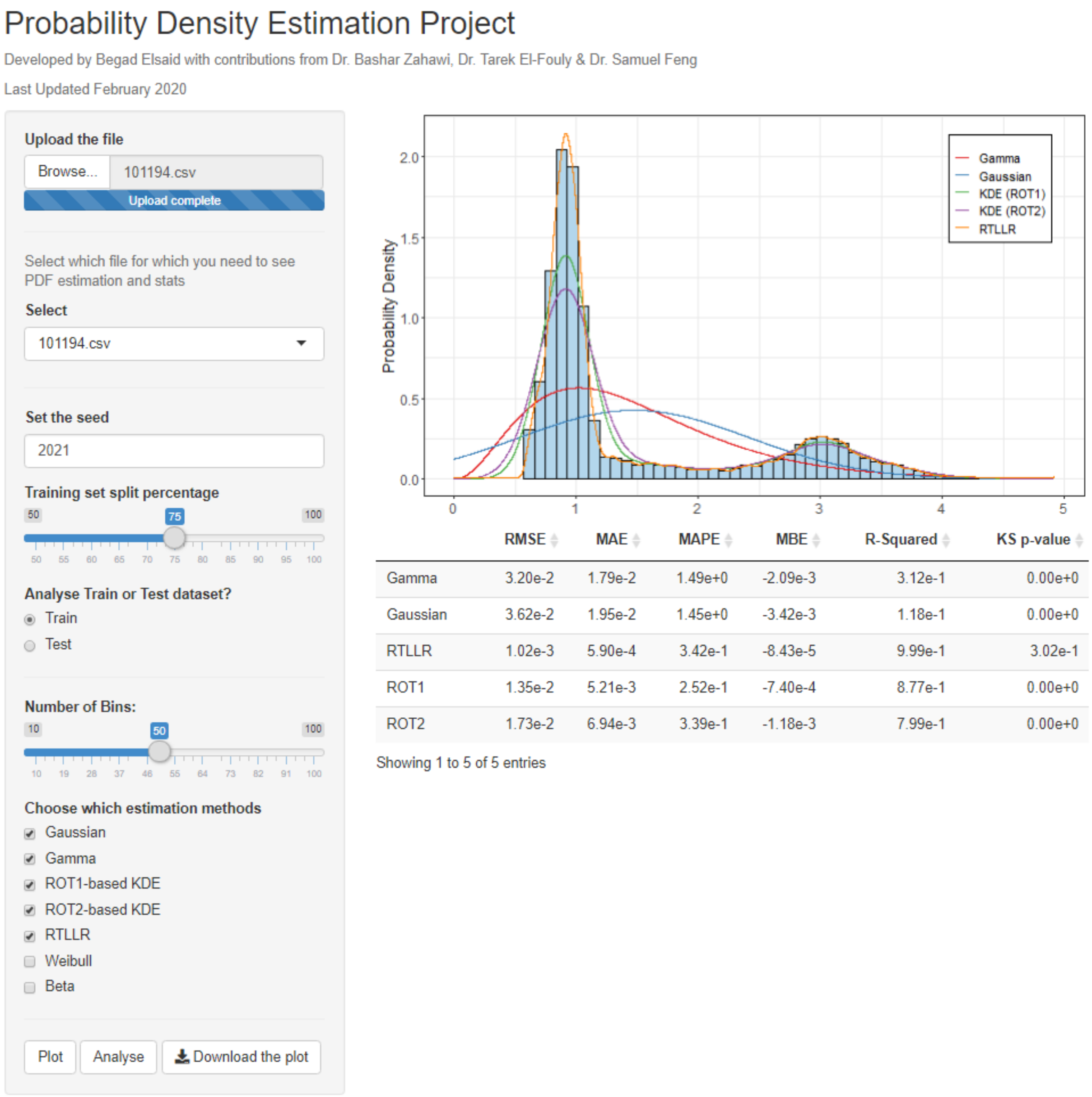
Another interesting error metric we explored is the One-sample Kolmogorov-Smirnov (KS) test. Kolmogorov-Smirnov tests if data is distributed according to a specific model. The KS test is done by finding the supremum distance statistic which is calculated by finding the difference between the data's Empirical CDF, an estimator of the data's CDF, is to the model's CDF. Then, the KS test p-value can be calculated using the supremum distance statistic which has an asymptotic CDF given by the KS function [23]. There is evidence that a model is a good fit for the data if the KS test's p-value is larger than the threshold α where we consider a threshold of $\alpha = 0.01$ in this report.

3.3 Data Splitting

A common practice in statistics that is not present in power systems engineering systems field is data splitting. When estimating the PDF of data to accurately test a model's fitness on the data, one should first split the data into two datasets test and train. Then, one fits their model on the training dataset and test how good the model is on the test set. This is because if a model is a good fit for the current data, it doesn't necessarily mean it is a good fit for other data. Therefore, we consider the test dataset as other data that we use to measure the model's quality. What exactly do we do? The train and test datasets were split randomly according to a 75%:25% split in all analysis in this paper.

3.4 Developed Software

We built an interactive web application on R Shiny called PDEP (Probability Density Estimation Project) which enables users to perform the same analysis that was presented in this paper for any type of univariate data. It is equipped with a wide array of features from Data Splitting, where the user can split their data into training and test datsets, as well as setting the seed and train split percentage. Furthermore, users can also plot histograms of the data with customizable binwidth and fit the data to multiple parametric models (Gaussian, Gamma, Beta & Weibull) as well as nonparametric models (KDE with ROT1 & ROT2 bandwidths and RTLLR). Moreover, the user is able to download the plots with their preferred models of fit as well as view the assessment metrics introduced in this paper which are the RMSE, MAE, MAPE, MBE, R^2 and KS p-value. Analysis of Figure 5, Table 1 and Table 2 were done through the application and Fig. 4 displays a screenshot design of the application with its current features. The application will be released once our team explores potential patents and/or journal publications.



Probability Density Estimation Project

Developed by Begad Elsaid with contributions from Dr. Bashar Zahawi, Dr. Tarek El-Fouly & Dr. Samuel Feng

Last Updated February 2020

Upload the file

Browse... 101194.csv

Upload complete

Select which file for which you need to see PDF estimation and stats

Select: 101194.csv

Set the seed: 2021

Training set split percentage: 75

Analyse Train or Test dataset?

● Train

○ Test

Number of Bins: 50

Choose which estimation methods

☑ Gaussian

☑ Gamma

☑ ROT1-based KDE

☑ ROT2-based KDE

☑ RTLLR

□ Weibull

□ Beta

	RMSE	MAE	MAPE	MBE	R-Squared	KS p-value
Gamma	3.20e-2	1.79e-2	1.49e+0	-2.09e-3	3.12e-1	0.00e+0
Gaussian	3.62e-2	1.95e-2	1.45e+0	-3.42e-3	1.18e-1	0.00e+0
RTLLR	1.02e-3	5.90e-4	3.42e-1	-8.43e-5	9.99e-1	3.02e-1
ROT1	1.35e-2	5.21e-3	2.52e-1	-7.40e-4	8.77e-1	0.00e+0
ROT2	1.73e-2	6.94e-3	3.39e-1	-1.18e-3	7.99e-1	0.00e+0

Showing 1 to 5 of 5 entries

Figure 4: Screenshot of the PDEP app reproducing Fig. 5 left panel and Tables 1 analysis.

4. RESULTS

In this section, the performance of the Gaussian and Gamma distributions are analyzed alongside the Root Transform Local Linear Regression method and two Kernel Density models with 2 different bandwidths. First, electric load data from a commercial enterprise site with single-tariff is fit with the models and the models are assessed. We, then, repeat the analysis of electric load data from all 2446 enterprise and residential locations.

Fig. 5 shows the histograms for both train and test electric load datasets fit with all 5 models. Visually the RTLLR model seems to fit the data best followed by the two KDE models & then parametric distributions. This finding is also supported by the results in Tables 1 and 2 where the following observations can be made:

1. RMSE and MAE values are lower for the RTLLR model followed by the KDE models and lastly parametric models.
2. The R^2 values for RTLLR are the highest followed by the KDE models then the parametric distributions
3. RTLLR is the only model that fails to reject the null hypothesis for the KS test.

In the remainder of this section, the performance of the parametric and nonparametric models is evaluated by the test RMSE (Fig. 6), test R^2 values (Fig. 7) and Train KS test p-values (Fig. 8) of all enterprise and residential sites.

The left panel of Fig. 6 presents the average scores of each model according to a scoring system from 1 to 5, 1 for the model with the highest Test RMSE and 5 the one with the least. The right panel of Fig. 6 displays a box plot for the relative percentage test RMSE improvement of the RTLLR, two KDE and Gamma models with the Gaussian model. The results show that RTLLR almost always outperforms all models (score = 4.89) and has a significant average relative percentage improvement (around 80%) with respect to the Gaussian distribution. The 2 KDE models also show promising results with average relative percentage improvement (around 50%) to the Gaussian distribution. However, the Gamma distribution while overall outperforming the Gaussian distribution, around 30% average percentage improvement, seems to be unreliable as it underperforms significantly in multiple locations, as seen in the numerous outliers below the bottom whisker.

Fig. 7 presents similar visual plots as Fig. 6 but with test R^2 values rather than test RMSE values. Results from the left panel of Fig. 7 shows that the RTLLR method outperforms all other techniques in most locations followed then by the KDE models and finally the parametric models. The right panel of Fig. 7 visually shows the reliability of the RTLLR method in explaining the data's variance since the mean R^2 value is almost 1 and it has the narrowest interquartile range of R^2 values among all other models. Moreover, the KDE models, $\mu_{R^2} \approx 0.77$, seem to perform significantly better than both the Gamma, $\mu_{R^2} \approx 0.70$, and Gaussian distributions, $\mu_{R^2} \approx 0.35$. In addition, there are also instances where the parametric models attain negative R^2 values signifying that a constant mean model would provide a better fit for the data.

Fig. 8 shows the train KS test p-values of all enterprise and residential sites. The figure suggests that electric load data does not seem to be distributed by a Gaussian nor a Gamma distribution since they reject the KS test null hypothesis for all sites (i.e. $p < 0.01$). The 2 KDE models do seem to be a good fit for a small number of locations. However, the RTLLR looks more promising as it does well in a good number of the locations and attains the highest p-values in our study.

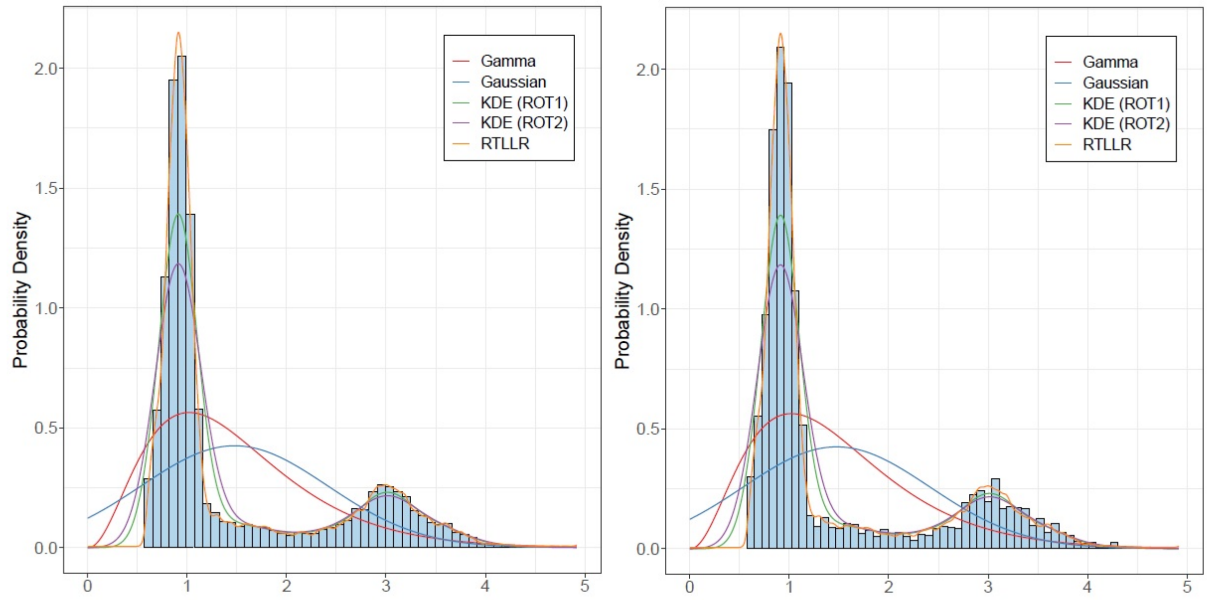


Figure 5: Histograms of electric load data from a commercial site with single tariff with parametric and nonparametric models fit. The training dataset is presented on the left while the test dataset is on the right.

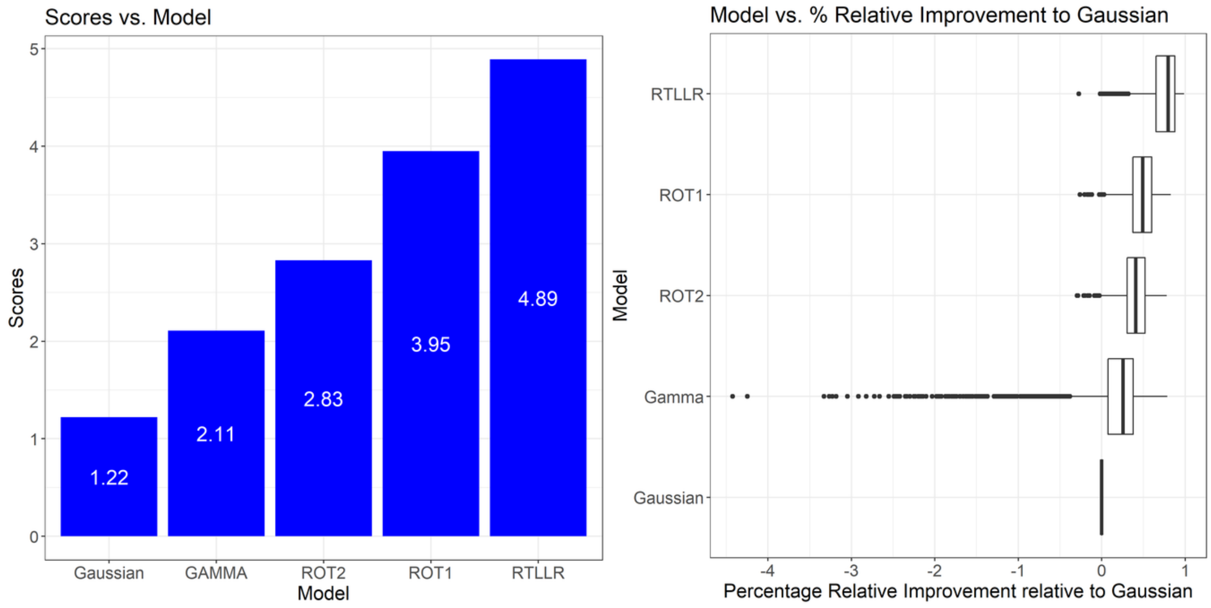


Figure 6: Left Panel: Average scores of each model from worst to best relative to Test RMSE. Right Panel: Percentage Improvement of each model vs. the Gaussian Distribution. Error metric analyzed is the Test RMSE.

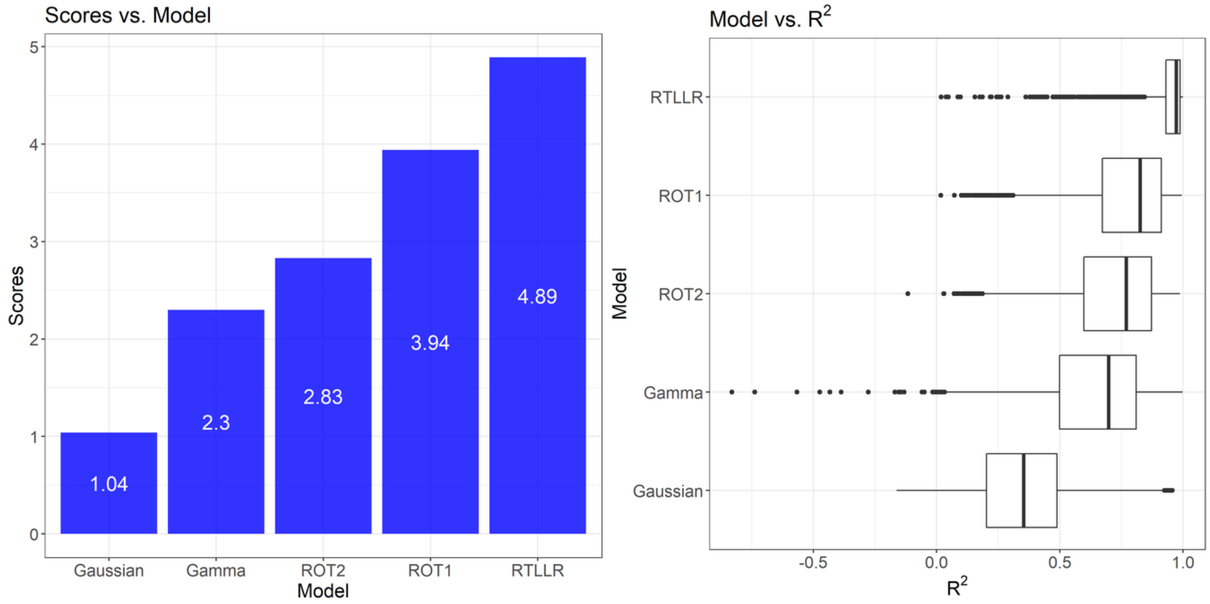


Figure 7: Left Panel: Average scores of each model from worst to best relative to Test R^2. Right Panel: R^2 values of each non-parametric method vs. the Gaussian Distribution. Error metric analyzed is the Test R^2.

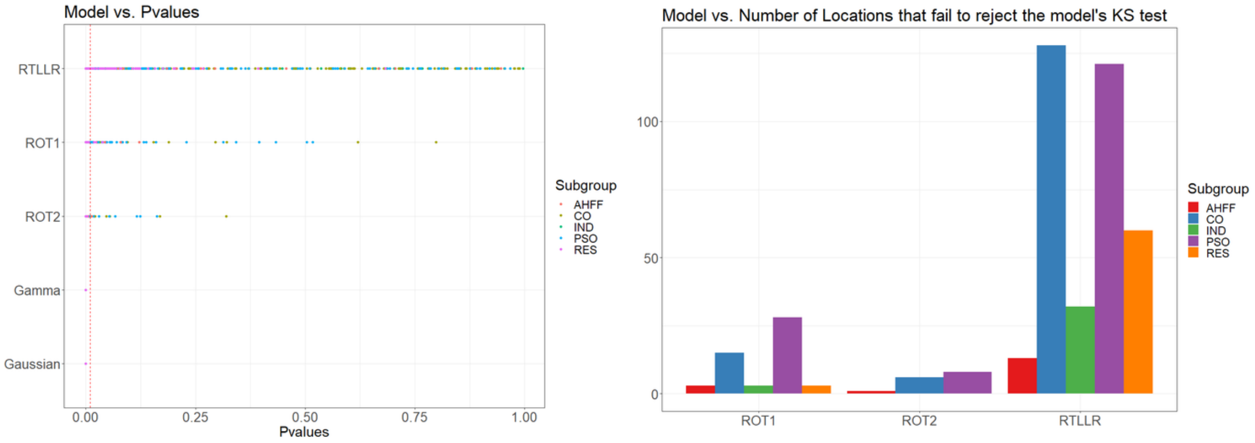


Figure 8: Left Panel: KS test p-values of all locations analyzed by subgroup. Right Panel: Number of locations that were fit well by each model, pass the p-value = 0.01 threshold, as suggested by the KS test.

	RMSE	MAE	MAPE	MBE	R^2	KS p-value
Gamma	3.20×10^{-2}	1.79×10^{-2}	1.49	-2.09×10^{-3}	3.12×10^{-1}	$< 2.2 \times 10^{-16}$
Gaussian	3.62×10^{-2}	1.95×10^{-2}	1.45	-3.42×10^{-3}	1.18×10^{-1}	$< 2.2 \times 10^{-16}$
RTLLR	1.02×10^{-3}	5.90×10^{-4}	3.42×10^{-1}	-8.43×10^{-5}	9.99×10^{-1}	3.02×10^{-1}
ROT1	1.35×10^{-2}	5.21×10^{-3}	2.52×10^{-1}	-7.40×10^{-4}	8.77×10^{-1}	$< 2.2 \times 10^{-16}$
ROT2	1.73×10^{-2}	6.94×10^{-3}	3.39×10^{-1}	-1.18×10^{-3}	7.99×10^{-1}	$< 2.2 \times 10^{-16}$

Table 1: Train electric load dataset assessment for the location in Fig. 5

	RMSE	MAE	MAPE	MBE	R^2	KS p-value
Gamma	3.07×10^{-2}	1.80×10^{-2}	1.45	-2.18×10^{-3}	2.86×10^{-1}	$< 2.2 \times 10^{-16}$
Gaussian	3.47×10^{-2}	1.99×10^{-2}	1.65	-3.46×10^{-3}	8.65×10^{-2}	$< 2.2 \times 10^{-16}$
RTLLR	1.99×10^{-3}	1.33×10^{-3}	1.37×10^{-1}	-1.54×10^{-4}	9.97×10^{-1}	3.86×10^{-1}
ROT1	1.33×10^{-2}	5.71×10^{-3}	2.63×10^{-1}	-7.88×10^{-4}	8.66×10^{-1}	$< 2.2 \times 10^{-16}$
ROT2	1.69×10^{-2}	7.18×10^{-3}	3.26×10^{-1}	-1.23×10^{-3}	7.84×10^{-1}	$< 2.2 \times 10^{-16}$

Table 2: Test electric load dataset assessment for the location in Fig. 5

CONCLUSION

Two nonparametric models Root Transform Local Linear Regression and Kernel Density Estimation are proposed for estimating electric load PDF over the Gaussian distribution used in literature to improve the accuracy of electric load modeling. The performance of the nonparametric techniques was compared alongside the Gaussian and Gamma distribution and assessed using electric load data from over 2400 enterprise and residential locations in the United Kingdom using RMSE, R^2, Kolmogorov-Smirnov test and data splitting. Root Transform Local Linear Regression had the best results across the board with the lowest Test RMSE values and with the most locations producing p-values greater than 0.01 when conducting the KS test followed by Kernel Density Estimation. The parametric distributions had overall the highest RMSE values per location and the KS test null hypothesis was rejected for all locations using those models. Further research areas would investigate the performance of RTLLR in power systems planning and optimization studies for predicting stochastic load.

ACKNOWLEDGMENTS

I would like to thank Dr. Samuel Feng for giving me the opportunity to work with him in and to develop my coding, writing and presentation skills as well as my statistical knowledge where he always provided me with support when I needed it.

Secondly, I would like to thank Dr. Alliyah Anderson for providing a different perspective to writing and motivation to improve my writing skills as well as Dr. Bashar Zahawi and Dr. Tarek al Fouly for guidance with finding the load data and selecting the appropriate locations for my data and Dr. Maisam Wahbah for helping with my code and setting up the Kolmogorov–Smirnov test. I would further like to thank Dr. Adriana Gabor for being the second reader ofthis report.

Finally, I would like to thank all my professors, friends and most importantly my parents and sister for providing me with happiness, support and motivation every day.

References

[1]Y. M. Atwa, E. F. El-Saadany, and A.-C. Guise, "Supply Adequacy Assessment of Distribution System Including Wind-Based DG During Different Modes of Operation", *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 78–86, Feb. 2010, doi: 10.1109/tpwrs.2009.2030282.

[2]G. Li, W. Lu, J. Bian, F. Qin, and J. Wu, "Probabilistic Optimal Power Flow Calculation Method Based on Adaptive Diffusion Kernel Density Estimation", *Frontiers in Energy Research*, vol. 7, Nov. 2019, doi: 10.3389/fenrg.2019.00128.

[3]N. M. Nor, A. Ali, T. Ibrahim, and M. F. Romlie, "Battery Storage for the Utility-Scale Distributed Photovoltaic Generations", *IEEE Access*, vol. 6, pp. 1137–1154, 2018, doi: 10.1109/access.2017.2778004.

[4]X. Guo, R. Gong, H. Bao, and Q. Wang, "Hybrid Stochastic and Interval Power Flow Considering Uncertain Wind Power and Photovoltaic Power", *IEEE Access*, vol. 7, pp. 85090–85097, 2019, doi: 10.1109/access.2019.2924436.

[5]H. Sheng and X. Wang, "Probabilistic Power Flow Calculation Using Non-Intrusive Low-Rank Approximation Method", *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3014–3025, Jul. 2019, doi: 10.1109/tpwrs.2019.2896219.

[6]A. Seppälä, "Load research and load estimation in electricity", PhD thesis, Helenski University of Technology, 1996.

[7]E. M. Authority, "A Household's Typical Electricity Consumption Pattern in a Day". .

[8]"Project Data - Customer-Led Network Revolution". .

[9]"The Consumer Classification Solution for Consistent Cross-Channel Marketing". .

[10]F. Li, S. Zhang, W. Li, W. Zhao, B. Li, and H. Zhao, "Forecasting Hourly Power Load Considering Time Division: A Hybrid Model Based on K-means Clustering and Probability Density Forecasting Techniques", *Sustainability*, vol. 11, no. 24, p. 6954, Dec. 2019, doi: 10.3390/su11246954.

[11]L. Wasserman, *All of Nonparametric Statistics*. Springer New York, 2006.

[12]B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, 1986.

[13]M. Wahbah, T. H. M. El-Fouly, B. Zahawi, and S. Feng, "Hybrid Beta-KDE Model for Solar Irradiance Probability Density Estimation", *IEEE Transactions on Sustainable Energy*, pp. 1–1, 2019, doi: 10.1109/tste.2019.2912706.

[14]L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer New York, 2001.

[15]L. Brown, T. Cai, R. Zhang, L. Zhao, and H. Zhou, "The root–unroot algorithm for density estimation as implemented via wavelet block thresholding", *Probability Theory and Related Fields*, vol. 146, no. 3-4, pp. 401–433, Jan. 2009, doi: 10.1007/s00440-008-0194-2.

[16]C. Davidson-Pilon, "Poissonization of Multinomials". https://dataorigami.net/blogs/napkin-folding/127970947-poissonization-of-multinomials.

[17]"Homoscedasticity - Statistics Solutions". https://www.statisticssolutions.com/homoscedasticity/.

[18]M. Znidaric, "Asymptotic Expansion for Inverse Moments of Binomial and Poisson Distributions", *The Open Statistics & Probability Journal*, vol. 1, no. 1, pp. 7–10, Jan. 2009, doi: 10.2174/1876527000901010007.

[19]M. Wahbah, S. F. Feng, T. H. M. EL-Fouly, and B. Zahawi, “Wind speed probability density estimation using root-transformed local linear regression”, *Energy Conversion and Management*, vol. 199, p. 111889, Nov. 2019, doi: 10.1016/j.enconman.2019.111889.

[20]M. Wahbah, S. Feng, T. H. M. EL-Fouly, and B. Zahawi, “Root-Transformed Local Linear Regression for Solar Irradiance Probability Density Estimation”, *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 652–661, Jan. 2020, doi: 10.1109/tpwrs.2019.2930699.

[21]O. El-Dakkak, S. Feng, M. Wahbah, T. H. M. EL-Fouly, and B. Zahawi, “Combinatorial method for bandwidth selection in wind speed kernel density estimation”, *IET Renewable Power Generation*, vol. 13, no. 10, pp. 1670–1680, Jul. 2019, doi: 10.1049/iet-rpg.2018.5643.

[22]D. Freedman and P. Diaconis, “On the histogram as a density estimator: L2 theory”, *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, Dec. 1981, doi: 10.1007/bf01025868.

[23]E. G. Portugués, “Notes for Nonparametric Statistics”. .