Root Transform Local Linear Regression

Begad Elhouty[1], Samuel F. Feng[2], Tarek El-Fouly[3], Bashar Zahawi[3]

[1] Learning Delivery & Field Support, Emirates School Establishment
[2] Department of Science & Engineering, Sorbonne University
[3] Advanced Power & Energy Center,
Department of Electrical Engineering and Computer Science, Khalifa University

AGENDA

1. Motivation
2. Electric Load Data
3. Statistical Models
4. Assessment Methodology
5. Results

Motivation

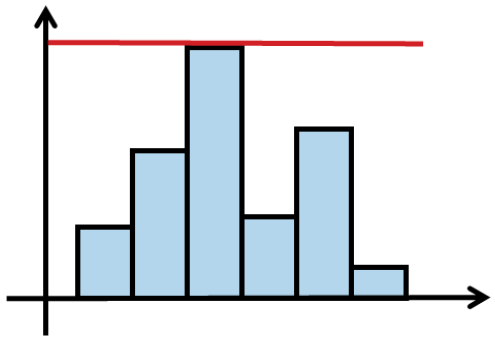
Electric Load Estimation

Problem Statement

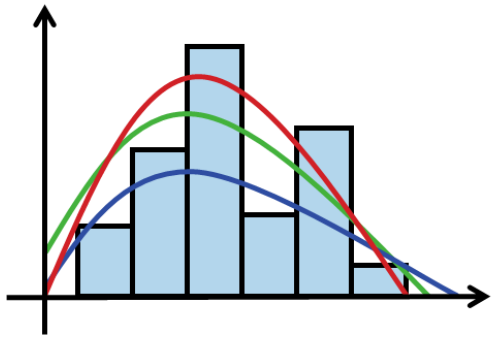
- Electric power networks try to strike a balance between meeting customer's energy demand and maintain supply quality.

- To combat this issue:

Traditionally



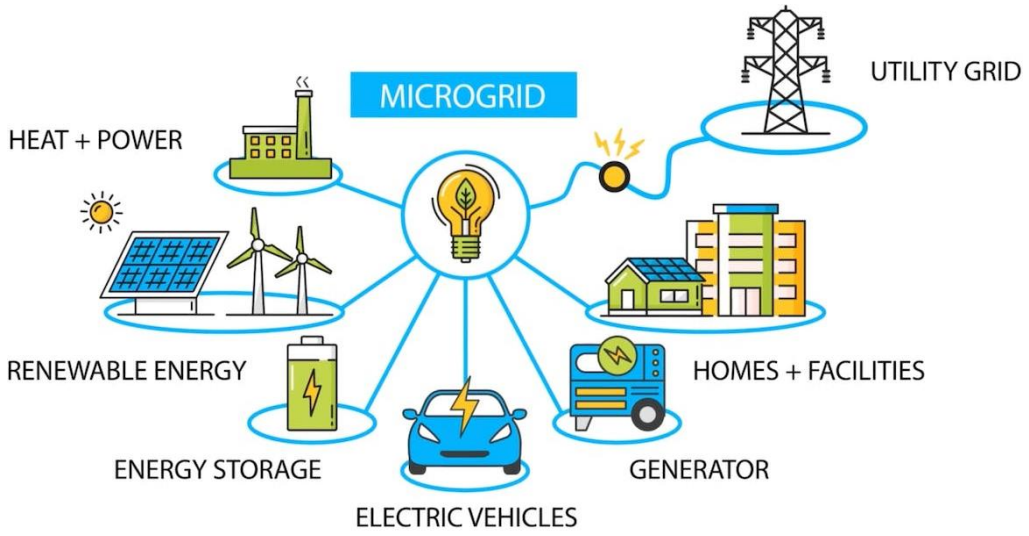
More Practical



Electric Load Estimation

Applications of Electric Load Estimation

- Planning, operating and optimizing microgrids.
- Optimal allocation of renewable distributed generation to minimize annual energy loss.
- Evaluating the reliability and accuracy of power systems with PV power generation.
- Integration of batteries with photovoltaic plants on a large scale.



Current Estimation Models

Parametric Models

Load Fluctuations are modelled in literature by various parametric distributions such as:

- Gaussian [2]
- Beta [3-4]
- Weibull [5]
- Log-Normal [5]
- Erlang [5]

Without **significant evidence**, assuming electrical load data follows a Gaussian or Beta distribution is dangerous as it may lead to wrong analyses.

Nonparametric Models

To address the rigid assumptions of parametric distributions and improve accuracy of estimation, electric load has been modelled by nonparametric distributions:

- Gaussian Mixture Models [6]
- Bayesian Approaches [7]

Proposed RTLLR Method

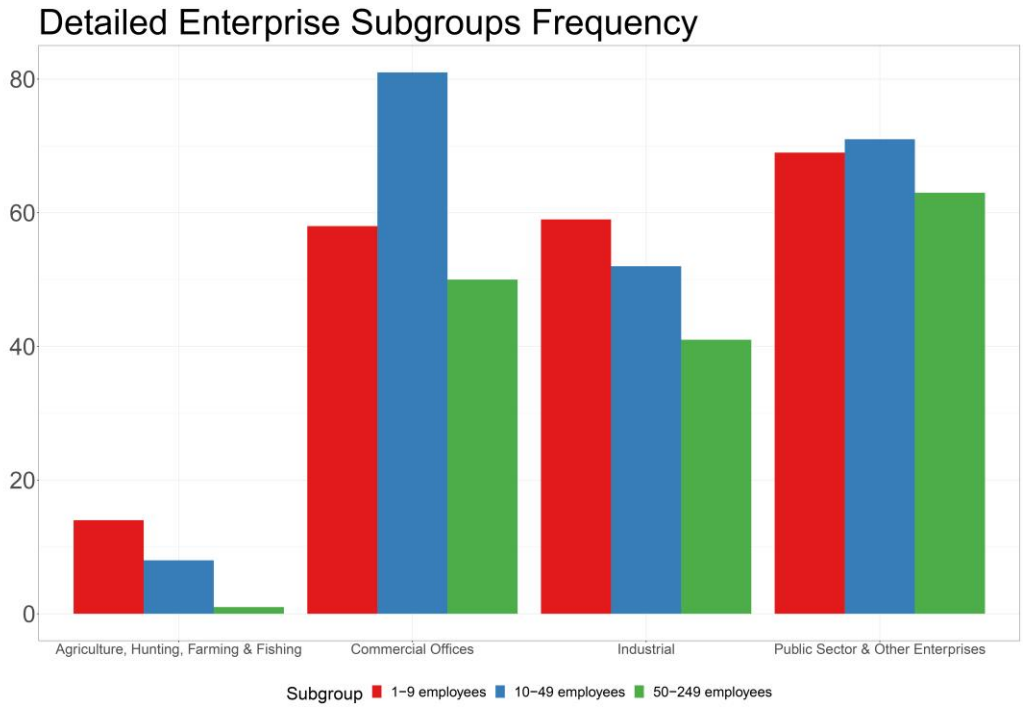
1. We propose a nonparametric method called Root Transform Local Linear Regression (RTLLR) for estimating electric load data. RTLLR is easy to use and computationally efficient.

2. Intensive comparative analysis with traditional parametric and nonparametric load density estimation models is conducted.

Electric Load Data

Electric Load Data

- Data collected from the Customer-Led Network Revolution based in UK.
- 567 Enterprise locations analyzed.
- Only locations with a year's worth of data analyzed
- Electric Power in kWh was collected per half-hour intervals.
- Enterprises were divided further into sectors and number of employees.



Statistical Models

The Statistical Problem

- Let $X_1, X_2, \ldots, X_n$ be the electric load data is collected as a time series with observations n. A yearly collection of electric load data yields 17520 electric load observations.

- Therefore, the goal is to find $P(a < X < b) = \int_a^b f_X \, dx$ for a random electric load point X in the future.

- However, f_X is unknown so we build estimates $\widehat{f_X}$ through statistical models.

Parametric Models

We studied the Gaussian & Gamma models.

$$\hat{f}_{Gaussian} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$ with mean μ and variance σ^2.

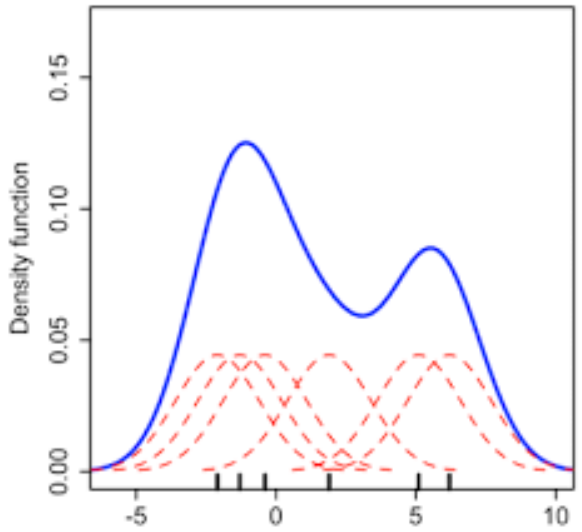
$$\hat{f}_{Gamma} = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$ with shape α and rate β parameters.

- Parametric models are interpretable and simpler to fit
- Parametric models have high bias due to having rigid assumptions.

Nonparametric Models

Kernel Density Estimators (KDE)

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



We use 2 Rule-Of-Thumb formulas for h:

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

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

- KDE has lower bias than parametric models.
- KDE suffers from boundary bias.

Root Transform Local Linear Regression

1. **Binning**. Electric load data is divided into $T \approx \frac{n}{10}$ bins, where n is the sample size. Let $Q_1, \dots, Q_T$ be the number of observations in each bin, and $x_1, \dots, x_T$ represent the centers of each of the bins.

2. **Variance Stabilizing Transform**. Calculate $y_i = \sqrt{Q_i + \frac{1}{4}}$, yielding a new paired data set with T observations: $(x_1, y_1), \dots, (x_T, y_T)$.

3. **Nonparametric Regression**. Any nonparametric regression can then be used on the paired data $(x_1, y_1), \dots, (x_T, y_T)$. We elect to use local linear regression because of its efficiency and accuracy in regression modelling.

4. **Unroot**. Reverse the root transform $\hat{f}_u(x) = \hat{r}(x)^2$.

5. **Normalize**. We normalize $\hat{f}_{RTLLR}$ so that $\hat{f}_{RTLLR} = \frac{\hat{f}_u(x)}{\int_0^1 \hat{f}_u(x)dx}$.

Assessment Methodology

Error Metrics

- Kolmogorov-Smirnov Test

- $RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(p_i - \widehat{p_i})^2}$

- $MAE = \frac{1}{T}\sum_{i=1}^{T}|p_i - \widehat{p_i}|$

- $R^2 = 1 - \frac{\sum_{i=1}^{T}(p_i - \widehat{p_i})}{\sum_{i=1}^{T}(p_i - \overline{p_i})}$, $\overline{p_i} = \frac{1}{T}\sum_{i=1}^{T} p_i$

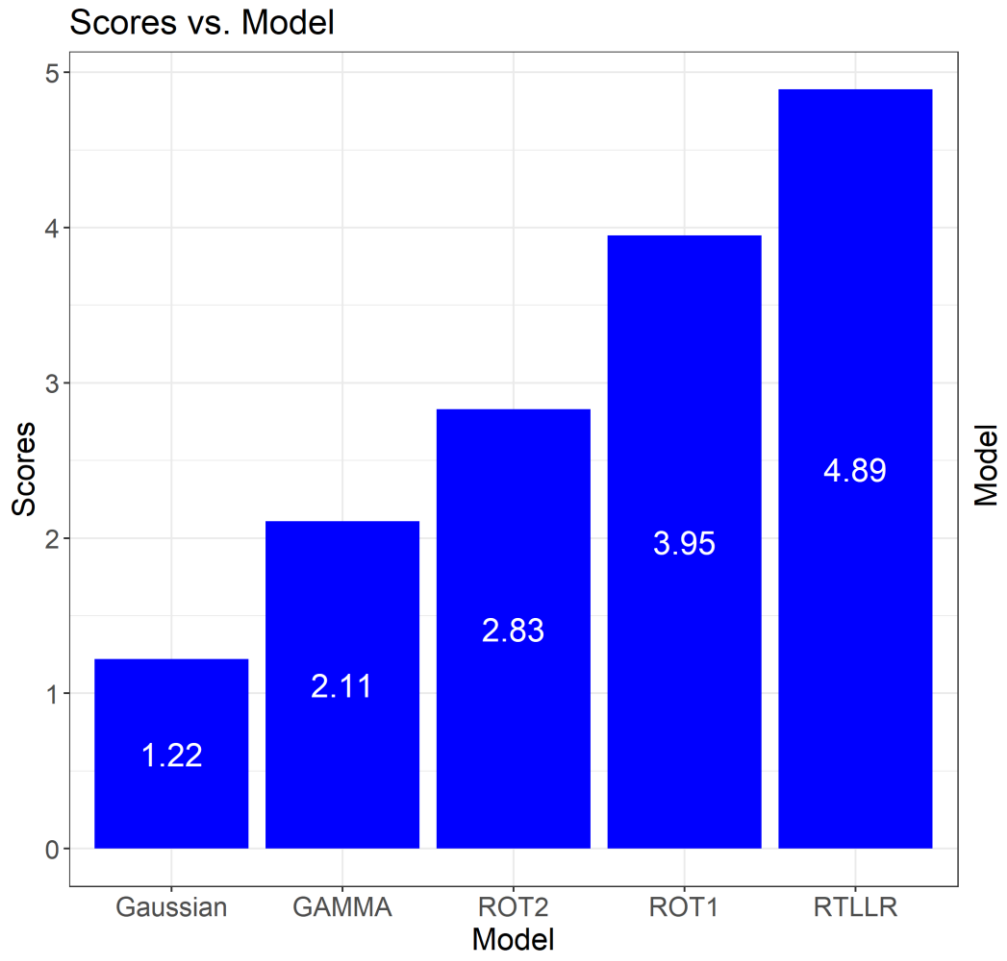
T is the number of bins,

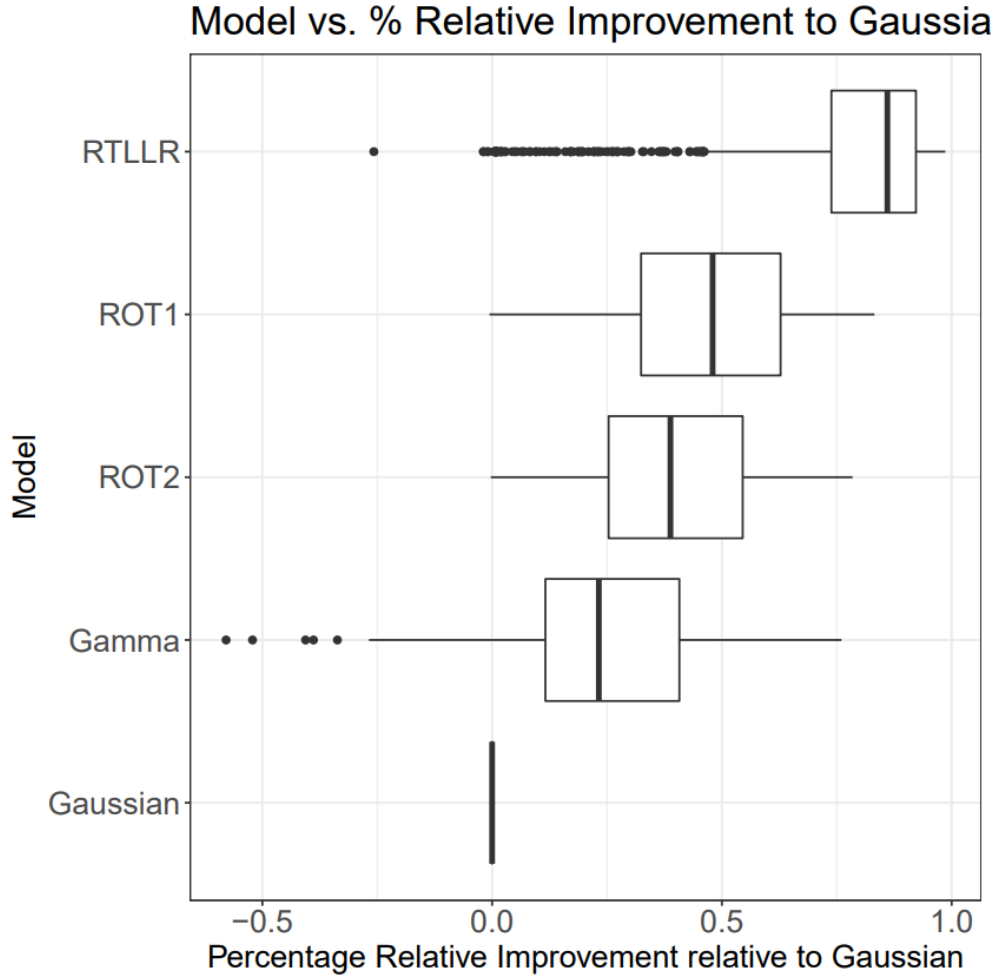
p_i is the probability that data fall in I_i,

$\widehat{p_i}$ is the estimated probability that data fall in I_i

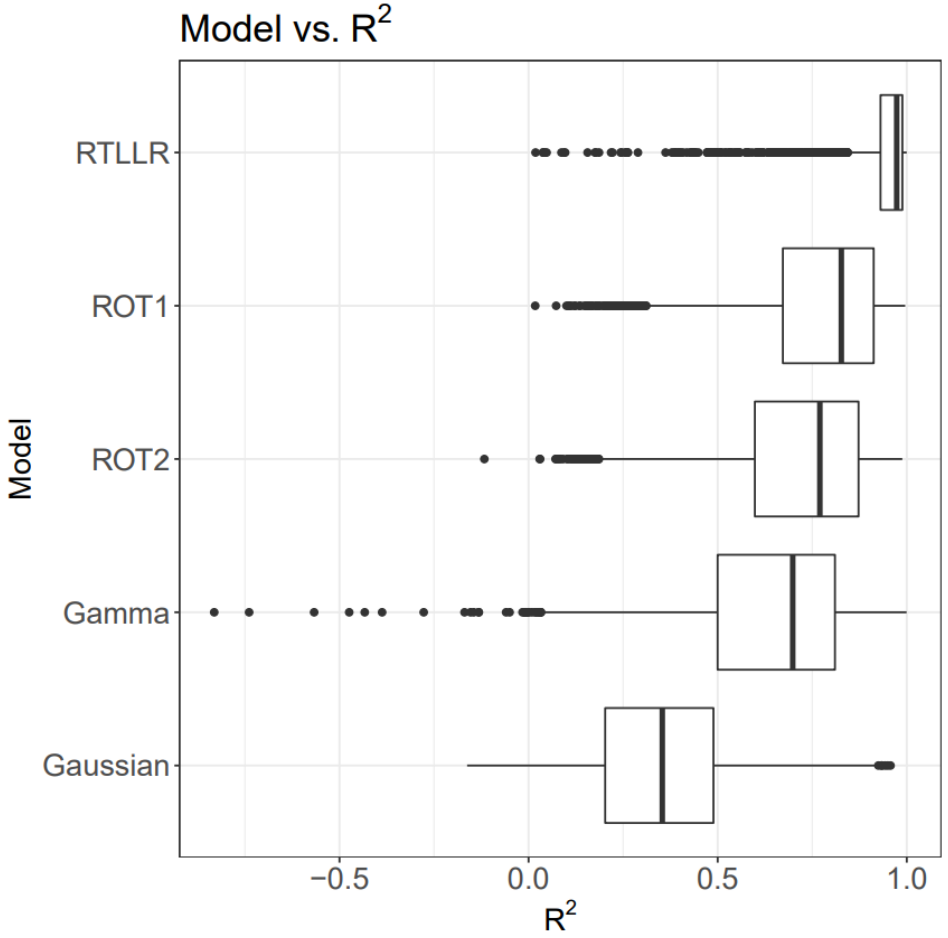
Results

RMSE Results





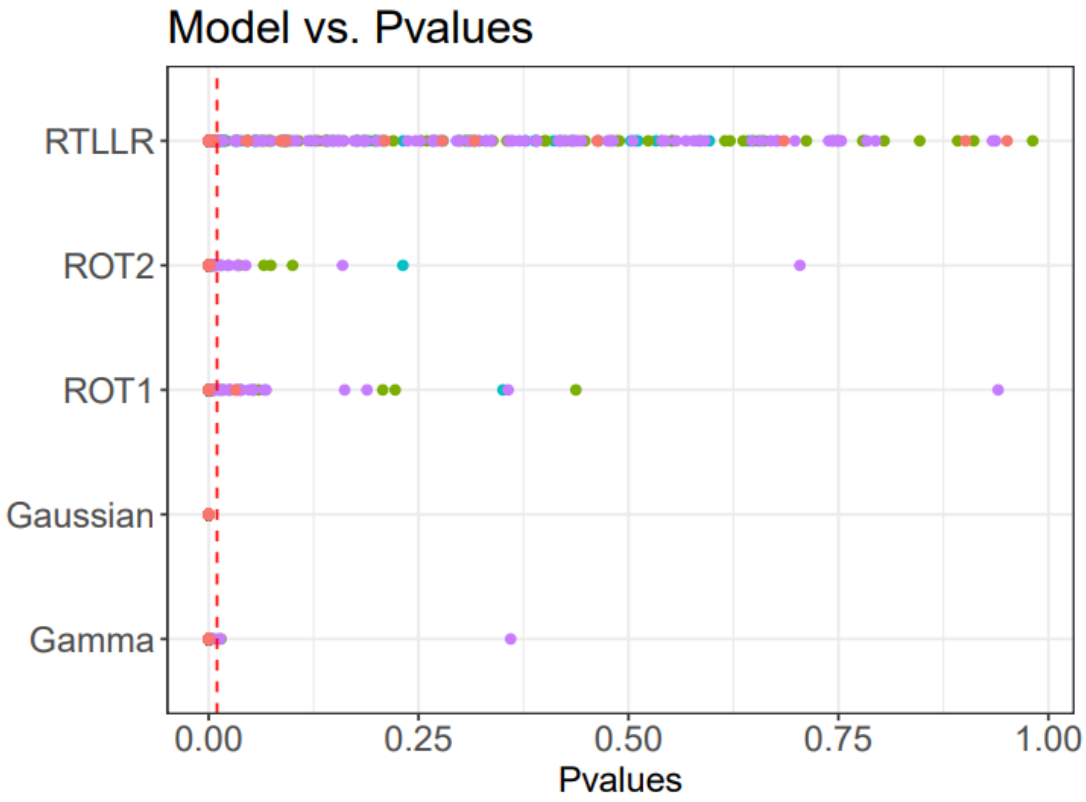
R^2 Results

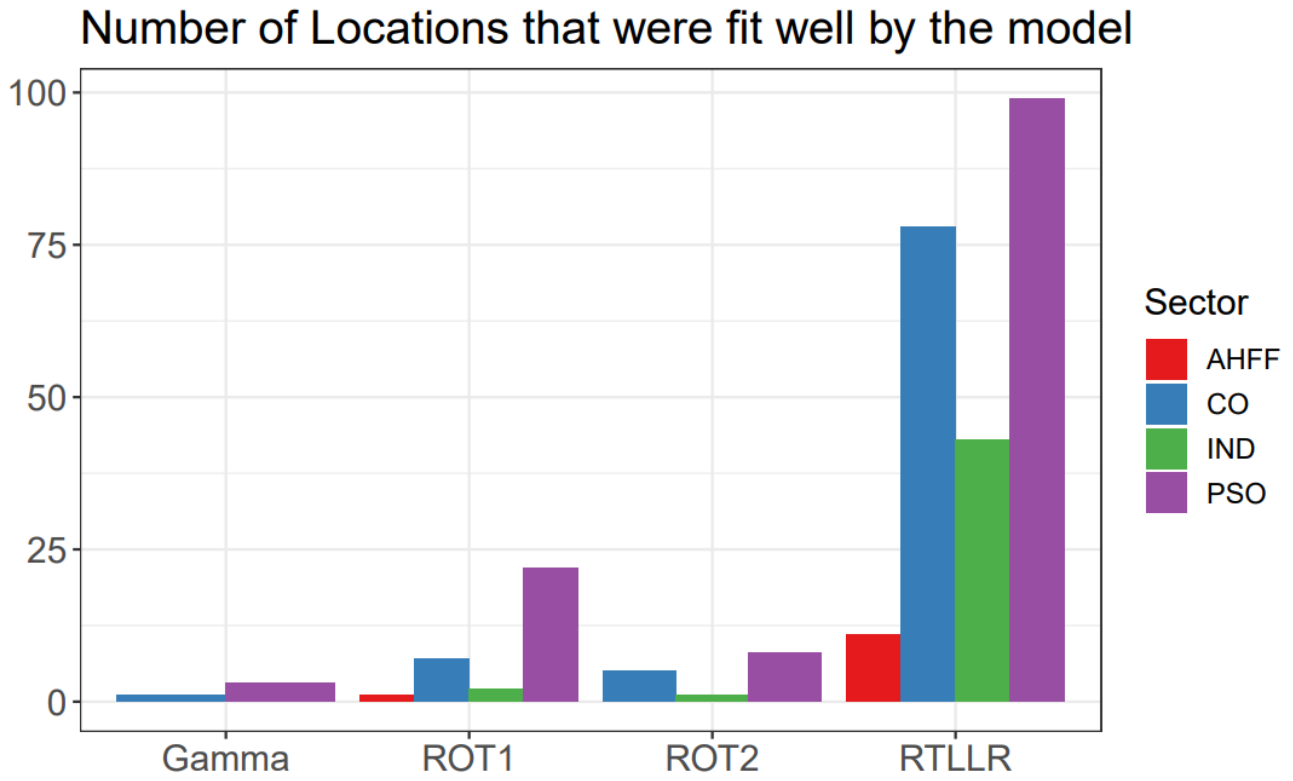


Enterprise	R^2	
	Average	Maximum
CO	91%	100%
IND	91%	100%
PSO	93%	100%
AHFF	98%	100%

Average and Maximum values for R^2

KS Test Results





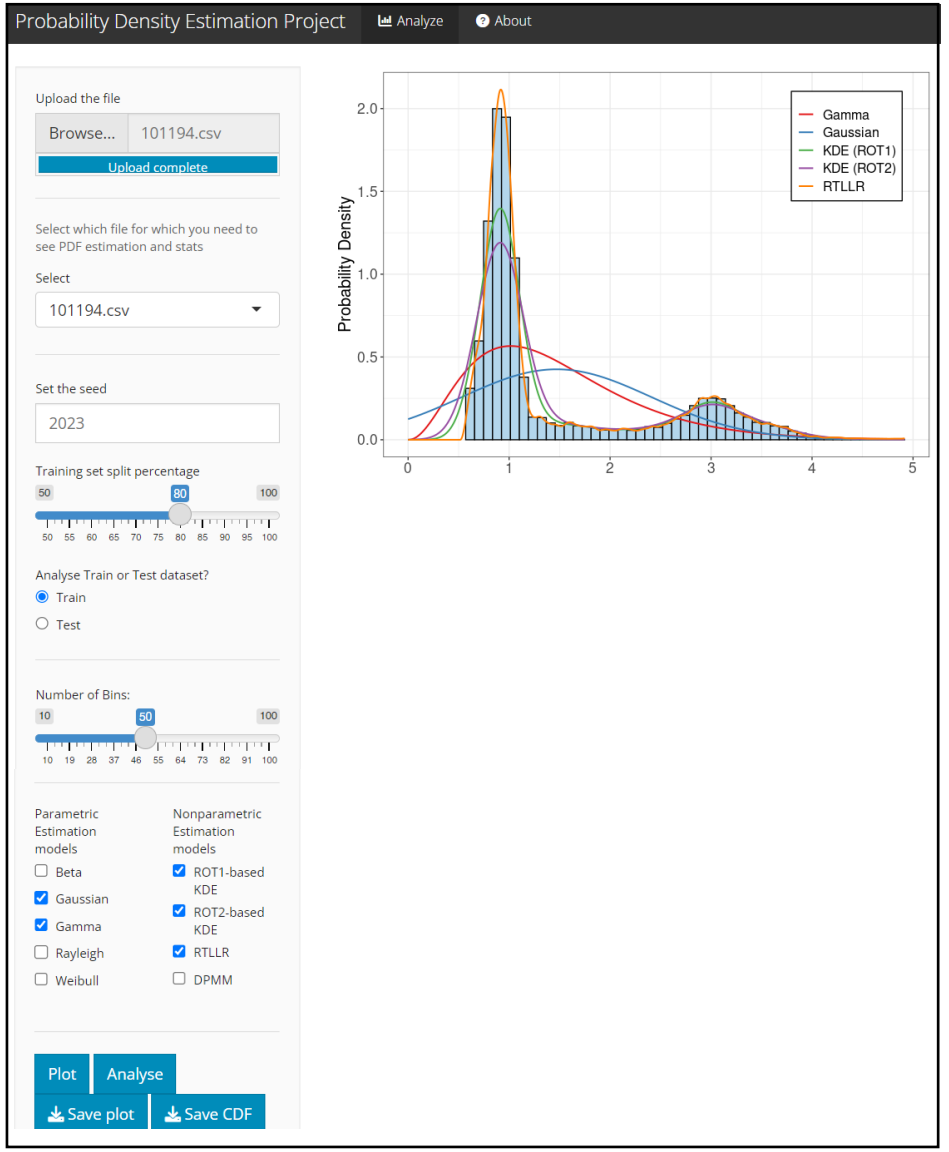
Probability Density Estimation Project

PDEP App

PDEP Web Application

Features of the application include:

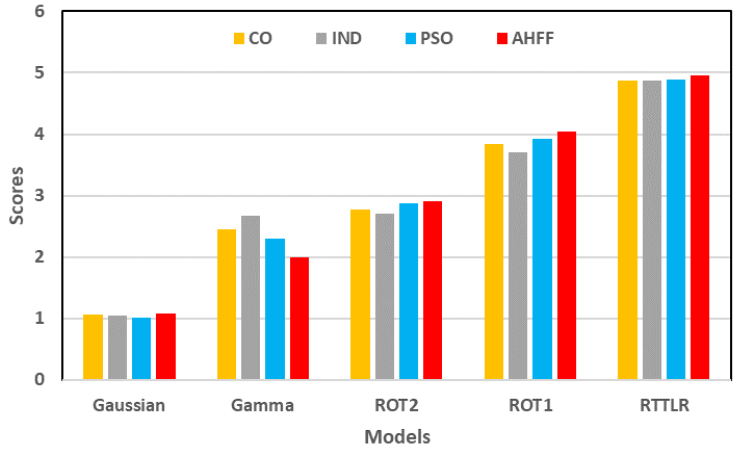
- Split provided dataset into train and test.
- Vary the train split percentage.
- Vary the seed
- Plot a selection of various parametric and nonparametric models
- Obtain error metrics for all models chosen
- Download plots
- Download CDF values for all models.



Conclusion

Conclusion

1. We recommend RTLLR for load density estimation as it:
   - avoids boundary bias in KDE models.
   - visually follows the shape of electric load distribution.
   - Has the lowest error metrics and highest R^2 values when compared to KDE and parametric models.



2. You can use an open-source web application to recreate our analysis on any data.

https://pdep.shinyapps.io/pdep_app/

Thank You

I would like to specially thank family, friends and the team for advising me throughout the project.

Appendix

Kolmogorov-Smirnov Test

Let $X_1, X_2, \ldots, X_n$ be iid with CDF F. Then, the null hypothesis for the Kolmogorov-Smirnov test is:

$$H_0: F = F_0$$

where F_0 is the CDF of the estimated model.

The Kolomogorov-Smirnov test statistic D_n is calculated by finding the supremum distance between the data's ECDF, an estimate of the true CDF, and the CDF of the model. The lower the value of D_n is, the better the model's fit is.

The p-value is then calculated from the statistic D_n where the model is a good fit for the data if the p-value is larger than the threshold $\alpha > 0.01$.