# ENHANCING PATTERN SEQUENCE-BASED FORECASTS: A MODIFIED STRATEGY RELATIVE TO ELECTRICAL LOAD

Suseelatha Annamareddi[1] and Sudheer Gopinathan[2,*]

•

Department of Mathematics, Gayatri Vidya Parishad College of Engineering for Women
Visakhapatnam, Andhra Pradesh, India
[1]suseelatha.a@gvpcew.ac.in, [2]g.sudheer@gvpcew.ac.in

**Abstract**

*A precise forecast of the one-day-ahead load is essential for the efficient management of modern power system operations. This paper proposes a univariate model for short term load forecasting (STLF) that improves the precision of the Pattern sequence forecasting (PSF) algorithm. An analysis was conducted to identify the underlying patterns in the electrical load data using K-means clustering and hierarchical clustering algorithms. The results demonstrate the efficacy of hierarchical clustering. The limitations of the original PSF algorithm, particularly in its clustering and prediction phases are addressed using hierarchical clustering and a new weighted average formula. The proposed method was validated using real-time series datasets and its performance was compared with those of three pattern sequence-based forecasting models. The performance is further evaluated on two electricity demand data sets and compared with bench mark models. The uncertainty and reliability of the forecast model was assessed using an error variance metric. The results show the superior forecast accuracy of the model.*

**Keywords:** short-term load forecasting, hierarchical clustering, pattern sequence, time series, weighted average.

## 1. Introduction

The growing concerns of society regarding sustainability, decarbonization, and environmental change have spurred technological advancements in electrification, electric vehicles, and renewable energy. These technological breakthroughs present substantial difficulties in the energy supply-demand balance, as electricity storage is difficult [1]. Consequently, electrical load forecasting is crucial for efficient electrical system management. Numerous load forecasting models have been proposed in the literature, depending on the time range of the future values to be predicted: short-term (intraday and day-ahead), medium-term (one week to several months ahead), and long-term (one or more years). Short-term load forecasts are critical for planning power system operations and for bidding strategies in deregulated electricity markets [2]. Load behavior is the fundamental driver of power pricing, therefore the level of accuracy in predicting future loads has a direct impact on the financial performance of energy businesses and other market participants [3].

Over the years, various techniques have been developed for Short Term Load Forecasting (STLF). These models include contemporary computational intelligence, machine learning, and

pattern recognition techniques in addition to traditional methods [4]. Among these, pattern recognition techniques leverage past data to identify load series patterns.

In the short term, load patterns are highly autocorrelated. Univariate models analyse past load patterns to predict future loads and do not depend on external factors. Consequently, univariate models can prevent inaccuracies caused by faulty or noisy exogenous data and produce more reliable and robust forecasts in STLF. This study proposes a univariate model for STLF that relies solely on the historical load series and does not incorporate any other information.

Pattern similarity is crucial in univariate models to ensure precise prediction. Understanding these patterns guarantees that the models capture the essential characteristics of the data, leading to more robust, interpretable, and applicable models across various domains [5]. Unsupervised learning techniques such as clustering reveal hidden patterns in data. This technique groups data points into meaningful clusters based on underlying patterns.

The Pattern Sequence Forecasting (PSF) technique [6] utilized clustering technique to identify patterns in time series data and then applied them to generate predictions. Owing to its efficacy and interpretability, it has gained prominence in a multitude of applications [7-10].

The PSF has certain limitations, although its performance in electrical load forecasting is encouraging. Some previous studies have addressed the limitations of the PSF algorithm and proposed improvements and modifications that are useful in increasing the forecast accuracy of electrical load data and in treating missing values and outliers [11-12]. The current study suggests alterations to the original PSF algorithm in both the clustering and prediction stages to improve the precision of electrical load forecasting. We performed a comprehensive analysis of the proposed methodology using publicly accessible Pennsylvania - New Jersey - Maryland (PJM) market demand data and compared it with benchmark models to ascertain its effectiveness.

The subsequent sections of the paper are organized as follows: Section 2 presents the original PSF algorithm, a literature review of the proposed PSF modifications, and scope for improvement. Section 3 outlines the proposed methodology. Section 4 reports and analyzes the performance of the proposed methodology. Section 5 summarizes the contributions of this study.

## 2. Pattern Sequence Similarity algorithm: Variations and scope for refinement

### 2.1. Original PSF algorithm

The PSF algorithm [6] can be divided into two phases: clustering and prediction. Phase one aims to assign each day, or a vector of 24 hours, to a cluster. The cluster pattern sequence prior to the day to be predicted was matched with the historical patterns, and future values were obtained by averaging the subsequent days of the matched patterns.

The different steps involved in both phases of the PSF algorithm are as follows:

The clustering component encompasses several activities, such as data normalization, determination of optimal number of clusters, and obtaining the cluster labels.

- Data normalization: Data normalization was used to smooth the trend from the original data. The transformation used in the original PSF algorithm is $x_j = \frac{x_j}{\frac{1}{N}\sum_{i=1}^{N} x_i}$ where $x_j$ is the demand of the jth hour of the day and $N$ is equal to 24 (the number of hours per day).
- Number of clusters: The optimal number of clusters is determined by the concordance between at least two of the following three indices: the Silhouette index, Dunn index, and the Davies-Bouldin index.
- Clustering/Labelling: K-means clustering was used to label each day with the optimal number of clusters. Clustering reduces the dimensionality of the data from 24 features to a single dimension, which enhances resilience by substituting the actual values with whole numbers (cluster labels).

The prediction phase in PSF consists of activities such as choosing the optimal window size,

identifying matching pattern sequences, and calculating the final forecasts.

Let $X(i) \in \mathbb{R}^{24}$ be a vector composed of 24-hourly demand of day $i$, and the corresponding cluster label is given by $L_i \in \{1, 2, \dots, K\}$, where $K$ is the number of clusters.

- Selection of optimal window size: The optimal window length $(w)$ of the pattern sequence must be determined prior to prediction. The calculation is performed using n-fold cross validation, and is selected at which prediction error $\sum_{t \in TS} \|\bar{X}(t) - X(t)\|$ is minimum during the training process. Here $\bar{X}(t)$ is the forecasted demand for day $t$, and $TS$ refers to the testing set.
- Identification of matching pattern sequences: If day $d$ is to be predicted, matchings for a sequence of labels $S_w^{d-1} = [L_{d-w}, L_{d-w+1}, \dots, L_{d-2}, L_{d-1}]$ of window length $w$, are searched in the labelled data. The search continues until at least one matching pattern sequence of the same length is discovered. If no replicates are identified, the window size is reduced by one unit. This guarantees the presence of at least one duplicate in a labelled sequence, with a minimum w value of 1.
- Forecasting: After identifying the matches, the subsequent 24 values that directly follow all coincidences are extracted to a vector $NS$. Finally, the values are averaged using the formula given in to anticipate the value of the future load.

$$\bar{X}(t) = \frac{1}{size(NS)} \sum_{j=1}^{size(NS)} NS(j) \quad (1)$$

## 2.2. Modifications Proposed in the literature

The literature proposes various modifications and improvements to the PSF algorithm. This section discusses some of the variations of the original PSF algorithm. The original PSF algorithm identifies analogous patterns in temporal data, although it had difficulties with specific instances. To address this issue, an enhanced version of the PSF algorithm was developed in [11] to predict anomalies in time series data with high accuracy. The method proceeds by using an additional measure to identify motifs or repetitive patterns in sequences that leads to improved predictions and capacity to identify potential outliers. A modification to PSF was proposed in [12], which uses nonnegative tensor factorization for clustering in PSF and is a promising direction for energy demand prediction. A novel method employing the PSF algorithm is presented in [13] wherein the accuracy of power demand forecasting was enhanced by employing distribution-based predictions and computing the frequency ratios of the cluster patterns. The imputePSF method suggested in [14] is a variation of the PSF algorithm that looks for recurring patterns in observed data to obtain a more accurate estimate of missing values. A novel hybrid algorithm, the funPSF, was designed to forecast functional time series, particularly in the context of electricity demand [15]. This algorithm combines functional data clustering with a forecasting strategy based on pattern sequence similarity. The bigPSF [16] and CUDA-bigPSF [17] algorithms, which build on the PSF method, design big data time series forecasting with notable improvements in scalability and accuracy. An improved version of the algorithm proposed in [10] which makes use of self-organizing maps and artificial neural networks and a genetic algorithm to determine the optimal hyperparameters of the model. MV-bigPSF algorithm [18], was proposed to forecast a multivariate time series. The model leverages the PSF algorithm, showcasing exceptional scalability and effectiveness in handling data sets consisting of millions of samples.

## 2.3. Scope for Refinement

In the clustering phase, the PSF algorithm employs K-means clustering. Despite its efficiency and simplicity, the K-means clustering algorithm has certain limitations, including sensitivity to the initialization of centroids, scale, and density. Also, ignoring the temporal order of the time series can hinder its efficiency when dealing with time series data or complicated pattern identification.

In the prediction phase, simply averaging the patterns observed immediately after a matched sequence may not be the optimal method. This is because the averaged pattern may not accurately depict the load pattern of the specific day under examination, particularly if the cluster patterns discovered differ from those of the previous working day [13].

## 3. Proposed Methodology

This section outlines the proposed methodology, which is based on the fundamental PSF algorithm. Below is a summary of the steps of the proposed methodology. Section 3.1 describes the data preprocessing; Section 3.2 details the clustering phase, which includes the determination of the clustering algorithm and the tuning of hyperparameters (k and w); and Section 3.3 presents the prediction phase. Fig. I illustrate the flow of the proposed method.

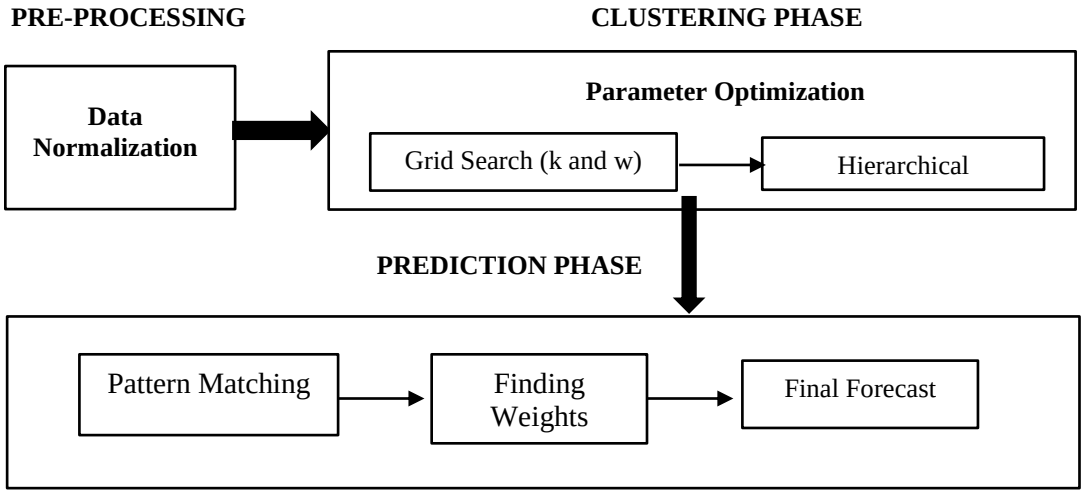


**Figure 1:** *Flow of the Proposed Methodology*

### 3.1 Pre-Processing Phase

Data normalization is a crucial technique in data pre-processing, particularly in clustering algorithms, as they rely on distance measurements to determine the similarity between any two data points. When features are not normalized, those with large scales can have a disproportionate impact on the distance calculations, resulting in biased or incorrect cluster labels.

The normalize technique used in this paper is

$$x_j' = \frac{x_j - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (2)$$

where $x_j'$ denotes the normalized value for $x_j$ and $i = 1, 2, \ldots, 24$.

This transformation is called Min-max normalization, which brings all the values into the range $[0,\ 1]$. Min-max normalization ensures that all features contribute equally to the clustering process, prevents any single feature from dominating due to its scale, and can lead to better and more interpretable clustering results [19].

### 3.2 Clustering Phase

The main objective of the clustering step is to classify the data into groups based on the behavior and underlying patterns in the time series. It provides a representation that preserves the original information and describes the shape of the time series data as accurately as possible.

The clustering phase consists of two steps: finding optimal values for the parameters and the clustering technique.

*3.2.1 Parameter Optimization (Optimal values of k and w)*

The proposed algorithm has two input parameters: the number of clusters ($k$) and the length of the window ($w$) that contains the search patterns. The optimal values of $k$ and $w$ are determined using a grid search over the training set. We split the original data set into training and testing sets. We further divided the training set into two sets, $y_T$ and $y_V$ one for training and the other for validation to fine-tune the hyperparameters. The proposed algorithm was applied to different combinations of $k$ and $w$ for prediction. Among all possible combinations of $k$ and $w$, the pair that results in the minimum prediction error on $y_V$ i.e., $\sum_{min} \|\overline{y_v}(t) - y_v(t)\|$ is considered as the optimal value of parameters $k$ and $w$.

*3.2.2 Clustering Technique*

This study used two different clustering algorithms: K-means and Hierarchical clustering. An analysis of both algorithms was performed to identify the patterns in the historical data.

*K-means Clustering:*
The primary concept underlying K-means clustering [20] is to establish k centroids, where each centroid represents a distinct cluster with k denoting the predetermined number of clusters. Each point in the given data set was assigned to its closest centroid. The centroids of these new clusters are recalculated and a new binding is performed between the same data points and the new centroids. Consequently, the location of the centroid's changes. This process was repeated until the centroids converged.

*Hierarchical Clustering:*
Hierarchical clustering generally falls into two types: the agglomerative (bottom-up) and the divisive (top-down). The agglomerative approach is the most common approach for hierarchical clustering. In agglomerative clustering, the clustering algorithm treats each point as an independent cluster and, iteratively merges the two most similar clusters into a single cluster at each step. It creates a tree-like structure called a dendrogram, which records sequences of merges or splits. Fig. 2 depicts a dendrogram with data points on the x-axis and cluster distance on the y-axis. The method of finding similarities between clusters results in the following hierarchical clustering variations: single, average, complete linkages, and Ward's method. Among them is the complete-linkage algorithm, which yields tightly bound clusters [21].

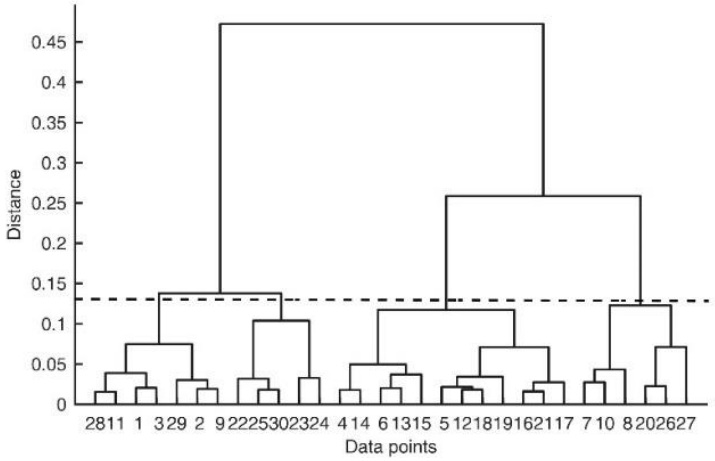


**Figure 2:** *Dendrogram*

## 3.3 Prediction Phase

This section proposes a new weighing prediction formula that addresses the limitations of the original PSF algorithm.

Let $O_w^i = [X(i-w+1), X(i-w+2), \dots, X(i-1),\ X(i)]$ be the vector composed of $w$ consecutive days prior to the day $'i'$. The distance between any pair of days $i, j$ is defined as $dist(i,j) = \left\| O_w^i - O_w^j \right\|$, where $\|.\|$ represents the Euclidean norm. The neighbors set of the day $'d-1'$ be $NS = \{q_1, q_2, \dots, q_m\}$ where $q_i$ is the day whose pattern sequence is matched with $S_w^{d-1}$ and $q_1$ and $q_m$ are the first and $m^{th}$ neighbor in order of distance calculated using the metric $'dist'$ . The weighted average of the load for the days following the nearest neighbors provides the prediction, assuming that load profiles that were similar in the past will likely be similar in the future. The prediction is given by Equation. (3)

$$X(d) = \frac{1}{\sum_{i \epsilon NS} \alpha_i} \sum_{i \epsilon NS} \alpha_i X(i+1) \tag{3}$$

where $\alpha_i$ are the weighting coefficients that can be obtained using any of the following schemes given below. The standard method of computing the weighting factors $\alpha_i$ as outlined in [22] is given by means of the Equation (4)

$$\alpha_i = \frac{dist(q_k, d-1) - dist(i, d-1)}{dist(q_k, d-1) - dist(q_1, d-1)} \tag{4}$$

# 4. Results and Discussion

This section outlines and analyses the performance of the proposed method. Section 4.1 describes the data set used to assess the effectiveness of the proposed method. Section 4.2 outlines the metrics used to measure the quality of the obtained results. Section 4.3 presents an analysis of the clustering techniques. Section 4.4 showcases the performance of the method using PJM market demand data for 2022, while, Section 4.5 outlines a comparative analysis with other methods proposed in the literature.

## 4.1 Data Set

Electricity demand data of the Pennsylvania - New Jersey - Maryland (PJM) market [23] on hourly basis for the year 2021 is considered to analyse the proposed methodology. The data set comprises 8760 data points with a mean $89.34 \times 10^3$ MW. Fig. 3 shows the hourly load data for the year 2021.

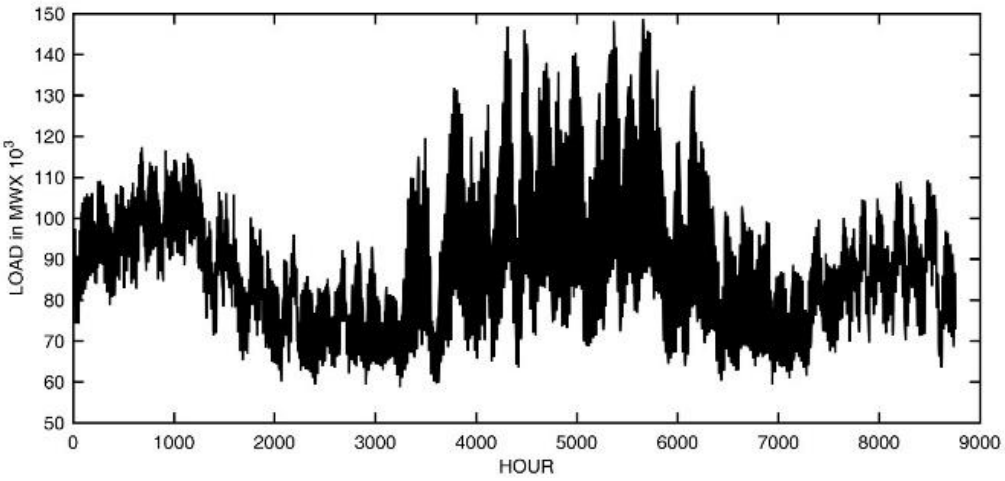


**Figure 3:** *Hourly demand data of PJM market in the year 2021*

One can observe a high demand during summer months specially from June through August. Before the clustering analysis the data was normalized as mentioned in Section 3.1.

## 4.2. Performance Metrics

The efficacy of the proposed methodology in obtaining day-ahead forecasts on the considered data was analysed using the forecast error metrics, Mean absolute percentage error (MAPE) in %, Root mean square error (RMSE) and Mean absolute error (MAE).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|l_i-\hat{l}_i|}{l_i}\times 100 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(l_i-\hat{l}_i\right)^2} \quad (6)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|l_i-\hat{l}_i\right| \quad (7)$$

where $l_i$ and $\hat{l}_i$ are the actual load and the forecast load at hour '$i$' respectively and $N$ is the number of predictions. In addition, the uncertainty in the forecasts can be estimated through the variance of the forecast error and evaluated using the metric VAR given by [24].

$$VAR = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|l_i-\hat{l}_i|}{l_i}-MAPE\right)^2 \quad (8)$$

## 4.3 Analysis of clustering techniques

In the clustering phase prior to the clustering analysis, the important step is to find the best values for the input parameters $k$ and $w$. We considered the yearly load data of the PJM market from 2021, which spans 365 days, for training, and use the 24-hourly load data from January 1, 2022, for validation. We varied $k$ from 2 to 7 and $w$ from 1 to 12 to measure the forecasting error when predicting the validation set. We found that $k = 4$ and $w = 5$ achieve the minimum RMSE, leading us to choose these values as optimal.

We used a sample data set of PJM load data from March 1, 2021, to May 31, 2021 (spring season) to demonstrate the effectiveness of the clustering techniques (k-means and hierarchical clustering). K-means and Hierarchical clustering were used to label each day in the sample data into 4 clusters. Tables I and II shows the percentage of days classified into four clusters. By observing the tables, we can clearly classify the clusters into two groups: workings days and weekends. From Table 1 and 2 it is evident that Cluster 1 represents a group of weekends and clusters 2, 3 and 4 represents a group of working days. However, from both the tables one can observe that some days are mislabelled owing to the complex behaviour of the load data.

**Table 1:** *The distribution of days of the week (in%) and clusters using k-means clustering*

| Cluster label | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| 1 | 7.14 | 0.00 | 0.00 | 0.00 | 0.00 | 61.54 | 69.23 |
| 2 | 64.29 | 69.23 | 69.23 | 53.85 | 69.23 | 23.08 | 15.38 |
| 3 | 21.43 | 23.08 | 15.38 | 30.77 | 23.08 | 7.69 | 7.69 |
| 4 | 7.14 | 7.69 | 15.38 | 15.38 | 7.69 | 7.69 | 7.69 |

**Table 2:** *The distribution of days of the week (in%) and clusters using Hierarchical clustering*

| Cluster label | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| 1 | 7.14 | 0.00 | 0.00 | 0.00 | 0.00 | 69.23 | 84.62 |
| 2 | 64.29 | 53.85 | 61.54 | 46.15 | 61.54 | 15.38 | 0.00 |
| 3 | 0.00 | 15.38 | 7.69 | 7.69 | 7.69 | 0.00 | 0.00 |
| 4 | 28.57 | 30.77 | 30.77 | 46.15 | 30.77 | 15.38 | 15.38 |

In Table 1, one working day and nine weekends were misclassified, and in Table 2, one working day and six weekends were misclassified. Upon thorough analysis of holidays during the above period, we find that the one mislabelled working day is a Monday, falling on May 31, 2021,

as a holiday. Therefore, out of 92 days (working days and weekends), k-means clustering mislabelled five Saturdays and four Sundays, whereas hierarchical clustering mislabelled four Saturdays and two Sundays. The relative errors for k-means clustering and hierarchical clustering were 9.78% and 6.52%. The above analysis reveals that the hierarchical clustering is effective in labelling the time series data.

## 4.4 Performance of the Proposed method

A case study is conducted by considering the hourly load data from the PJM market for the year 2022. The methodology is used to forecast day-ahead load data by considering the historical load of one year prior to the day in which the load is to be forecast. We advance the one-year training window for a specific day by one day, resulting in forecasts for the next 24 hours. This process yields forecasts for an entire year. We calculated and presented the monthly MAPE and error variance in Table 3 to evaluate the model's performance across all the months of 2022. The results are compared with the results of the model (K-means) obtained by using K-means clustering in the clustering phase. According to the results in Table 3, it is evident that the proposed methodology performs significantly better than the K-means model. The best and worst predictions occur on 7th of July and 30th of May with 0.2972 and 9.8249 MAPE (%) respectively. Fig. 4 and 5 shows the original day versus the predicted load.

**Table 3:** *Monthly MAPE (MMAPE) and Error variance (VAR) for all the months of the year 2022*

| Month | Proposed Methodology | | K-means | |
|---|---|---|---|---|
| | MMAPE | VAR | MMAPE | VAR |
| January | 2.71 | 6.46e-4 | 3.43 | 10e-4 |
| February | 2.41 | 4.33e-4 | 3.08 | 5.92e-4 |
| March | 2.54 | 5.40e-4 | 3.02 | 7.59e-4 |
| April | 2.46 | 6.15e-4 | 2.74 | 6.02e-4 |
| May | 2.22 | 6.97e-4 | 2.26 | 6.57e-4 |
| June | 2.18 | 4.60e-4 | 2.97 | 12e-4 |
| July | 2.01 | 4.50e-4 | 2.58 | 6.55e-4 |
| August | 1.84 | 4.34e-4 | 2.27 | 5.06e-4 |
| September | 2.17 | 3.96e-4 | 2.42 | 6.03e-4 |
| October | 1.53 | 1.80e-4 | 1.51 | 1.82e-4 |
| November | 2.2 | 4.79e-4 | 2.08 | 3.88e-4 |
| December | 2.18 | 5.79e-4 | 2.97 | 9.73e-4 |
| **Average** | **2.20** | **4.92e-4** | **2.61** | **6.76e-4** |

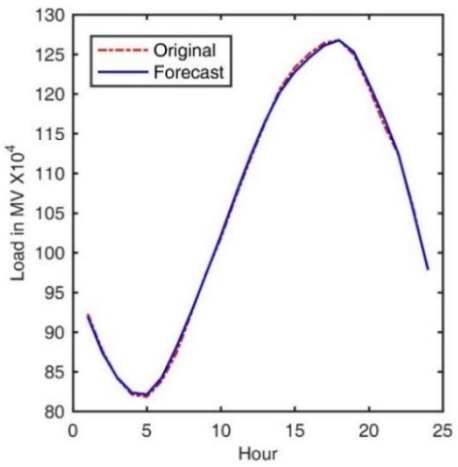


**Figure 4:** *Best Prediction in PJM load 2022*

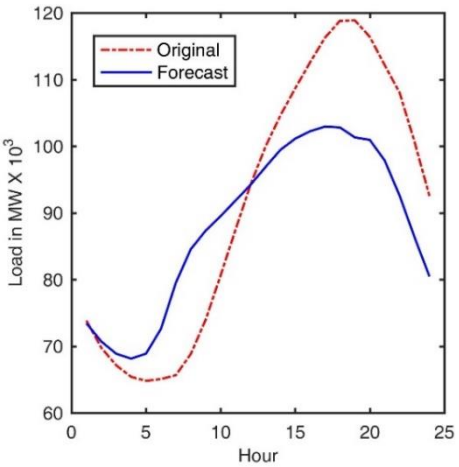


**Figure 5:** *Worst Prediction in PJM load 2022*

## 4.5 Comparative Analysis

We compared the proposed methodology to three main approaches: Bokde & Asencio et al.'s R package (2017) [19], which includes the basic PSF function (RPSF) ; Shende et al.'s (2022) Python package [25], which includes the basic PSF function (PPSF) and DPSF functions; and modified PSF algorithm proposed in [13].

First, we validated the proposed methodology using real-time series "nottem" and "CO2" datasets. The "nottem" time series contains the average air temperatures at Nottingham Castle in degrees Fahrenheit over 20 years, and the "$CO_2$" dataset consists of atmospheric concentrations of $CO_2$ expressed in parts per million (ppm). We conduct a comparative analysis against RPSF, PPSF, DPSF functions using the metrics root-mean-square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Both the time series data sets were partitioned into training and testing datasets. The training set comprised the time series data, with the exception of the final 12 values. The testing dataset comprised of final 12 values. The values of the error metrics are recorded in Table 4.

**Table 4:** *A comparative analysis of real-time series forecasting results*

| Time Series | Error Metric | RPSF | PPSF | DPSF | Proposed Method |
|---|---|---|---|---|---|
| nottem | RMSE | 2.24 | 1.84 | 5.27 | 1.81 |
| | MAE | 1.94 | 1.54 | 4.77 | 1.35 |
| | MAPE | 4.14 | 3.23 | 9.43 | 2.89 |
| $CO_2$ | RMSE | 5.93 | 1.42 | 0.41 | 1.21 |
| | MAE | 5.91 | 9.27 | 0.32 | 1.13 |
| | MAPE | 1.62 | 2.67 | 0.08 | 0.61 |

From the results, it is evident that the proposed methodology is performed well. The DPSF function yields better results for a data with positive or negative trend.

Secondly, the methodology was compared with the model proposed by Jin et al (2014) [13] using the load data of NYISO market over the year 2006 [26] with the error metric MAE. Load data of 2005 is used as training set and the forecast of 24-hours ahead is calculated. Further forecasts can be found by shifting the window of the training set to next day. The error metric MAE is evaluated and recorded in Table 5 against the measures obtained in [13].

**Table 5:** ***A*** *comparative analysis of the forecast results using MAE of NYISO load data from 2006*

| Month | **Mean Absolute Error (MAE)** | | | | |
|---|---|---|---|---|---|
| | RPSF | PPSF | DPSF | Modified PSF | Proposed Method |
| January | 6.71 | 9.61 | 6.7 | 3.45 | 2.18 |
| February | 6.91 | 9.84 | 8.07 | 3.8 | 2.38 |
| March | 5.20 | 9.01 | 7.87 | 3.59 | 2.58 |
| April | 9.15 | 8.87 | 12.01 | 3.32 | 2.18 |
| May | 9.62 | 12.35 | 10.37 | 3.67 | 2.07 |
| June | 8.06 | 15.27 | 11.05 | 4.53 | 3.35 |
| July | 9.60 | 15.91 | 10.41 | 5.84 | 3.37 |
| August | 8.75 | 13.33 | 12.83 | 4.07 | 2.17 |
| September | 8.45 | 10.39 | 8.65 | 2.6 | 2.18 |
| October | 4.22 | 9.97 | 7.7 | 2.92 | 1.67 |
| November | 4.80 | 9.12 | 6.87 | 3.47 | 2.54 |
| December | 7.65 | 11.73 | 8.58 | 3.77 | 2.52 |
| **Average** | **7.43** | **11.28** | **9.26** | **3.75** | **2.43** |

## 5. Conclusion

The paper presents a methodology that enhances the forecast capability of the PSF algorithm. The modifications to the PSF algorithm that includes a judicious use of hierarchical clustering algorithm in its clustering phase and a weighted average formula in the prediction phase has led to improved accuracy in day ahead load forecasts. Alongside MAE, RMSE, MAPE, the error variance (VAR) has been used for a comprehensive evaluation of the model's performance. The proposed model outperforms benchmark models in terms of forecasting accuracy, as evidenced by the performance metrics calculated from the real-time series data. The findings highlight the effectiveness of the proposed approach in enhancing the precision of day-ahead load forecasts, making it a valuable tool for efficient power system management and operational planning in electricity markets.

## References

[1] Aquila, G., Morais, L.B.S., de Faria, V.A.D., Lima, J.W.M., Lima, L.M.M., and de Queiroz, A.R. (2023). An Overview of Short-Term Load Forecasting for Electricity Systems Operational Planning: Machine Learning Methods and the Brazilian Experience, Energies, 16(21) : 7444. https://doi.org/10.3390/en16217444.

[2] Liu, Y., Dutta, S., Kong, A. W. K., and Yeo, C. K. (2023). An Image Inpainting Approach to Short-Term Load Forecasting, IEEE Transactions on Power Systems, 38(1): 177-187. https://doi: 10.1109/TPWRS.2022.3159493.

[3] Dudek, G. (2015). Pattern similarity-based methods for short-term load forecasting – Part 1: Principles, Applied Soft Computing, 37: 277-287. https://doi.org/10.1016/j.asoc.2015.08.040.

[4] Hong. T., Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review, International Journal of Forecasting, 32(3): 914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011.

[5] Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D.L.R., Thompson, E.B., and Ashraf, I.A. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective, Symmetry, 15(9):1679. https://doi.org/10.3390/sym15091679.

[6] Martínez-Álvarez, F., Troncoso, A., Riquelme, J., and Aguilar-Ruiz, J.S. (2011). Energy Time Series Forecasting Based on Pattern Sequence Similarity, IEEE Transactions on Knowledge and Data Engineering, 23:1230-1243. https://doi: 10.1109/TKDE.2010.227.

[7] Bokde, N., Troncoso, A., Asencio-Cortés, G., Kulat, K. and Martínez-Álvarez, F. (2017). Pattern sequence similarity based techniques for wind speed forecasting, in: Proceedings of the International Work-Conference on Time Series, 2: 786–794.

[8] Gupta, A., Bokde, N., Kulat, K.D. (2018). Hybrid leakage management for water network using PSF algorithm and soft computing techniques, Water Resourse Management, 32(3):1133–1151. https://doi.org/10.1007/s11269-017-1859-3

[9] Zhu, K., Geng, J., Wang, K. (2021). A hybrid prediction model based on pattern sequence-based matching method and extreme gradient boosting for holiday load forecasting, Electric Power Systems Research, 190: 106841. https://doi.org/10.1016/j.epsr.2020.106841.

[10] Criado-Ramón,D., Ruiz, L.G.B., Pegalajar, M.C. (2023). An improved pattern sequence-based energy load forecast algorithm based on self-organizing maps and artificial neural networks, Big Data and Cognitive Computing, 7 (2) :92. https://doi.org/10.3390/bdcc7020092.

[11] Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S. (2011). Discovery of motifs to forecast outlier occurrence in time series, Pattern Recognition Letters, 32(12): 1652–

1665. https://doi.org/10.1016/j.patrec.2011.05.002.

[12] Fujimoto, Y., Hayashi, Y. (2021). Pattern sequence-based energy demand forecast using photovoltaic energy records, 2012 International Conference on Renewable Energy Research and Applications (ICRERA), Nagasaki, Japan, pp. 1-6. https://doi:10.1109/ICRERA.2012.6477299.

[13] Jin, C.H., Pok, G., Park, H.-W. and Ryu, K.H. (2014). Improved pattern sequence-based forecasting method for electricity load, IEEJ Transactions on Electrical and Electronics Engineering, 9 (6): 670–674. https://doi.org/10.1002/tee.22024.

[14] Bokde, N., Beck, M.W., Martínez-Álvarez, F., Kulat, K. (2018). A novel imputation methodology for time series based on pattern sequence forecasting, Pattern Recognition Letters, 116: 88-96. https://doi.org/10.1016/j.patrec.2018.09.020.

[15] Martínez-Álvarez, F., Schmutz, A., Asencio-Cortés,G., Jacques, J. (2019). A novel hybrid algorithm to forecast functional time series based on pattern sequence similarity with application to electricity demand, Energies. 12 (1): 94–111. https://doi.org/10.3390/en12010094.

[16] Perez-Chacon, R., Asencio-Cortes, G., Martínez Alvarez, F., & Troncoso, A. (2020). Big data time series forecasting based on pattern sequence similarity and its application to the electricity demand, Information Sciences, 540:160–174. https://doi.org/10.1016/j.ins.2020.06.014.

[17] Criado-Ramon D., Ruiz, L.G.B., Pegalajar, M.C. (2023). CUDA-bigPSF: An optimized version of bigPSF accelerated with graphics processing unit, Expert Systems with Applications. 230: 120661. https://doi.org/10.1016/j.eswa.2023.120661.

[18] Perez-Chacon, R., Asencio-Cortés, G., Troncoso, A., Martínez-Álvarez, F. (2024). Pattern sequence-based algorithm for multivariate big data time series forecasting: Application to electricity consumption, Future Generation Computer Systems, 154: 397-412. https://doi.org/10.1016/j.future.2023.12.021.

[19] Bokde, N., Asencio-Cortés, G., Martínez-Álvarez, F., Kulat, K. (2017). PSF: Introduction to R package for pattern sequence based forecasting algorithm, The R Journal, 9 (1): 324–333. http://dx.doi.org/10.32614/RJ-2017-021.

[20] J Hartigan, J. A., Wong, M. A. (1979). Algorithm AS136 : A k-means clustering algorithm, Applied Statistics, 28: 100-108. https://doi.org/10.2307/2346830.

[21] Maimon, O., Rokach, L. (2010). Data mining and Knowledge Discovery Handbook, Springer, second edition.

[22] Lora, A.T., Santos, J. M. R., Exposito, A. G., Ramos, J. L. M., and Santos., J. C. R. (2007). Electricity market price forecasting based on weighted nearest neighbors techniques, IEEE Transactions on Power Systems, 22(3): 1294–1301. doi: 10.1109/TPWRS.2007.901670.

[23] Pjm market website: https://pjm.com (Accessed on 04 April, 2024).

[24] O. Abedinia Abedinia, O., Amjady, N. (2017). A new feature selection technique for load and price forecast of electrical power systems, IEEE Transactions on Power Systems, 32(1): 62-74. https://doi.org/10.1109/TPWRS.2016.2556620

[25] Shende, M.K., Salih, S. Q., Bokde, N. D., Scholz, M., Oudah, A. Y., Yaseen, Z.M. (2022). Natural Time Series Parameters Forecasting: Validation of the Pattern-Sequence-Based Forecasting (PSF) Algorithm; A New Python Package. Applied Science, 12: 6194. https://doi.org/10.3390/app12126194.

[26] New York ISO: https://www.nyiso.com/ (Accessed on 04 April 2024).