Multi-year Long-term Load Forecast for Area Distribution Feeders based on Selective Sequence Learning

Ming Dong\textsuperscript{a,+}, Kaigui Xie\textsuperscript{b}, QingXin Shi\textsuperscript{c}

\textsuperscript{+} Corresponding author email: mingdong@ieee.org

\textbf{Abstract} - Long-term load forecast (LTLF) for area distribution feeders is one of the most critical tasks frequently performed in electric distribution utility companies. For a specific planning area, cost-effective system upgrades can only be planned out based on accurate feeder LTLF results. In our previous research, we established a unique sequence prediction method which has the tremendous advantage of combining area top-down, feeder bottom-up and multi-year historical data all together for forecast and achieved a superior performance over various traditional methods by real-world tests. However, the previous method only focused on the forecast of the next one-year. In our current work, we significantly improved this method: the forecast can now be extended to a multi-year forecast window in the future; unsupervised learning techniques are used to group feeders by their load composition features to improve accuracy; we also propose a novel selective sequence learning mechanism which uses Gated Recurrent Unit network to not only learn how to predict sequence values but also learn to select the best-performing sequential configuration for each individual feeder. The proposed method was tested on an actual urban distribution system in West Canada. It was compared with traditional methods and our previous non-selective sequence prediction method. It demonstrates the best forecasting performance as well as the possibility of using sequence prediction models for multi-year component-level load forecast.

KEY WORDS: Long-term Load Forecast, Artificial Intelligence, Sequence Prediction, Gated Recurrent Unit Neural Network, Clustering
1. Introduction

The purpose of long-term load forecast (LTLF) is to forecast electric power demand in more than one-year planning horizon for different areas in a power system [1-2]. It is the foundation of system planning work. For electric distribution utility companies, LTLF of distribution feeder peak demand is especially important because it generates the necessary input for assessing the distribution power delivery capacity during normal operation and the restoration capability during contingencies for the next few years. Only based on accurate load forecast results, utility companies can plan long-term infrastructure upgrades or modifications in a cost-effective way [1-2]. In practice, there are three approaches for doing LTLF [3-5]:

1. Top-down approach: this approach focuses on the forecast of electric power demand at an area-level such as the total load of all customers or all residential customers in an area [4]. These methods often apply univariate or multivariate regression models to analyze the long-term trend of area loading or the long-term relationship between external variables (economy, demographics and weather) and area loading [6-10]. One advantage of these methods is the statistical explicable. Utility companies can leverage the forecast of external variables made by government or third-party agencies for area load forecast. However, these methods cannot effectively handle area components such as individual distribution feeders in a planning area. This is because strong variation exists at the component level and it is unrealistic to assume all components simply comply with the area-level total load behaviour. In comparison, concrete planning actions usually must be made at the component level. Due to this drawback, top-down forecasting is often only used as a reference to ensure component-level forecast does not contradict with the area overall characteristics [3,5].

2. Bottom-up approach: this approach requires gathering customer load information for building a higher level forecast and has been often used for component-level forecast. The gathering of customer-load information can be conducted through utility surveys or customer interviews. For each feeder, customer long-term information such as expected sizes of new loads, load maturation plan and/or long-term production plan can be obtained, summarized and estimated to be yearly loading change. In practice, load information gathering is only done for major customers since it is less practical to gather from all customers [3]. In spite of the tremendous communication effort with major customers, forecast accuracy can still be in question due to unreliable customer information and change of customer plans over the forecasting horizon. It is very common that customers lack sufficient understanding of the area economics and either overestimate or underestimate their load plans.

3. Hybrid approach: this approach aims to combine the advantages of top-down and bottom-up forecasting and overcome their drawbacks. Not many researches are found in this direction: [11] proposed a statistically-
adjusted end-use model for household-level load forecast. For area distribution feeder LTLF, our recent research [12] made the first attempt by using sophisticated sequence prediction models including Long Short-term Memory Network (LSTM) and Gated Recurrent Network (GRU). There are two significant advantages with this method:

- It can seamlessly combine top-down area information with bottom-up feeder information in one mathematical model for power demand forecast;
- It converts the LTLF problem to a sequence prediction problem – this type of problem intends to use historical values in a sequence to predict the next value(s) in a sequence. A typical example is predicting the next word or a few words in a sentence given the first few words and this application has been used widely in text input enhancement today [13-14]. Using sequence prediction for LTLF is a very powerful idea because similar to word prediction, the next year’s loading may not only be related to current year but also the past few years. For example, sometimes a loading increase momentum may last for a few years; sometimes a slow build up when economy conditions are not ideal may turn to a dramatic load increase when economy recovers. Our method can comprehend information embedded in multiple years for forecasting purpose.

In [12], we tested the sequence prediction method on actual data coming from a major urban grid in West Canada and demonstrated superior performance over various traditional methods. However, [12] focuses on the prediction of only next one year while in reality multiple future years may need to be forecasted. To solve this problem and further enhance the existing method, this paper introduces the following major improvements:

- It now extends sequence prediction from one year to multiple years in the future and evaluates the performance of sequence prediction on such basis;
- It introduces unsupervised learning to the learning process to automatically group feeders based on their load composition features. Then each group of feeders learns in a separate way;
- It introduces a novel learning mechanism to automatically learn and remember the best-performing sequential configuration among many-to-one, many-to-many, shifted many-to-many configurations for each feeder. Future forecast adopts the best-performing sequential configuration accordingly for each feeder.

The structure of the proposed method is shown in Fig.1:
In the beginning, all interested feeders in a planning area are clustered into different groups by their load compositions. Then feeder features are constructed and fed into the Virtual Feeder Conversion (VFC) module to eliminate data noises resulted from historical load transfer events; in parallel, area economic and temperature features go through Principal Component Analysis to reduce dimensions. In the end, processed area features and feeder features are combined to construct three different sequential datasets, corresponding to many-to-one, many-to-many and shifted many-to-many sequential models. Each sequential model is trained. Then for each feeder, the best-performing sequential model is identified through special evaluation and labeled for this feeder. Finally, for future forecast, three sequential models will be selected alternatively for different feeders to achieve the best overall forecasting accuracy for the planning area.
This paper firstly introduces the step of feeder clustering. It then introduces the area and feeder features used in the method with some special processing techniques. It then reviews the non-gated Recurrent Neural Network (RNN) and gated RNN as well as three sequential configurations and the format of their datasets. Section 5 explains the unique learning process for best-performing sequential configurations. In the end, the approach is applied to a large urban grid in West Canada and discussed with case studies. The proposed method achieves better performance than our previous method and other traditional methods. It also demonstrates a general sequence prediction method for multi-year component-level load forecast.

2. Feeder Clustering

Clustering is an unsupervised learning technique. It learns from data that is not pre-labeled [15]. Clustering analyzes the commonalities between data points and groups similar data points together. In the proposed method, feeders in one area will be clustered based on the feeder load composition. Typically, each feeder contains residential, commercial and industrial loads. These three types of loads mix on a feeder according to certain percentages. Because each type of load responds to economy and temperature in different ways, our method first groups feeders with similar load compositions so that different prediction models can be established down the road for each group of feeders. This step can enhance the prediction accuracy and is compared with treating all feeders in an area as only one group in Section 6.

K-Means clustering is chosen as the clustering method. It is a widely used method and has great efficiency and simplicity [16]. It requires only one input parameter $K$ which is the expected number of clusters. To optimize the clustering performance, this paper further explains the use of Silhouette analysis used as a clustering quality evaluation method to help select $K$ [17].

2.1. Feeder Load Composition

Generally, there could be three types of loads on feeders: residential, commercial and industrial loads. Some feeders such as dedicated feeders may have only one type of loads while more commonly, many feeders contain more than one type. Feeder load composition can be described by the percentages of each type of loads. Residential load percentage $R$, commercial load percentage $C$ and industrial load percentage $I$ comply with:

$$R+C+I=1$$  \hspace{1cm} (1)

Therefore, to reduce the dimensionality and complexity of clustering, only two percentage numbers are required to characterize a feeder’s composition. This means clustering can be performed on a 2-D basis. For example, we assume $R$ and $C$ are selected. $R$ can be calculated by:
\[ R = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{n} \frac{L_{i,t}^R}{P_t} \times 100\% \]  

where \( P_t \) is the summer/winter peak load of the feeder in \( t_{th} \) historical year; \( L_{i,t}^R \) is the loading of residential load \( i \) at the feeder’s peaking time in \( t_{th} \) historical year; \( n \) is the total number of residential loads on this feeder; \( N \) is the number of historical years used for learning.

Similarly, commercial peak load percentage of a feeder is calculated by:

\[ C = \frac{1}{N} \sum_{t=1}^{N} \sum_{i=1}^{n} \frac{L_{i,t}^C}{P_t} \times 100\% \]  

where \( L_{i,t}^C \) is the loading of commercial load \( i \) at the feeder’s peaking time in \( t_{th} \) historical year; \( n \) is the total number of commercial loads on this feeder.

After the above calculations, a feeder can be characterized with a vector \((R, C)\).

2.2. K-means Clustering

Mathematically, K-Means clustering is described as below: given a set of data points \((X_1, X_2, ..., X_n)\), where each data point is a q-dimensional real vector, K-Means clustering aims to group \( l \) data points into \( K \) (\( \leq l \)) clusters \( S = \{S_1, S_2, ..., S_K\} \) so as to minimize the within-cluster variances of \( S \). Formally, the objective function is defined as:

\[ \arg \min_{k=1}^{K} \sum_{x \in S_k} \| x - \mu_k \|^2 = \arg \min_{l} \sum_{i=1}^{K} |S_i| \text{ Var } S_i \]  

where \( \mu_k \) is the mean of data points in cluster \( S_k \) [16]. The steps of K-Means are described in Algorithm 1.

To signify the numerical differences, the raw \( R \) and \( C \) percentage numbers can be further normalized using Min-Max normalization [15]:

\[ R_i' = \frac{R_i - \text{Min}(R)}{\text{Max}(R) - \text{Min}(R)} \]  

where \( R_i \) is the raw percentage number of residential load on \( i_{th} \) feeder; \( \text{Min}(R) \) and \( \text{Max}(R) \) are the minimum and maximum values of all feeders’ residential load percentage. Feature \( C \) can be normalized in the same way.
The distance between any two feeders $F_1$ and $F_2$ can be calculated using standard Euclidean distance as below:

$$d(F_1,F_2) = \sqrt{(R'_{F_1} - R'_{F_2})^2 + (C'_{F_1} - C'_{F_2})^2}$$  \hspace{1cm} (6)$$

where $R'_{F_1}$, $R'_{F_2}$, $C'_{F_1}$, $C'_{F_2}$ are the normalized load composition features for two feeders $F_1$ and $F_2$.

2.3. Clustering Quality Evaluation and Determination of Parameter $K$

Silhouette analysis as a clustering quality evaluation method can be used to determine the optimal parameter $K$ from an initial range of $K$ values [17]. In this analysis, Silhouette coefficient $Q_r$ is used as an index to evaluate clustering quality. For a given data point $r \in S_r$, its $Q_r$ can be mathematically calculated following the steps below:

$$Q_r = \frac{b_r - a_r}{\max(a_r, b_r)}$$

$$a_r = \frac{1}{|S_r| - 1} \sum_{s \in S_r, r \neq s} d(r, s)$$

$$b_r = \min \left[ \frac{1}{|S_v|} \sum_{v \in S_v} d(r, v) \right]$$

\hspace{1cm} (7)
where \(|S_r|\) is the number of members in cluster \(S_r\) (i.e. cardinality); \(S_v\) is any other cluster in the dataset; data point \(v\) belongs to \(S_v\); \(d\) is the Euclidean distance between two data points measured by (6).

Equation (7) evaluates both the compactness and separation of produced clusters by K-means: for compactness, \(a_r\) is the average distance of data point \(r\) to all other points in the same cluster \(S_r\). It reflects the intra-cluster compactness; \(b_r\) is the smallest average distance of \(r\) to all points in every other cluster that does not contain \(r\). It reflects the inter-cluster separation. A big \(b_r\) indicates a large inter-cluster separation seen from point \(r\); in the end, \(Q_r\) combines \(a_r\) and \(b_r\). A good intra-cluster compactness and inter-cluster separation together will results in a big \(Q_r\).

(7) is the calculation for any single data point \(r\). To evaluate the clustering quality of the entire dataset, average Silhouette coefficient is used and is given as below:

\[
Q_{avg} = \frac{1}{m} \sum_{i=1}^{m} Q_i
\]

where \(m\) is total number of data points in this dataset.

The steps of using Silhouette analysis to determine optimal cluster number \(K\) are given as follows:

**Algorithm 2: Silhouette Analysis**

**Input:** \(D=\{x_1, x_2, ... x_m\}\) # dataset contains \(m\) data points  
\(K \in \{1, 2, ... N\}\) # Initial range for \(K\)  
**Output:** Optimal cluster number \(K_p\)

1: for \(i \leftarrow 1\) to \(N\)  
2: Apply K-means clustering (assuming \(K=i\))  
3: Calculate \(Q_r\) for each \(x \in D\)  
4: Calculate \(Q_{avg}\) for \(D\)  
5: end for  
6: \(K_p \leftarrow K\) with maximum \(Q_{avg}\)

Through K-means clustering and Silhouette analysis, a number of feeders can be grouped automatically based on their load compositions. An example of clustering 300 feeders to 4 clusters is shown as below:
Figure 2. Example of clustering 300 feeders to 4 clusters by load composition features

3. Feature Selection and Processing

As a hybrid LTLF approach, both area features (top-down) and feeder features (bottom-up) are incorporated into modelling. By employing domain knowledge, useful raw features related to distribution feeder LTLF are selected. They need to be processed before fed into the sequence learning step.

3.1. Area Features

Area features describe the overall drivers in the planning area. Typical area features are listed in Table 1. The economic and demographic features have been explained in detail in our previous paper [12]. “Extreme Temperature Above Average” feature is the difference between maximum (summer)/minimum (winter) temperature of the current year and the historical average such as the 10-year average [18-19].

| Feature Name                           | Category           |
|----------------------------------------|--------------------|
| Real GDP Growth (%)                    | Economy            |
| Total Employment Growth (%)            | Economy            |
| Industrial Production Index            | Economy            |
| Commodity Price                        | Economy            |
| Population Growth (%)                  | Demographics       |
| Net Migration                          | Demographics       |
| # of Housing Starts                    | Demographics       |
| Extreme Temperature Above Average     | Temperature        |
3.2. Feeder Load Features

Feeder load features describe the detailed feeder-level load information.

- Major Customer Net Load Change: this feature is the estimated net load change of all major customers on the feeder. Every year, some customers report growth while some report reduction. The aggregated net change is the summation of all these reported load changes from major customers on the feeder.

- Distributed Energy Resource (DER) and Electrical Vehicle (EV) Adoption Change: DER and EV can be considered in certain planning areas that have such concerns. Future DER and EV annual adoption can be forecasted in separate tasks and given as input [20-21].

- Previous-Year Peak Demand: the previous year’s summer or winter peak demand is required in this model. It provides a baseline reference while most of other area and feeder load features focus on the change from the forecast year.

3.3. Principal Component Analysis (PCA) for Area Features

Many economic and demographic area features are highly correlated. To improve the prediction accuracy, PCA shall be applied to reduce the dimensionality. This process has been explained in detail in [12].

3.4. Virtual Feeder Conversion for Feeder Features

When dealing with long-term historical feeder loading data, it is inevitable to encounter load transfer events which can suddenly disrupt the original trend of feeder loading and introduce interference. [12] proposed the idea of combining two or more feeders with load transfer events to one virtual feeder so that the transfers between them can be ignored. This processing technique can effectively eliminate the data noise caused by load transfers and continues to be used in this research.

4. Sequence Learning for LTLF

This section briefly reviews the theory of sequence learning models. We start from non-gated RNN and go on to explain gated RNN. Three different sequential configurations as well as their dataset formats are also discussed.

4.1. Non-gated RNN

As shown in Figure 3, an RNN is a group of feed-forward neural networks (FNN) connected in series. Hidden neurons of the FNN at a previous time step are connected with the hidden neurons of the FNN at the following time step. This can make hidden state at the last time step $H_{t-1}$ pass into the current time step.
$H_{t-1}$ is then combined with the current input $X_t$ to produce the current hidden state $H_t$ through trained weights $W_h$ and $W_x$. This process continues to the next time step until the end of the sequence. In this unique way, RNN is able to make use of historical information and does not treat one time step as an isolated point. This made RNN suitable for forecasting tasks such as word prediction and load forecast where the output of current time step is not only related to the current input but also previous time steps. An unfolded RNN structure is shown in Figure 4.

![Figure 3. Illustration of an unfolded RNN structure](image)

In spite of the obvious advantages, the training of non-gated RNN can be unstable due to an intrinsic problem called vanishing/exploding gradient [22-23]. During back propagation of RNN, gradient value may become too small to drive the network update or too large to stabilize the training. This problem leads to the invention of gated RNNs which successfully solve the vanishing/exploding gradient problem through sophisticated gate controls [23-25]. In recent years, gated RNNs have replaced non-gated RNN as the industry standard.

### 4.2. Gated RNN

![Figure 4. A GRU unit diagram](image)
GRU network and LSTM are two commonly used gated RNNs. In [12], we already compared LSTM and GRU through multiple tests and it is found GRU has a performance similar to LSTM but is much faster. This is consistent with comparison in [26]. Hence, in this paper, we only adopts GRU as the sequence learning model. In addition to the chain-structure of non-gated RNN, a GRU network has a GRU unit diagram shown in Figure 4.

An element of the reset gate \( r_t \) is calculated by:

\[
    r_t = \sigma(W_r \cdot [H_{t-1},X_t] + b_r)
\]

where \( [H_{t-1},X_t] \) is the concatenated vector of hidden state vector \( H_{t-1} \) at the previous time step and the input vector \( X_t \) at the current time step; \( W_r \) and \( b_r \) are the weight vector and bias.

An element of the update gate \( u_t \) is calculated in a very similar way, with different weight vector \( W_u \) and bias \( b_u \):

\[
    u_t = \sigma(W_u \cdot [H_{t-1},X_t] + b_u)
\]

According to the information flow illustrated in Figure 4, a temporary value \( h_t \) is calculated by:

\[
    h_t = tanh(W_h \cdot [r_t \otimes H_{t-1},X_t]+b_h)
\]

In the end of the information flow, the hidden state vector \( H_t \) at the current time step \( t \) is generated by using hidden state vector \( H_{t-1} \), update gate vector \( u_t \) and temporary vector \( h_t \) through pointwise multiplication and addition:

\[
    H_t = (1 - u_t) \otimes H_{t-1} + u_t \otimes h_t
\]

Through (9)-(12), hidden state \( H_t \) is updated from one time step to the next until the end of the sequence.

4.3. Sequential Configurations

An RNN can be implemented in different sequential configurations. For the multi-year forecast problem discussed in this paper, there are three suitable sequential configurations: many-to-one, many-to-many and shifted many-to-many [23-24]. Their schematics are shown in Figure 5.
Figure 5. Three different sequential configurations

The theoretical differences of these configurations lie in the network loss functions. For all neural networks, network loss is converted to gradient and drives the training of neural network through back propagation and mathematic chain-rule. Here, Mean Absolute Error (MAE) is chosen to construct the network loss function. For many-to-one configuration, its loss function $L_1$ is:

$$
L_1 = \frac{1}{n} \sum_{j=1}^{n} |Y_t^j - \hat{Y}_t^j|
$$

(13)

where $n$ is the training batch size; $Y_t^j$ is the forecast year $t$’s actual peak demand in $j_{th}$ record in the training batch; $\hat{Y}_t^j$ is the forecast year $t$’s forecasted peak demand in $j_{th}$ record in the training batch.

The loss function $L_2$ for many-to-many configuration is:

$$
L_2 = \frac{1}{nt} \sum_{j=1}^{n} \sum_{i=1}^{t} |Y_i^j - \hat{Y}_i^j|
$$

(14)

where $n$ is the training batch size; $Y_i^j$ is $i_{th}$ year ’s actual peak demand in $j_{th}$ record in the training batch; $\hat{Y}_i^j$ is $i_{th}$ year ’s forecasted peak demand in $j_{th}$ record in the training batch.

Both the above many-to-one and many-to-many configurations are recursive forecasting configurations as they only forecast the next one-year [12]. This means in order to forecast multiple years ahead, for example the next 3 years, $Y_t$ is firstly forecasted and then it is used as a known input to predict $Y_{t+1}$; similarly, $Y_{t+1}$ is then used to forecast $Y_{t+2}$. Different from these two configurations, the shifted many-to-
many configuration forecasts $Y_t, Y_{t+1}$ and $Y_{t+2}$ all at once. It is more efficient for multi-year forecast because it is not a recursive forecasting method. Its loss function $L_3$ is:

$$L_3 = \frac{1}{nT} \sum_{j=1}^{n} \sum_{i=t}^{t+T-1} |Y_i^j - \hat{Y}_i^j|$$

(15)

where $n$ is the training batch size; $Y_i^j$ is $i_{th}$ year’s actual peak demand in $j_{th}$ record in the training batch; $\hat{Y}_i^j$ is $i_{th}$ year’s forecasted peak demand in $j_{th}$ record in the training batch; $T$ is the length of forecast horizon in number of years.

4.4. Sequential Datasets

To fit into the above three RNN sequential configurations, data records must be grouped by a fixed number of time steps following specific formats. This is different from traditional single-row datasets that are commonly used for other types of supervised learning methods. Table 2 to Table 4 are examples for many-to-one, many-to-many and shifted many-to-many configurations.

For explanation purpose, we take “Data Record ID1” as an example. In many-to-one configuration, the goal focuses on the forecast of 2011’s peak load by using previous-year peak demand in 2008, 2009 and 2010 as well as the yearly economic and temperature features in 2009, 2010 and 2011 (forecasted). EP1 and EP2 are the processed economic-population growth features after applying PCA; in many-to-many configuration, the goal is not only to forecast 2011’s peak load but also reproduce peak load in 2009 and 2010. The loss function takes all three years’ errors into consideration and is therefore less biased if 2011’s loading experienced an unusual change. However, compared to many-to-one, its focus on year 2011 is weaker; in shifted many-to-many configuration, the goal is not only to forecast 2011 but also 2012 and 2013 all at once.

Table 2: Dataset Example for Many-to-One Configuration

| Data Record ID | Feeder ID | Input Year | Previous-Year Peak Demand | Yearly Features | Forecast Year - Actual Peak Demand |
|----------------|-----------|------------|---------------------------|----------------|------------------------------------|
|                |           |            |                           |                |                                    |
|                |           |            |                           |                | 2011- 550 A                       |
| 1              | 1001      | 2009       | 433 A                     | EP1 -0.64      | 0.7°C 42 A                        |
|                | 1001      | 2010       | 502 A                     | EP2 0.44       | -1.3°C 34 A                       |
|                | 1001      | 2011       | 554 A                     | Extreme Temperature Over Average 3.4°C 0 A           |
| 2              | 1001      | 2010       | 502 A                     | -0.16 0.31     | -1.3°C 34 A                       |
|                | 1001      | 2011       | 554 A                     | -0.31 3.4°C    | 0 A                               |
|                | 1001      | 2012       | 550 A                     | -0.06 -0.17    | -2.2°C -21 A                      |
| ...            | ...       | ...        | ...                       | ...            | ...                                |

...
Table 3: Dataset Example for Many-to-Many Configuration

| Data Record ID | Feeder ID | Input Year | Previous-Year Peak Demand | Yearly Features | Forecast Year - Actual Peak Demand |
|----------------|-----------|------------|---------------------------|----------------|----------------------------------|
|                |           |            |                           | EP1 | EP2 | Extreme Temperature Over Average | Major Customer Net Load Change |                          |
| 1              | 1001      | 2009       | 433 A                     | -0.64 | 0.44 | 0.7°C                             | 42 A                          | 2009-502 A                |
|                | 1001      | 2010       | 502 A                     | -0.16 | 0.31 | -1.3°C                            | 34 A                          | 2010-554 A                |
|                | 1001      | 2011       | 554 A                     | 0.33  | -0.31 | 3.4°C                             | 0 A                           | 2011-550 A                |
| 2              | 1001      | 2010       | 502 A                     | -0.16 | 0.31 | -1.3°C                            | 34 A                          | 2010-554 A                |
|                | 1001      | 2011       | 554 A                     | 0.33  | -0.31 | 3.4°C                             | 0 A                           | 2011-550 A                |
|                | 1001      | 2012       | 550 A                     | -0.06 | -0.17 | 3.4°C                             | 0 A                           | 2011-537 A                |

Table 4: Dataset Example for Shifted Many-to-Many Configuration

| Data Record ID | Feeder ID | Input Year | Previous-Year Peak Demand | Yearly Features | Forecast Year - Actual Peak Demand |
|----------------|-----------|------------|---------------------------|----------------|----------------------------------|
|                |           |            |                           | EP1 | EP2 | Extreme Temperature Over Average | Major Customer Net Load Change |                          |
| 1              | 1001      | 2009       | 433 A                     | -0.64 | 0.44 | 0.7°C                             | 42 A                          | 2011-550 A                |
|                | 1001      | 2010       | 502 A                     | -0.16 | 0.31 | -1.3°C                            | 34 A                          | 2012-521 A                |
|                | 1001      | 2011       | 554 A                     | 0.33  | -0.31 | 3.4°C                             | 0 A                           | 2013-537 A                |
| 2              | 1001      | 2010       | 502 A                     | -0.16 | 0.31 | -1.3°C                            | 34 A                          | 2012-521 A                |
|                | 1001      | 2011       | 554 A                     | 0.33  | -0.31 | 3.4°C                             | 0 A                           | 2013-537 A                |
|                | 1001      | 2012       | 550 A                     | -0.06 | -0.17 | -2.2°C                            | -21 A                         | 2014-549 A                |

5. Learning of Best-performing Sequential Configuration

For different feeders, the three sequential configurations presented in Section 4 may have different performances. For example, many-to-one configuration places all emphasis on the next one-year and can be more accurate if the external drivers and customer load information (including assumed values for future years) are accurate; many-to-many configuration is less biased when dealing with feeders with large yearly fluctuation as it considers three years’ average performance; shifted many-to-many configuration can be more accurate for saturated feeders which are less related to external drivers and feeders which have inaccurate customer load information. This is because it is not a recursive forecasting method and relies on less future information for multi-year forecast - when accuracy cannot be guaranteed, using less information can become an advantage.

However, it is difficult to manually analyze each feeder’s characteristics and manually selects the best-performing sequence. This paper proposes an automatic way of learning the best-performing Sequential Configuration for individual feeders. Figure 6 shows a training set example with 20-year data. The goal is to produce a three-year window forecast. This window slides from the fourth year to eighteenth year and for
each slide, a MAE is calculated for each sequential configuration. Therefore, it produces in total 15 MAEs for performance evaluation.

\[
P = \frac{1}{N - 2T'} + \sum_{j=1}^{N-2T'+1} \sum_{i=1}^{T'} |Y_i^j - \hat{Y}_i^j|
\]

where \( N \) is the total number of years in training set; \( T' \) is the forecast window length; \( Y_i^j \) is the actual peak demand of \( i_{th} \) year in the \( j_{th} \) window; \( \hat{Y}_i^j \) is the forecasted peak demand of \( i_{th} \) year in the \( j_{th} \) window.

For each feeder, (16) is used to calculate performance indexes for all three sequential configurations. The sequential configuration with the smallest \( P \) index is selected and labeled for this particular feeder. This configuration will be used on test set and future forecast for the feeder.

6. Test and Case Studies

The discussed method was applied to an urban distribution system in West Canada to forecast a three-year window peak demand in both summer and winter. In total 403 distribution feeders and their past 18-year annual data (2001-2018) were used to create the dataset. 2001-2015 data were used to form the training set and 2016-2018 three-year data were used to form the test set. Feeders were clustered into 4 groups using the proposed clustering techniques. For each sequential configuration, a neural network with a 5-neuron input layer and a 10-neuron GRU hidden layer is implemented. For the output layer, many-to-one configuration uses 1 neuron while many-to-many configuration and shifted many-to-many configuration use 3 neurons. All neurons use ReLU activation function. A 20% neuron dropout rate is also applied to avoid overfitting. Since Mean Absolute Percent Error (MAPE) is a common metric for measuring prediction error for a group of objects, we use the average of MAPE (AMAPE) in the three test years 2016-2018 to evaluate the forecast performance for each feeder. Mathematically, it is given by:
where $m$ is the number of feeders in the planning area; $\hat{Y}_i^j$ is the actual peak demand of $i_{th}$ feeder in $j_{th}$ year (starting in 2016); $Y_i^j$ is the forecasted peak demand of $i_{th}$ feeder in $j_{th}$ year.

6.1. Effect of Selective Sequence Learning

To test the effect of the proposed selective sequence learning (SSL) method, the proposed method is compared to using only single sequential configurations for both summer and winter. It should be noted that single sequential configuration many-to-one and many-to-many are the configurations used in [12]. The results are shown in Figure 7.

![Figure 7. Performance comparison with using single sequential configuration](image)

As can be seen, the selective sequence learning method demonstrates a lower AMAPE against all single configurations. It is expected as each feeder now learns the best-performing Sequential Configuration during the training process and uses it for forecast.

6.2. Effect of Feeder Clustering

Another contribution of this paper is the proposed feeder clustering step. This step groups feeders with similar load compositions together. The rationale behind this is that feeders with similar load compositions respond to external drivers in similar ways and therefore have similar growth patterns. Therefore, if separate
learning can be established for separate groups, better performance could be achieved. This is proven by comparing AMAPE after using feeder clustering and AMAPE without using feeder clustering (treating all area feeders as one group). The results are shown in Figure 8.

Figure 8. Comparison of AMAPE after feeder clustering and AMAPE without feeder clustering

6.3. Comparison with Traditional Models

In the end, the proposed model was compared to various other models specified as below. To ensure fairness, the same feature processing techniques discussed in Section 3 are also used for traditional models.

- Bottom-up model: as discussed in Section 1, this is the method to build a feeder load forecast based on major customer load information. Mathematically, the “Major Customer Net Load Change” feature from 2016 to 2018 was gathered and added to the previous year’s peak demand recursively to approximately estimate the following year’s peak demand.

- ARIMA model: for each feeder, its peak demand data between 2001 and 2015 were fed into an ARIMA model for training. ARIMA (2,0,0) was chosen because it gives the best forecast result among different ARIMA order parameters for the dataset. Then the peak demand values between 2016 and 2018 were calculated recursively.

- One-input year recursive FNN (ORF): for each feeder, only one-input year’s features are used for forecast. The features are the same as used in the proposed method, i.e. “Previous-Year Peak Demand”, EP1, EP2, “Extreme Temperature over Average” and “Major Customer Net Load Change”. A
A traditional FNN model is used, with a 5-neuron input layer, two 6-neuron hidden layers and a 1-neuron output layer. Peak demand values between 2016 and 2018 were forecasted recursively.

- Three-input year recursive FNN (TRF): a traditional FNN model is used to incorporate all the features from three input years (in total 15 features) to forecast the third year’s peak demand. The FNN has a 15-neuron input layer, two 10-neuron hidden layers and a 1-neuron output layer. Peak demand values between 2016 and 2018 were forecasted recursively. The main difference between this method and the proposed many-to-one sequence prediction is that it uses the FNN structure instead of RNN structure.

- Three-input year non-recursive FNN (TNF): a traditional FNN model is used to incorporate all the features from three input years (in total 15 features) to forecast the three forecast years’ peak demand. The FNN has a 15-neuron input layer, two 12-neuron hidden layers and a 3-neuron output layer representing peak demands in three years. Peak demand values between 2016 and 2018 were forecasted all at the same time. The main difference between this method and the proposed shifted many-to-many sequence prediction is that it uses the FNN structure instead of RNN structure.

The comparison is summarized in Figure 9. As can be seen, the proposed method outperforms all 5 non-sequence prediction based traditional methods.

![Figure 9. Performance comparison with 5 non-sequence prediction based traditional methods](image)

7. Conclusions

This paper discussed about a multi-year long-term load forecast method for area distribution feeders. This method is based on selective sequence learning. It has the following obvious advantages:
• As a hybrid LTLF method, it can seamlessly integrate top-down area features and bottom-up feeder features to improve forecasting performance;

• As a sequence prediction method, it can comprehend multiple years of historical data to improve forecasting performance;

• It combines unsupervised learning and supervised learning techniques. Area feeders are clustered by load composition to improve forecasting performance;

• Three sequential configurations can be learned and alternatively selected for different feeders to improve forecasting performance;

• It can produce multi-year forecast results in the future.

The proposed method was tested on an actual urban distribution system in West Canada and achieved better performance than traditional methods and our previous non-selective sequence prediction method.

References

[1] H. L. Willis, Power Distribution Planning Reference Book, CRC press, 1997.

[2] F. Elkarmi, ed., Power System Planning Technologies and Applications: Concepts, Solutions and Management, IGI Global, 2012.

[3] W. Simpson and D. Gotham, “Standard approaches to load forecasting and review of Manitoba Hydro load forecast for needs for and alternatives To (NFAT),” The Manitoba Public Utilities Board, Winnipeg, MB, Canada. [Online] http://www.pub.gov.mb.ca/nfat/pdf/load_forecast_simpson_gotham.pdf.

[4] “Peak forecasting methodology review,” Midwest ISO, Carmel, IN, USA. [Online] https://legalectric.org/f/2011/10/peak-forecasting-methodology-review-whitepaper-2011.pdf.

[5] “Electric load forecast: Fiscal 2013 to Fiscal 2033,” BC Hydro, Vancouver, BC, Canada. [Online] https://www.bchydro.com/content/dam/BCHydro/customer-portal/documents/corporate/regulatory-plFNNing-documents/integrated-resource-plans/current-plan/2012-electric-load-forecast-report.pdf.

[6] T. Al-Saba and I. El-Amin, “Artificial neural networks as applied to long-term demand forecasting,” Artificial Intelligence in Engineering, 13(2), pp.189-197, 1999.
[7] L. Ekonomou, “Greek long-term energy consumption prediction using artificial neural networks,” Energy, 35(2), pp.512-517, 2010.

[8] M. Askari and A. Fetanat, “Long-term load forecasting in power system: Grey system prediction-based models,” Journal of Applied Sciences, 11(16), pp. 3034-3038, 2011.

[9] J. Wang, L. Li, D. Niu and Z. Tan, “An annual load forecasting model based on support vector regression with differential evolution algorithm,” Applied Energy, 94, pp.65-70, 2012.

[10] D. Burillo, M. V. Chester, S. Pincetl, E. D. Fournier and J. Reyna, “Forecasting peak electricity demand for Los Angeles considering higher air temperatures due to climate change,” Applied Energy, 236, pp.1-9, 2019.

[11] K. Train, J. Herriges and R. Windle, “Statistically adjusted engineering models of end-use load curves,” Energy, 10(10), pp. 1103-1111, 1985.

[12] M. Dong and L. S. Grumbach, “A Hybrid Distribution Feeder Long-Term Load Forecasting Method Based on Sequence Prediction,” IEEE Transactions on Smart Grid, 2019, Early Access, DOI: 10.1109/TSG.2019.2924183.

[13] S. Mori, M. Nishimura, and N. Itoh, “Systems and methods for natural spoken language word prediction and speech recognition,” International Business Machines (IBM) Corp., U.S. Patent 7,359,852, 2008.

[14] R. Sun and C. L. Giles, “Sequence learning: from recognition and prediction to sequential decision making,” IEEE Intelligent Systems, 16(4), pp. 67-70, July-Aug. 2001.

[15] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.

[16] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-means clustering algorithm,” Applied Statistics, 28(1), pp. 100-108, 1979.

[17] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” Journal of Computational and Applied Mathematics, 20, pp.53-65, 1987.
[18] “Electricity use during cold snaps,” Hydro Quebec, Montreal, QB, Canada. [Online] http://www.hydroquebec.com/residential/customer-space/electricity-use/winter-electricity-consumption.html.

[19] “Historical Climate Data,” Government of Canada, Canada. [Online] http://climate.weather.gc.ca/.

[20] “Distributed Energy Resources Customer Adoption Modeling with Combined Heat and Power Application,” Berkeley Lab, Berkeley, CA, USA. [Online] https://escholarship.org/uc/item/874851f9.

[21] A. Soltani-Sobh, K. Heaslip, A. Stevanovic, R. Bosworth and D. Radivojevic, “Analysis of the Electric Vehicles Adoption over the United States,” Transportation Research Procedia, 22, pp. 203-212, Jan.2017.

[22] J. F. Kolen and S. C. Kremer, “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies,” A Field Guide to Dynamical Recurrent Networks, IEEE, 2001.

[23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, 9, pp.1735-1780, 1997.

[24] I. Sutskever, O. Vinyals and Q. V. Le. “Sequence to sequence learning with neural networks,” Advances in Neural Information Processing Systems, pp. 3104-3112, 2014.

[25] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv Preprint, arXiv:1406.1078, 2014.

[26] J. Chung, C. Gulcehre, K. Cho and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv Preprint, arXiv:1412.3555, 2014.