A Scaling Law of Short Term Electricity Load Forecasting on Varying Levels of Aggregation

Raffi Sevlian and Ram Rajagopal

Abstract—We propose a simple empirical scaling law that describes load forecasting accuracy at different levels of aggregation. The model is justified based on a simple decomposition of individual consumption patterns. We show that for different forecasting methods, aggregating more customers improves the relative forecasting performance up to specific point. Beyond this point, no more improvement in relative performance can be obtained. A benchmarking procedure for applying the scaling law to different forecasting models is presented. The model and procedure are evaluated with experimental data.

I. INTRODUCTION

To meet the challenges posed by a significant increase in distributed energy resources, there is a growing need for active control at the distribution level. The emergence and deployment of improved sensing and control technologies allow a variety of applications in the distribution system. As an example, smart meter data has been proposed to be used in various planning and operational applications. In applications such as outage detection or state estimation, load forecasts of 1 hour up to a day ahead are needed to provide pseudo-measurements. For the design of the next generation of distribution system applications, understanding the variability of these pseudo-measurements is of importance. The focus of this paper is the following: relying on empirical analysis and theoretical modeling to develop a simple intuitive scaling law for load-forecasts on varying levels of aggregation.

The field of load forecasting is very mature with numerous methodologies having been proposed throughout the years. These works have focused primarily on the level of large substations servicing tens of megawatts or up to an entire transmission system which has a load of tens of gigawatts. With recent advances in communication infrastructure for remote measurement and automated metering, there is an abundance of new data from homes and commercial buildings. The proliferation of this more granular data has led to an increase in forecasting research at these lower levels aggregation. Typical home loads are 1 to 2 kWh, while commercial buildings can be 100 times that amount. The relative forecasting errors typically seen at the level of substations and power systems has been quite low (1% – 2%). Also, recent work showing forecasting performance at the individual level show much higher errors (up to 30%).

Clearly, there is a discrepancy in forecasting performance at the level of individuals and of the entire power grid. This is the effect of load aggregation on forecasting loads. Our work aims to quantify the effect of aggregation by proposing a scaling law relating kWh load level and forecasting error.

Using data from hundreds of thousands of residential customers and commercial buildings, we construct datasets of varying aggregation levels and verify the proposed model. The laws are obtained by assuming a simple underlying consumption model for each individual then decomposing the aggregate error into bias and variance terms. We observe that the law holds for different methods and can be utilized to compare methods across datasets.

The paper is organized as follows. Section III develops a model for aggregation and proposes a set of scaling laws for two common forecasting metrics. The scaling laws are then verified in Sections IV and V including a review of the relevant literature. Section VI concludes the paper and discusses some future work.

II. LITERATURE REVIEW

The literature on load forecasting is very extensive. We focus our survey on both general methods used in short term forecasting in Section II-B as well as individual work compared in Section V-E.

A. Short Term Load forecasting

A general overview of short term load forecasting state of the art is provided in [16], [15]; and more classic surveys are given in [12], [30] and [13]. We review a few popular methods that we utilize as procedures in this paper. Seasonal ARMA and other linear modeling approaches are considered in [25]. Seasonal Vector modeling with segment identification is considered first in [9]. Neural networks have been applied to load forecasting for quite some time. Some early papers are [20], [31], [5], [21]. In [14], the author provides a comprehensive survey of the current state of Neural networks applied to load forecasting. Support vector regression has recently been applied to load forecasting as well with substantial work done by [23] and [24].

B. Hour Ahead Forecasting on Individual and Large Aggregates

Recent work shows a fundamental limitation to the predictability of individual customers. [8] Performs one hour ahead forecasting based on hourly data utilizing machine learning. The methods achieve MAPE of 1.61% to 13.41% for a 700kWh commercial building and between 15% to 30% for three homes with mean consumption close to 1.5kWh. In [32], machine learning methods are compared on data from three homes with mean consumption 1 to 2kWh achieving MAPE close to 25%. In [29], various methods are utilized to forecast peak demand for individual homes. The authors conclude that seasonal autoregressive models achieve the best performance, with MAPE of 30%. In [11], a Kalman filter based forecaster is applied to single home data with mean consumption of 0.8kWh to achieve MAPE of 30%.

Very low errors are reported at high aggregation levels. In [1] the authors use an artificial neural network to forecast a mean
load of 2.5 GWh, with MAPE ranging from 1.73% to 3.02%. In [6] the authors apply wavelet multi-scale decomposition based autoregressive approaches. They report MAPE values of 0.7% to 3.5% depending on the method used on a dataset with mean load 9GWh. In [26] artificial neural networks are applied to data with a mean consumption of 800 MWh achieving MAPE errors from 1.11% to 1.63%. Similarly, [2] obtains MAPE in the range 0.81% to 1.21% utilizing neural networks on a 8GWh load data. In [7] a novel ANN architecture is applied to two datasets with peak load 4.4GWh and report MAPE between 0.8% and 1.5%. Finally, [20] applies artificial neural networks to attain an error rate of 1.7% for a load of 7 GWh. In the experimental comparisons in this paper we benchmark hour ahead forecasting of one hour intervals of consumption.

C. Recent Work on Aggregation Forecasting

Initial work on developing a model for forecasting accuracy with an explicit scaling law was done by the authors in [27]. The model was limited to 2,000 residential customers and was unable to capture scaling behavior at large aggregate levels. Other work also rely on small datasets and show similar results as in [27]. In [22] the authors aggregate up to 1000 customers for one hour ahead forecasting and provide a qualitative rationale for the effect. In [17] the authors demonstrate an empirical plot of normalized root mean squared error against number of customers and show it decreases. This work aggregates up to 782 homes. In [28] the authors show that mean absolute percentage error (MAPE) decreases with the number of customers and use it for examining electricity market trading performance.

This paper differs from prior work since we extend the aggregation to over 100,000 customers (100 MWh) and point out the crucial point that errors no longer improve beyond a critical load. We then propose a statistical linear consumption model based on data results from smart meter clustering analysis [19]. This model is used to derive a benchmark formula that describes the relationship between MAPE or the coefficient of variation and the aggregation size. The model is fit to experimental performance data of several forecasters to uncover aggregation relationships. To the best of our knowledge, this is the first paper modeling forecasting error scaling with aggregation size. This work connects individual consumption models to aggregate forecasting and provides a simple mechanism to benchmark any forecasting algorithm as it is applied to varying levels of aggregation.

III. MODELING LOAD AGGREGATION

Aggregation reduces the inherent variability in electricity consumption resulting in increasingly smooth load shapes. Figure 1 illustrates this effect where it is clear that the higher aggregation levels are easier to predict. A colloquial explanation is that the law of large numbers smooths out the signal, therefore justifying why gigawatt level forecasting is very accurate. Yet, it is less clear how to quantify the improvements in forecasting. The main goal of this paper is to develop an appropriate scaling model for forecasting performance with respect to aggregation size. In particular, we identify the most appropriate way to measure aggregation and then propose a simple model to explain the scaling phenomena. We start by reviewing typical performance metrics used for quantifying forecasting performance.

A. Forecast Accuracy Performance Metrics

The two performance metrics most commonly used in forecasting literature are Coefficient of Variation (CV) and Mean Absolute Percentage Error (MAPE). Coefficient of variation measures the ratio of the prediction error standard deviation to the signal mean. Consider two time series $x(t)$ and its forecast $\hat{x}(t)$ for $t = 1, ..., T$. The empirical coefficient of variation (CV) measures the difference between these time series and is computed as

$$ CV(x, \hat{x}) = \frac{100}{\frac{1}{T} \sum_{t=1}^{T} x(t)} \left( \frac{1}{T} \sum_{t=1}^{T} \frac{|x(t) - \hat{x}(t)|}{x(t)} \right) (\%) . \tag{1} $$

Likewise the mean absolute percentage error (MAPE) is defined as

$$ MAPE(x, \hat{x}) = \frac{100}{T} \sum_{i=1}^{T} \left( \frac{x(t) - \hat{x}(t)}{x(t)} \right) (\%) . \tag{2} $$

CV and MAPE are relative error metrics traditionally reported in the literature although MAPE is more commonly reported. It is assumed they allow comparison of performance in different datasets. This paper investigates whether this fact is true for the case of electricity consumption.

B. Forecasting Scaling Laws

Consider a set of $N$ customers with consumption given by a time-series $x_n(t)$. The mean consumption for each customer is $W_n = \frac{1}{T} \sum_{t=1}^{T} x_n(t)$. We select a subset $A \subseteq \{1, 2, \ldots, N\}$ of customers and form a group that consumes according to the aggregate-time-series

$$ x_A(t) = \sum_{n \in A} x_n(t), \tag{3} $$

Fig. 1. Hourly electricity consumption for various aggregation levels. Consumption pattern of a single customer generally has little structure to be exploited. Aggregating more and more customers “smoothes” the signal so that it can be more predictable. Aggregation level of 20 or more residential customers shows a predictable pattern. Plots are not in the same scale.
with mean consumption
\[ W_A = \sum_{n \in A} W_n. \] (4)

In load forecasting, we build a predictor for \( x_A(t) \) that outputs the predicted sequence \( \hat{x}_A(t) \) and evaluate \( CV(x_A, \hat{x}_A) \) and \( MAPE(x_A, \hat{x}_A) \). Suppose such forecaster is built for every such group of customers. Then the expected coefficient of variation of this group of customers can be bounded according to Theorem 1.

**Theorem 1.** Consider all sets \( A \) of consumers \( x_A(t) \) with mean consumption \( W_A = W \). The average coefficient of variation at the \( W \) level of aggregation is upper bounded by
\[ E[CV(x_A, \hat{x}_A)|W_A = W] \leq \sqrt{\alpha_0 + \frac{\alpha_1}{W}}, \] (5)

where the expectation is taken over all sets \( A \) of mean consumption \( W \).

**Proof:** See Appendix B.

Theorem 1 gives a set of sufficient conditions under which the coefficient of variation can be bounded from above. However this motivates an empirical scaling law which fits the expected value of CV and MAPE based on the theoretical upper bound. We postulate that the population average \( MAPE \) scales as a function of \( W_A \) according to
\[ CV(W) = E[CV(x_A, \hat{x}_A)|W_A = W] \]
\[ = \sqrt{\frac{\alpha_0}{W^p} + \alpha_1} \, \% . \] (6)

Some key observations about this empirical law are:
(a) \( \overline{CV} \) is the average CV over all groups with mean consumption \( W \).
(b) The choice of using mean consumption as a proxy for group size is deliberate. An alternative choice would be the number of consumers in the group. But we observe that in fact customers with higher absolute consumption tend to be more predictable.
(c) \( \alpha_0/W^p \) measures improvement in relative error from aggregation. It captures the reduction in underlying variance due to aggregation and \( \alpha_1 \) is the irreducible relative error.
(e) The exponent \( p \) measures the rate of reduction of relative variance due to aggregation.
(f) The ideal aggregation condition occurs when \( \alpha_1 = 0 \) and \( p = 1 \).

Assuming \( p = 1 \), the proposed scaling law segments the forecasting problem into two regimes:

1) **Scaling:** When \( \alpha_0/\sqrt{W} \gg \alpha_1 \), relative error improves considerably due to aggregation. Eq. (7) can be approximated as
\[ CV(W) = \sqrt{\frac{\alpha_0}{W}}. \]

2) **Saturation:** When \( \alpha_0/\sqrt{W} \ll \alpha_1 \), there is no improvement in forecasting from aggregation and \( CV(W) \approx \sqrt{\alpha_1} \).

We can extend our proposed empirical model to MAPE calculation as well. The population average \( MAPE \) scales as a function of \( W_A \) according to
\[ \overline{MAPE}(W) = E[MAPE(x_A, \hat{x}_A)|W_A = W] \]
\[ = \sqrt{\frac{\beta_0}{W^p} + \beta_1} \, \%. \] (8)

We focus our experimental analysis on MAPE since it is more often reported in forecasting literature. The theoretical analysis of MAPE is left for future work.

**C. Aggregation benchmarking**

We introduce a benchmarking procedure that can be applied to characterize the aggregation performance of a forecasting algorithm. The workflow diagram is shown in Figure 2, and is applied to a specific selected forecaster using the following steps:

1) The dataset to be analyzed is selected. For example we focus on the PG & E Northern California region.
2) Second, the set of aggregates are generated using Eq. (3).
3) The actual forecast \( x_A(t) \) is computed for each aggregate signal and each time \( t \) utilizing the selected forecasting algorithm.
4) Error metrics are computed with either \( CV(x, \hat{x}) \) or \( MAPE(x, \hat{x}) \).
5) Aggregation error curves are generated by fitting the proposed scaling law

**IV. EXPERIMENT SETUP**

**A. Description of Data**

This study uses anonymized electricity data for residential customers of Pacific Gas & Electric (PG&E). The data spans a one year period from 08/01/2010 to 07/31/2011. The residential data captures hourly electricity consumption for 116,423 customers. The data for each customer is temporally aggregated to form hourly interval measurements. Both datasets cover more than 408
zip codes and a variety of climate zones and business types. Figures 3(a) displays a week long consumption of an individual household customer. The mean consumption of the data is of importance for this work. Figure 3(b) shows the population distributions of mean hourly consumption for households. The hourly population mean consumption for residential customers is 1.05 kWh.

B. Generating Aggregate Consumption

Residential aggregate consumption time series were generated by forming groups of randomly selected customers. Fifty six group sizes were chosen ranging from one to 100,000 customers. Fifty random groups for each size were generated by uniformly selecting customers. The mean hourly consumption of these groups ranged from 1kWh to 100MWh. The largest mean hourly consumption for each size ranged from 3 kWh to 180 MWh.

C. Forecasting models

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}_1$</td>
<td>SARMA(1,0)×(1,0)$_{24}$</td>
</tr>
<tr>
<td>$\mathcal{M}_2$</td>
<td>SARMA(2,0)×(1,0)$_{24}$</td>
</tr>
<tr>
<td>$\mathcal{M}_3$</td>
<td>SARMA(3,0)×(1,0)$_{24}$</td>
</tr>
<tr>
<td>$\mathcal{M}_4$</td>
<td>SVR - Radial Basis Function</td>
</tr>
<tr>
<td>$\mathcal{M}_5$</td>
<td>FFNN - Logistic Activation Function</td>
</tr>
</tbody>
</table>

The proposed scaling laws are studied using three commonly used methods for short term load forecasting: (SARMA) Seasonal Auto Regressive Moving Average, (SVR) Support Vector Regression and (FFNN) Feed Forward Neural Networks. The models and their training/testing procedures are described in detail in the Appendix.

V. EXPERIMENTAL RESULTS

A. Empirical MAPE with Aggregation Level

The first analysis investigates the performance of the SARMA(1,0)×(1,0)$_{24}$ ($\mathcal{M}_1$) in the one hour ahead prediction task. Aggregate time series $x_A$ were generated according to Section IV-B. For each time-series, the corresponding mean load $W_A$, forecast $\hat{x}_A$ and performance MAPE($x_A, \hat{x}_A$) were computed. The aggregation-error curve is the plot of the pairs $(W_A, \text{MAPE}(x_A, \hat{x}_A))$ for all generated groups A as shown in Figure 4. The scaling law in Eq. (8) was fit to this data using least-squares.

The fit (solid line) and the ideal aggregation (dashed line) scaling laws are displayed in Figure 4. Fit parameters are shown in Table II, III.

B. Critical Load $W^*$

The scaling law can be decomposed into scaling and saturation regimes as shown in Sec. III-B. The transition point between regimes is defined as the load aggregation level $W^*$ where the regime approximations are equal. The critical load $W_{\text{MAPE}}^*$ is the positive solution $W$ to

$$\sqrt{\frac{\beta_0}{W^p}} = \sqrt{\beta_1}. \quad (9)$$

$W_{\text{CV}}^*$ is defined in a similar way. The critical load for model $\mathcal{M}_1$ is 2.2MWh (Table III). The scaling regime extends from 1kWh to 2.2MWh aggregate loads, and the saturation regime extends from 2.2MWh to 100MWh.

C. Linear Scale Observations

Linear scale analysis investigates the forecasting performance scaling with respect to number of customers that form a group. Analyses with 2000 customers was presented in [27] and with 200 customers in [28]. However, analysis with such limited group sizes are unable to identify the various regimes present in the scaling law. More importantly, they also fail to identify the nature of the law that is driven by average kWh consumption rather then numbers. This observation is important when considering heterogeneous customers. If a 500kWh average consumer forms a group with ten 1 kWh consumers, the MAPE is predicted at a level 510 kWh.
D. Comparison of Different Models

In this section, we validate that the scaling law holds for a variety of typically utilized models in short term load forecasting. The scaling law parameters also provide a way to compare the performance of these different models. Five models were selected for comparison (Table I) after testing a variety of different choices. The resulting scaling law fits for MAPE and CV are shown in Tables III and II.

The variation of the scaling error $\sqrt{\alpha_0}$ and $\sqrt{\beta_0}$ between different models is small. The irreducible errors $\sqrt{\alpha_1}$ and $\sqrt{\beta_1}$ are quite different between the models. Combining these two important observations leads to identifying the irreducible errors as a fundamental performance metric for model comparison.

Using this metric, we see that with sufficient training, the SVR and FFNN models perform quite well. These models show irreducible CV errors of 1.961% and 1.338%, respectively, and MAPE errors of 1.210% and 1.423%. We should note however, that these models take considerably more training since a number of running parameters are fit in the validation step prior to a single test sample is evaluated.

The critical load value $W^*$ can be compared between different models. It can be seen to depend almost exclusively on the irreducible error since the reducible errors are close to each other. This observation leads to the conclusion that forecasters with low irreducible error benefit more from aggregation. Model $M_5$ has critical load 16MWh and its flat region in the scaling law is far to the right of the graph.

Table II, III show the fitted values as well as 95% confidence intervals which are computed by bootstrap resampling the experimental points. Two key points are shown by the confidence interval:

1) The confidence intervals of $\sqrt{\alpha_0}$, $\sqrt{\alpha_1}$ are quite wide and intersect in the intervals $\sqrt{\alpha_0} \in [50.0, 53.8]$ and $\sqrt{\alpha_1} \in [50.0, 50.6]$. This means that for example, a null hypothesis of $\sqrt{\alpha_0} = 52.4$ and $\sqrt{\alpha_1} = 50.3$ for all the models would not be rejected. This validates the analytical model relating this reducible term to a consumption profile independent parameter.

2) The value’s of $p$ for each fit are interesting since for an individual dataset, setting $p = 1$ leads to an unseemly fit. However, for half of the models, the 95% confidence interval contains the value $p = 1$. Therefore we conclude, that for a given dataset, generating an accurate fit of the aggregation-error curve requires a $p \neq 1$. However, we should keep in mind that there is no model basis for this parameter.

E. Short Term Load Forecasting Benchmarking

The MAPE for different studies can be compared utilizing the MAPE and the approximate aggregate size. We focus on selected papers that perform hour ahead forecasts of hourly intervals. Figure 6 displays the results and compares them to the scaling law obtained for model $M_3$. In the lower aggregation range, the results fall within the 95% confidence interval and at the GWh aggregation level the accuracy is related to that predicted by the model.

F. Multiple Hour Ahead Forecasting

This section tests the scaling law with respect to forecasting multiple hours ahead to verify its validity for different horizons. The model $M_3$ is utilized for this purpose. Figure 7(a) and Table IV displays the results. It is clear that the irreducible error increases with forecasting horizon, reducing the benefit from aggregation. It indicates that for more complex tasks such as day ahead forecasting, the forecasters need to be designed carefully to achieve low irreducible error.
when compared to each other.

to 1MWh) need to consider the improvements due to aggregating tasks of similar average sizes. Larger buildings can be thought of as an aggregation of smaller buildings, so the number of tasks scale linearly and so does average consumption. This leads to kWh providing the proper scaling for forecasting. Finally, data suggests that previous work agrees with model suggested here.

Forecasting studies in commercial buildings (mean loads 10 kWh) validate the choice of kWh average to drive the scaling raw. This observation is quite consistent despite the mean loads of residential building (residential or SMB) consumes electricity as a service.

An intuitive interpretation of the scaling law is that every 1 kWh 100 kWh 10 MWh 1 GWh, the number of tasks of similar average sizes. Larger buildings can be thought of as an aggregation of smaller buildings, so the number of tasks scale linearly and so does average consumption. This leads to kWh providing the proper scaling for forecasting. Finally, we note that for the SMB data the critical load is close to 10MWh. Forecasting studies in commercial buildings (mean loads 100kWh to 1MWh) need to consider the improvements due to aggregation when compared to each other.

H. Subpopulation Comparison

The benchmarking design in Section III-C is performed on the PG & E climate zone populations. First, the total population is split between the inland and coastal customers. Then the aggregates are randomly generated from each climate subpopulation. The climate subpopulation is separated by the climate zone that each customers zip code falls into. Following the California PG & E climate zone designations [10], we define "coastal populations" as those in climate zones 1, 2, 3, 4 and "inland populations" as those in climate zones 11, 12, 13, and 14. The total mean consumption available for the coastal and inland climate zone are 50 MWh and 80 MWh respectively. This corresponds to 43,558 coastal customers and 72,558 inland customers. We then generate aggregates ranging from single users to 40 thousand users for the coastal customers and 65 thousand for the inland customers.

Figure 8 shows the aggregation error curve for \( M_3 \) applied to the inland and coastal users. The results show under subpopulation breakdown, a similar aggregation error curve.

### Table IV
<table>
<thead>
<tr>
<th>Horizon (hours ahead)</th>
<th>MAPE ( \sqrt{\alpha_0} )</th>
<th>CV ( \sqrt{\beta_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72.5 1.28</td>
<td>56.3 1.06</td>
</tr>
<tr>
<td>2</td>
<td>85.2 5.30</td>
<td>75.1 4.26</td>
</tr>
<tr>
<td>3</td>
<td>50.2 8.92</td>
<td>86.6 8.35</td>
</tr>
<tr>
<td>4</td>
<td>68.3 10.94</td>
<td>78.6 10.11</td>
</tr>
</tbody>
</table>

### Table V

<table>
<thead>
<tr>
<th>Pop.</th>
<th>Model</th>
<th>( p )</th>
<th>( \sqrt{\alpha_0} )</th>
<th>( \sqrt{\alpha_1} )</th>
<th>( \sqrt{\beta_0} )</th>
<th>( \sqrt{\beta_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.P.</td>
<td>( M_1 )</td>
<td>0.91 (0.86 1.05)</td>
<td>63.76 (58.1 66.8)</td>
<td>2.48 (2.45 2.50)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M_2 )</td>
<td>0.96 (0.92 1.02)</td>
<td>61.14 (57.1 66.5)</td>
<td>1.43 (1.38 1.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M_3 )</td>
<td>0.90 (0.82 0.93)</td>
<td>61.90 (57.6 66.8)</td>
<td>1.39 (1.34 1.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I.P.</td>
<td>( M_1 )</td>
<td>0.97 (0.94 1.01)</td>
<td>47.06 (44.3 50.17)</td>
<td>1.94 (1.90 1.99)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M_2 )</td>
<td>0.95 (0.90 1.03)</td>
<td>46.32 (43.2 50.09)</td>
<td>1.35 (1.26 1.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( M_3 )</td>
<td>0.94 (0.89 0.98)</td>
<td>46.47 (42.9 49.95)</td>
<td>1.31 (1.22 1.40)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table V shows the tree linear models applied to each set
of aggregates. The linear models were chosen since they were much less computationally expensive to train and test. The results indicate that in both climate zones, model $M_3$ outperforms the other linear models. Also, the inland dataset has lower irreducible error in the inland dataset as opposed to the coastal dataset. This is the case even though the coastal dataset has a higher maximum mean load. Since the critical loads are 4.8 MWh and 6.2 MWh for inland and coastal populations, there more than enough samples in the irreducible regime.

I. Robustness of the Scaling Law

![Fig. 9. Comparing quantiles of forecast errors for each aggregation group. The scaling law is robust to the mechanism utilized to generate groups.](image)

It's important to verify whether the previous results were obtained due to the method of generating aggregates, that is following a random assignment. Here we compare the aggregation error curve for the aggregates with best performance and worst performance at each group level. The choice is determined by setting a quantile for MAPE error at each group size for residential data. Figure 9 displays the result. The scaling law is observed at the different performance quantiles, thus different aggregation mechanisms will obtain results similar to that reported in this paper.

VI. CONCLUSION

This paper introduces the idea of the effect of aggregation on load forecasting. We show that forecasting accuracy, as measured in relative error (MAPE/CV) improve with larger mean load until a critical load. We verify this model with empirical experiments. We also provide sufficient conditions leading to the observed aggregation-error curves introduced in the paper.

Currently, various papers focus on new model formulations to describe individual electricity consumers (e.g. [4], [3], [18]). These models can be utilized to justify a detailed understanding of how aggregate consumption patterns are formed and verified on higher resolution data. Moreover, novel ideas can be investigated for aggregate forecasting based on models induced by aggregating this individual consumption models.

The aggregation phenomena is also likely to be observed in other types of forecasting processes, such as for example day ahead load forecasting, wind forecasting and electric vehicle availability. Determining the scaling parameters for these problems is an important task as it can lead to new concepts on the limitations of forecasting big and small aggregates.

REFERENCES


APPENDIX

A. Forecasting models

1) Seasonal Auto Regressive Moving Average (SARMA): SARMA [2] predicts the electricity consumption in the next time step as a linear function of prior consumption values and forecast errors. Seasonality is considered by including additional predictors at a fixed prior period. A model SARMA$(p, q) \times (P, Q)_s$ has autoregressive (AR) order $p$ and moving average (MA) order $q$. It uses a seasonal component with a cycle of $s$ time steps, with AR order $P$ and MA order $Q$. This work considers a restricted class without MA component so $q = 0$ and $Q = 0$. The resulting model for the time-series $y(t)$ is

$$ y[t] = \sum_{k=1}^{P} \theta_k y[t-k] + \sum_{k=1}^{s} \phi_k y[t-sk] + \epsilon(t). $$

It is usual to assume $\epsilon(t) \sim N(0, \sigma^2)$ is an independent and identically distributed normal variable. Values of $p \geq 4$ are not recommended because it can lead to overfitting [3]. The seasonality is set to $s = 24$ hours and AR order $P = 1$. The adaptive SARMAX model learns the parameters $\theta$ and $\phi$ but keeps the parameters $p, P$ and $s$ fixed. The SARMA model is applied at each time step by learning the linear model using a pre-set model size. This constitutes an adaptive SARMA model.

2) Support Vector Regression: Support Vector Regression (SVR) works by building a non-linear learning method for a training dataset $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. The training set comprises of $N$ response $y_i$ and predictor $x_i$ pairs. The SVR data fitting method solves the following optimization:

$$ \min_{w, C, \xi, \epsilon} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) $$

subject to:

$$ y_i - w^T \Phi(x_i) - b - \epsilon \leq \xi_i, \quad i = 1, \ldots, N $$

$$ w^T \Phi(x_i) + b - y_i + \epsilon \geq \xi_i^*, \quad i = 1, \ldots, N $$

$$ \xi_i, \xi_i^* \geq 0, \quad 0 \leq \xi_i^* \leq \xi_i $$

Given training predictor $x_i$, there are given set of kernel functions $\Phi(x_i)$ which map $x_i$ to a high dimensional space. The kernel function is fixed and predictions are computed by:

$$ \hat{y}_i = w^T \Phi(x_i) + b. $$

The variables $w$, $\Phi$, and $b$ are used to map predictors to response. However, fitting the training data only generates vector $w$ and scalar $b$ subject to a set of constraints. The SVR will solve for $w$ such that it minimizes the sum of the norm of $w$: $\frac{1}{2} \|w\|^2$ as well a fitting error $C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$. The variables $\epsilon, \xi_i, \xi_i^*$ quantify the fitting error. Any deviation $|y_i - (w^T \Phi(x_i) - b)| \leq \epsilon$ incur no penalty. However, any deviation outside this dead band $(\xi_i, \xi_i^*)$ will incur linear cost yielding $C \sum_{i=1}^{N} (\xi_i + \xi_i^*)$. Under this model, the values of $C$, $\epsilon$, kernel function $\Phi$ and additional parameters to $\Phi$ must be specified.

The support vectors, as well as the constants $C$ and $\epsilon$ are learned adaptively in each training round. In the training data, we use $3/4$ of the data for training, and $1/4$ for validation of the support vector and constants. We should note that this method proves computationally expensive but outperforms statically trained models and SARMAX models usually. For a given kernel function, cross validation is performed to determine the parameters specific to the kernel. Then using a moving window we train the SVR and forecast one sample ahead as done in the SARMA model.

3) Feed Forward Neural Network (FFNN): FFNNs (e.g. [1]) provide a popular alternative to define the nonlinear map between $x(t)$ and $y_k(t - 1)$ used in the SVR model description. They are subset of artificial neural networks where neurons with a chosen activation function connect to each other in layers without feedback. The number of neurons, layers, choice of activation function and network parameters are learnt from the training set.

In this model, the training data is used with the 3/4, 1/4 split to learn model parameters like in the SVR case.

B. Analytic Model of Aggregation

In this section we propose a theoretical analysis to validate the empirically tested scaling law and observed aggregation-error curves. We present a set of sufficient conditions for the underlying consumption stochastic process that lead to the proposed scaling. The CV error metric is chosen due to its mathematical properties that leads to a simple analysis. The analysis is developed in multiple steps: a general individual behavioral model, a statistical analysis of aggregate forecasting, and specific statistical assumptions on the components of the behavioral model.

1) Individual Consumption Behavior: Electricity consumption for individuals $n$ for a specific time is given by $x_n(t)$ and the profile spans a 24 hour period. The consumption is decomposed as $x_n(t) = P_n(t) + e_n(t)$. The individual chooses a daily profile for each day $t \in R^T (T = 24)$ and deviates from it according to $e_n$. Therefore a dataset spanning many days is composed of two different stochastic processes: an individual dependent unique shape generation process and a random deviation stochastic process $e_n$. We detail next the forecasting process applied to aggregates and relate the error to $p_n$ and $e_n$.

2) Aggregate Forecasting Error Decomposition: An aggregation of $N$ consumers is given by $x_N = \sum_{n=1}^{N} x_n$. Assume a forecaster that produces the corresponding forecast $\hat{x}_N$. The forecaster is a function of the previous data available at a particular time. The forecast residual is $r_N = x_N - \hat{x}_N$. The following assumptions hold:

A1 The consumption signal minimum support grows linearly with the number of customers: $\sum_{n \in A} \frac{T}{\tau} \sum_{t} x_n(t) = N \mu$.

A2 The error residuals from the forecaster are uncorrelated in time, i.e. the covariance $\text{COV} (r_N(t), r_{N}(t+\tau)) = 0$ for $\tau \neq 0$.

Denote by $P_n \in R^N \times T$ the matrix of the profile patterns $p_n$ for all $N$ customers. The mean squared error (MSE) for the forecaster can be decomposed by conditioning on the (random) profile matrix using the tower property:

$$ \mathbb{E}[\text{MSE}(x_N, \hat{x}_N)] = \mathbb{E}[\mathbb{E}[\text{MSE}(x_N, \hat{x}_N) | P_N]]. $$
The conditional MSE on profiles (\(E[MSE(x_N, \hat{x}_N)|P_N]\)) can be further decomposed according to

\[
E[MSE(x_N, \hat{x}_N)|P_N] = \frac{1}{T} \sum_{t=1}^{T} E \left[ \frac{e_N^2(t)}{P_N} \right]
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \frac{1}{P_N} \left( \frac{1}{N} \sum_{n=1}^{N} e_n(t) - \frac{\hat{x}_N(t)}{N} \right)^2
\]

\[
= N^2 \left( \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{N} \sum_{n=1}^{N} e_n(t) \right)^2 \right)
\]

\[
+ \frac{1}{T} \sum_{t=1}^{T} \text{VAR} \left( \sum_{n=1}^{N} e_n(t) \right)
\]

where \(\text{VAR}\) denotes the variance function. The assumptions on the forecaster and the random error process \(r_n\), lead to estimates of the scaling of these quantities as explained in the next subsections.

### C. Additive Error and Properties of \(e_n\)

We consider different correlation structures on the random error \(e_n\) that are commonly considered in practice. The main observation is that the various structures generate a scaling of the error variance of the following type

\[
\text{VAR} \left( \sum_{n=1}^{N} e_n(t) \right) = \kappa N^2 + \sigma'^2 N,
\]

where the parameters \(\kappa\) and \(\sigma'\) depend on the specific error structure. We identify these parameters for two cases. Consider a finite correlation structure where \(e_n\) correlates to a constant fixed number \(K\) of other individual errors. In this case, \(\text{VAR} \left( \sum_{n=1}^{N} e_n(t) \right) = (\sigma^2 + K \rho)N = N\sigma'^2\). In particular this captures the case where errors are uncorrelated \((K = 0)\). Assume there is a correlation structure where any two individuals can have a positive covariance with some probability. That is \(\text{COV}(e_m(t), e_n(t)) = \rho \sigma^2\) with probability \(\gamma\) for any two individuals. Consider the indicator random variable \(I_{m,n}\) which is equal to 1 with probability \(\gamma\) or zero otherwise. Using the law of total variance:

\[
\text{VAR} \left( \sum_{n=1}^{N} e_n(t) \right) = E \left[ \text{VAR} \left( \sum_{n=1}^{N} e_n(t) | I \right) \right]
\]

\[
= E \left[ N\sigma^2 + \rho \sigma^2 \gamma \{ \text{m,n pairs equal to one} \} \right]
\]

\[
= \left( \gamma \rho \sigma^2 \right) N^2 + N \left( \sigma^2 - \gamma \rho \sigma^2 \right)
\]

\[
= \kappa N^2 + \sigma'^2 N.
\]

### D. Population Bias and Forecasters

The quantity

\[
\delta^2(\hat{x}_N/N, P_N)^2 = \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{N} \sum_{n=1}^{N} p_n(t) - \frac{\hat{x}_N(t)}{N} \right)^2
\]

represents how well any function can match the mean profile of the population. Notice that this depends on the set of profiles \(P_N\) and the estimating function. If we an unbiased estimate of the population generating process. That is \(\hat{x}_N = \bar{p}\), the mean profile of all customers. Under such estimator:

\[
E[N^2 \delta(\bar{p}, P)^2] = N^2 E \left[ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{N} \sum_{n=1}^{N} p_n(t) - \frac{\bar{p}(t)}{N} \right)^2 \right]
\]

\[
= \frac{N^2 \text{VAR}(P)}{N}
\]

\[
= N \text{VAR}(P).
\]

If on the other hand, a forecaster has non-zero expected population bias as:

\[
E[N^2 \delta(\bar{x}, P)^2] = N^2 \delta^2(\bar{x})
\]

### E. CV Scaling Law

We can combine the two terms which results in the following:

\[
E[MSE(x_N, \hat{x}_N)] = E \left[ E[MSE(x_N, \hat{x}_N)|P_N] \right]
\]

\[
= (\delta^2(\bar{x}) + \kappa) N^2 + \sigma'^2 N
\]

The coefficient of variation (CV) can be now bounded as a function of MSE. First using Jensen’s inequality:

\[
E[CV(x_N, \hat{x}_N)] \leq \sqrt{E[CV(x_N, \hat{x}_N)^2]}
\]

Finally combining the results:

\[
E[CV(x_N, \hat{x}_N)] \leq (E[CV^2(x_N, \hat{x}_N)])^{1/2}
\]

\[
= \left( E \left[ \frac{MSE(x_N, \hat{x}_N)}{\left( \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} x_n(t) \right)^2} \right] \right)^{1/2}
\]

\[
\leq \left( E \left[ \frac{MSE(x_N, \hat{x}_N)}{\mu^2 N^2} \right] \right)^{1/2}
\]

\[
= \left( \frac{N^2 \delta^2 + N \sigma'^2}{\mu^2 N^2} \right)^{1/2}
\]

\[
= \left( \frac{\delta^2}{\mu^2} + \frac{\sigma'^2}{N \mu^2} \right)^{1/2}
\]

The inequality in e.q. (4) comes from the minimum support condition for the group mean. Therefore we have the upper bounding envelope function:

\[
E[CV(x_N, \hat{x}_N)] \leq \sqrt{\frac{\delta^2}{\mu^2} + \frac{\sigma'^2}{N \mu^2}}
\]

Using assumption A1, the relationship can be put as a function of mean load yielding:

\[
E[CV(x_N, \hat{x}_N)] \leq \sqrt{\frac{\delta_0}{\alpha_0} + \frac{\alpha_1}{\bar{W}}}
\]
The result can also be understood according to the following intuitive analysis. Coefficient of variation is approximately \( \sqrt{\text{residual variance}/\text{signal mean}} \). If the residual mean grows as \( \kappa N^2 + \sigma^2 N \), the quadratic component of the growing MSE will counterbalance the growing signal mean. In particular, the gain in aggregation is due to reduction of the effect due to the variance of \( e_n \), but its limited by bias and variance terms that grow quadratically with aggregation numbers.

REFERENCES

