Predictability, Constancy and Contingency in Electric Load Profiles

Chenye Wu[†] Wenyuan Tang[†] Kameshwar Poolla[†] Ram Rajagopal[‡]

Abstract—Electric load profile forecasting is among the most important tasks in power system operation. This task has been performed reasonably well with minimal disruption from renewables over the past decades. The expected high penetration of renewables in the system is challenging this classical task. Moreover, the blossom of demand side management programs warrants the necessity of load profile forecasting even for individual (residential as well as industrial) consumers. In this paper, we seek to define predictability in a rigorous way from an information theoretic perspective, which shed light on comparing the predictability between different kinds of loads. Then, we identify two separable components of predictability - constancy and contingency. We further exploit the key parameters in deciding the predictability, which challenge the commonly-held beliefs. For example, does a finer granularity imply a lower predictability? To what degree does the law of large numbers help improve the predictability?

I. INTRODUCTION

Electricity sector has been working admirably over the past century, with minimal disruption from renewables. Researchers and power system engineers used to forecast the loads in aggregate (e.g., at the feeder or substation level) very accurately, with an average prediction error less than 3% [1]. This, unfortunately, is now challenged by the expected high penetration of wind farms and photovoltaic stations in the coming years. The renewables are bringing huge uncertainties to the system and are making load forecasting - the classical task - very delicate. On the other hand, the deployment of various demand side management programs warrants conducting the load forecasting not only at the substation level, but also for each individual consumer, preferably with a fine granularity. This leads to the blooming of short-term forecasting methods for individual consumers over the past few years.

A. Opportunities and Challenges in Load Forecasting

With the help of more open access data sets (Pecan Street data set [2], the open energy information initiatives [3], etc.), we seek to answer a more theoretical question: how should we evaluate the predictability of individual load profiles? Intuitively, one may propose to use the forecasting accuracy to measure the predictability, but this metric will be highly

dependent on the selected forecasting methods. More importantly, this metric may not fully capture the information in the data.

In this paper, we try to make the initial attempt toward defining predictability from an information theoretic perspective. A natural solution is to use the concept of entropy to measure how much information is included in each individual load profile. More information seems to imply more uncertainty, which may make the load profile more difficult to predict. However, we have to be careful when applying results from information theory since unlike the standard stationary stochastic process, each load profile is by nature periodic. In this paper, we propose a neat yet insightful predictability metric to capture most characteristics of load profiles. Based on the proposed metric, we make several interesting observations from the data. Many of our findings challenge the commonly-held beliefs: does a finer granularity imply the load profile will be necessarily harder to forecast? Does the law of large numbers always help improve the predictability?

We would like to emphasize that a rigorous and suitable definition of predictability in the electricity sector is not only of theoretical interests, but also may make many great opportunities happen:

- A suitable predictability definition will help better understand upper bounds of forecasting methods' performance.
- Better predictability metric may also provide better classifications for loads. Is the current classification (residential loads, commercial loads, industrial loads, agricultural loads, etc.) good enough to manage the system?
- The proposed predictability metric will provide the system operators suggestions on how to design specific products for each kind of loads. This will better enable the demand side management programs.

B. Related Work

Electric load forecasting has been playing a vital role in the electric power industry. In the past decades, a variety of forecasting techniques, for both short-term and long-term horizons, have been proposed in the literature [4], [5]. While load forecasting traditionally refers to point forecasting, probabilistic forecasting has become increasingly important with the emerging requirements such as renewable energy integration. In [6], Hong *et al.* offer a tutorial review of probabilistic electric load forecasting.

A closely related problem is electricity price forecasting, which can be based on load forecasting. This task is even
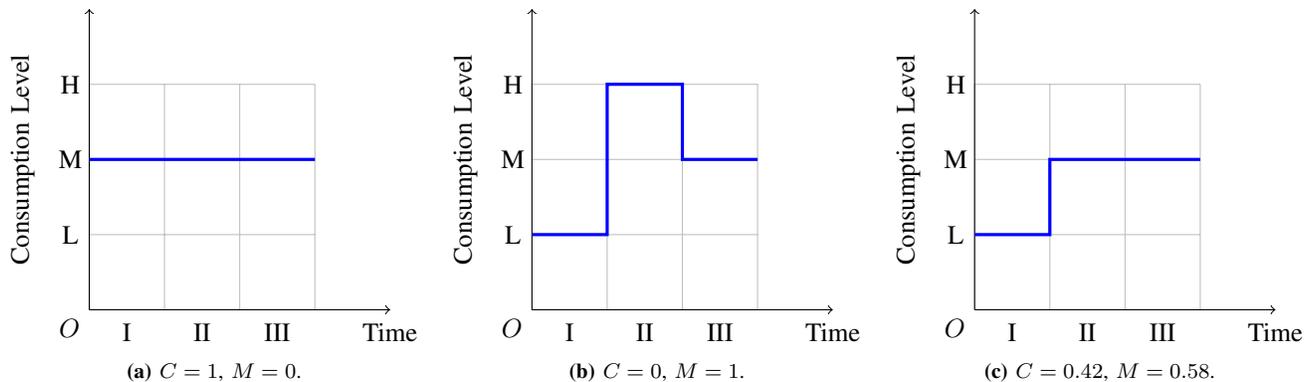
Fig. 1: Load profiles with maximal predictability. x-axis represents time periods (I: 0:00 - 8:00, II: 8:00 - 16:00, III: 16:00 - 24:00). y-axis shows the electricity consumption levels (H: high, M: medium, L: low). The load profiles repeat for 30 days. Thus, predictability (P) is 1 for each case but with different constancy (C) and contingency (M). These metrics are defined in Section II.

more challenging, since prices are much more volatile than loads. The majority of the methods proposed in the literature focus on point forecasting of day-ahead prices. See [7] for a comprehensive review. Compared with day-ahead prices, real-time (e.g. 5-minute ahead) prices are more difficult to forecast. In [8], Ji *et al.* propose a methodology for probabilistic forecast, based on a multiparametric programming formulation that partitions the uncertainty parameter space into critical regions. In this way, the authors estimate the conditional probability of the real-time locational marginal price.

The various forecasting methods for electric load, mainly based on statistical and machine learning approaches, can also be applied to the forecasting of other utilities [9]. Understanding predictability, for both periodic and general aperiodic phenomena, is an evolving topic in many other fields such as meteorology [10] and finance [11].

To quantify the effect of aggregation on load forecasting, an empirical scaling law is proposed in [12], which describes load forecasting accuracy at different levels of aggregation. Our proposed predictability metric enjoys advantages that previous works lack: our information theoretic definition for predictability does not rely on the forecasting methods.

C. Our Research Contributions

In this paper, we study the predictability of electric load profiles as well as the key parameters of predictability. The principal contributions of this paper are:

- *Predictability Decomposition*: We propose an information theoretic definition for predictability. In Section II, we explain how to decompose predictability into two separable components - constancy and contingency. This decomposition will help us better understand many counter-intuitive observations in Section IV.
- *Key Parameter Characterization*: Data helps us exploit the key parameters in deciding predictability from two aspects - granularity and data aggregation. 1) Predictability may not necessarily decrease with a finer granularity. 2) The aggregation of several hundreds of loads may not improve the predictability.

- *Predictability Comparison*: By comparing the predictability between different kinds of loads, we conclude albeit more predictable, the commercial buildings' predictability is contributed by different levels of constancy and contingency.

The remainder of this paper is organized as follows. In Section II, we describe our mathematical formulation for predictability, constancy, and contingency. Then, we treat the binning problem to select the appropriate number of electricity consumption levels to discretize the load profile in Section III. We investigate the key parameters of predictability in Section IV. Section V compares the load profile predictability between residential loads and commercial buildings. Concluding remarks and future directions are given in Section VI.

II. MATHEMATICAL FORMULATION

We will first exemplify the guidelines of our formulation, and then formally define predictability and its two components.

A. Intuitive Examples

Suppose we are interested in the daily load profile of some individual. If it has a very regular life pattern, then its load profile should be maximally predictable. This is because we can expect its load profile is repeated from day to day. What kind of load profile is minimally predictable? One cannot obtain any useful information if its consumption at each time slot independently follows the identical uniform distribution.

We propose that predictability could be decomposed into two separable components - constancy and contingency. Constancy measures if the load profile is flat while contingency measures if the load profile changes rapidly in a deterministic way (in other words, highly contingent on time). Thus, maximum predictability can be obtained as a result of either complete constancy (as in Fig. 1(a)), complete contingency (as in Fig. 1(b)), or a combination of constancy and contingency (as in Fig. 1(c)). In the case of complete constancy, the electricity consumption is maintained at the same level throughout the day. In the case of complete contingency, the electricity consumption changes from time to time, but
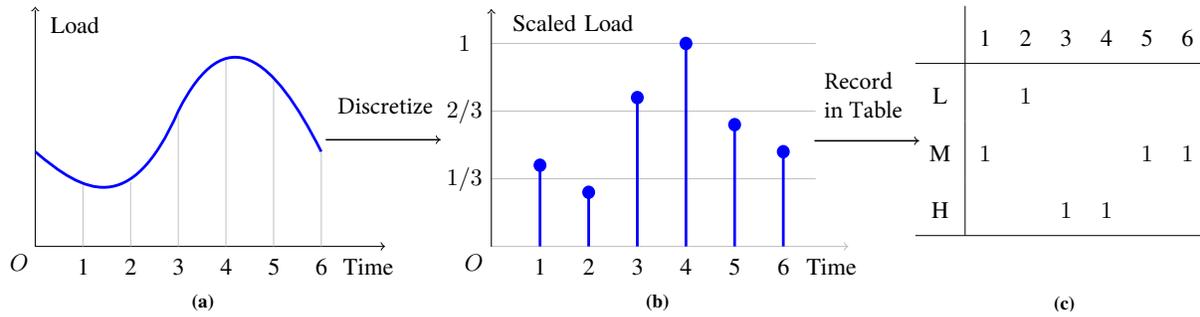
Fig. 2: Illustration of converting an individual's daily load profile into a frequency matrix.

the pattern remains the same everyday. We will use these observations as guidelines when designing the metrics.

B. Predictability Metric

Predictability and uncertainty are always coupled together. Therefore, it is natural to utilize *information theory* to design the predictability metric (as well as its two components). These metrics have been proposed in [13] by Colwell for general periodic phenomena.

Before the formal definitions, we will describe how to construct the frequency matrix for each individual load profile. For a given daily load profile, time of the day is divided into T time slots ($T = 6$ in Fig. 2). Then, we can calculate the total energy consumption during each time slot. Fig. 2(b) illustrates the discretized load profile obtained from Fig. 2(a). To better capture the common features across different days, we also normalize the discretized load profile between $[0,1]$[1]. Then, we need to further identify the number of states of interests. For illustration purposes, we bin the load profile into 3 states (low, medium, high). This yields the one-day frequency matrix in Fig. 2(c). After processing the one-day frequency matrix spanning a certain period of time, we can obtain the desired frequency matrix.

In general, consider a frequency matrix with T columns (number of time slots in a day) and S rows (number of bins to capture the electricity consumption levels). We denote N_{ij} the number of periods for which the electricity consumption was at level i at time j. Define the column sum X_j, row sum Y_i, and the grand total Z as

$$X_j = \sum_{i=1}^{S} N_{ij}, 1 \le j \le T, \tag{1}$$

$$Y_i = \sum_{j=1}^{T} N_{ij}, 1 \le i \le S. \tag{2}$$

$$Z = \sum_{i=1}^{S} \sum_{j=1}^{T} N_{ij}. \tag{3}$$

Then, using standard entropy definition, we can define the uncertainty with respect to time:

$$H(X) = -\sum_{j=1}^{T} \frac{X_j}{Z} \log \frac{X_j}{Z}; \tag{4}$$

the uncertainty with respect to electricity consumption level:

$$H(Y) = -\sum_{i=1}^{S} \frac{Y_i}{Z} \log \frac{Y_i}{Z}. \tag{5}$$

[1]We normalize the discretized load profile by its maximum of the day.

Moreover, the uncertainty with respect to both time and consumption level can be measured by:

$$H(XY) = -\sum_{i=1}^{S} \sum_{j=1}^{T} \frac{N_{ij}}{Z} \log \frac{N_{ij}}{Z}. \tag{6}$$

Based on these quantities, we propose to define the predictability (P) as follows:

$$P = 1 - \frac{H(XY) - H(X)}{\log S}. \tag{7}$$

Intuitively, $H(XY) - H(X)$ demonstrates for a given time period, the conditional uncertainty with regard to the level of electricity consumption. Its maximum ($\log S$) is achieved when all its consumption levels throughout the period are uniformly distributed for all time slots. Thus, the second term in (7) is the normalized uncertainty. We argue that predictability is the converse of uncertainty, which yields the predictability definition in (7).

Contingency represents the degree to which the electricity consumption level is contingent on time. In information theory, contingency is measured by mutual information $I(XY)$ [14]:

$$I(XY) = H(X) + H(Y) - H(XY). \tag{8}$$

Normalizing this quantity gives us one definition of contingency metric (M):

$$M = \frac{H(X) + H(Y) - H(XY)}{\log S}. \tag{9}$$

By definition, constancy and contingency are two separable components of predictability. Thus, the constancy metric C is:

$$C = P - M = 1 - \frac{H(Y)}{\log S}. \tag{10}$$

Intuitively, $H(Y)$ measures the uncertainty with respect to the consumption levels. Its opposite exactly corresponds to constancy C.

Fig. 3 gives the P, C, M values to the corresponding sample frequency matrix. We want to close this section with the following observations. Constancy is minimal when all electricity consumption levels occur equally - that is, the row sums of the matrix are all the same (e.g., matrices (b), (d) in Fig. 3). On the other hand, contingency is minimal when the electricity consumption distribution is independent of time of the day - that is, when the columns of the matrix are all the same (e.g., matrices (a), (d) in Fig. 3).

Sample a

	I	II	III
L	5	5	5
M	10	10	10
H	15	15	15

P=0.08, C=0.08, M=0

Sample b

	I	II	III
L	20	0	10
M	10	0	20
H	0	30	0

P=0.61, C=0, M=0.61

Sample c

	I	II	III
L	0	3	12
M	5	7	10
H	25	20	8

P=0.28, C=0.13, M=0.15

Sample d

	I	II	III
L	10	10	10
M	10	10	10
H	10	10	10

P=0, C=0, M=0

Fig. 3: Sample frequency matrices using 30-day data.
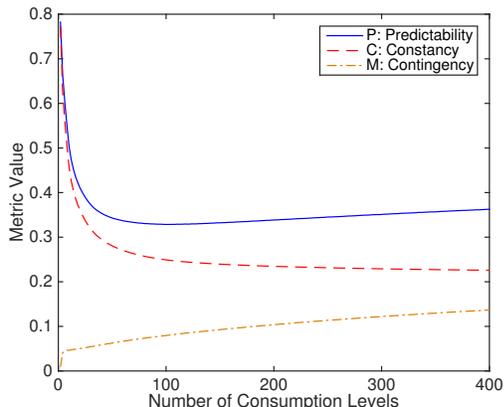
Fig. 4: Binning efforts on evaluating the metrics.

Fig. 5: Minimal P value distribution.

Fig. 6: C and M distribution at minimal P.

III. Binning the Electricity Consumption Levels

We have avoided discussing the binning problem in the previous section. Intuitively, this is a trivial task since one would imagine that the more consumption levels we have, the more meaningful or reliable predictability metrics we will attain. Bearing this intuition in mind, we will try to decide on how many electricity consumption levels that we need to calculate the meaningful predictability metrics.

In this section, we analyze the Pecan Street data set [2]. We select 378 users' load profile data with 1-minute resolution for 120 days. Fig. 4 shows how the three metrics evolve with more consumption levels and casts doubt on our intuition. If our intuition were right, then the metrics should approach their limits with more levels, which has not been observed from the figure. One clue as to how this happened lies in the fact that too many electricity consumption levels may lead to 'over-fitting'. That is, due to the limited size of data (120-day data for each individual), too many consumption levels will make most consumption levels seem to be contingent on the time slots. This is why the contingency metric (M) continues growing with the increase of number of levels.

Therefore, the appropriate binning number should well balance two aspects - capturing more information and avoiding over-fitting. We prop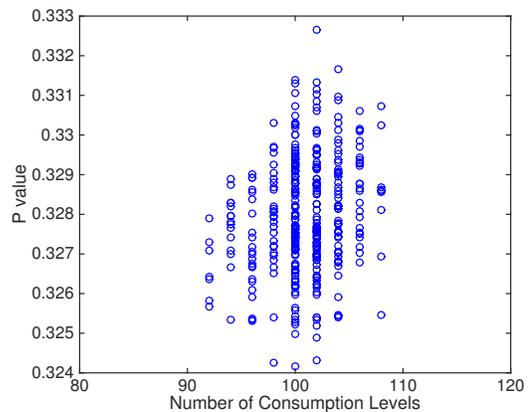ose using the number of levels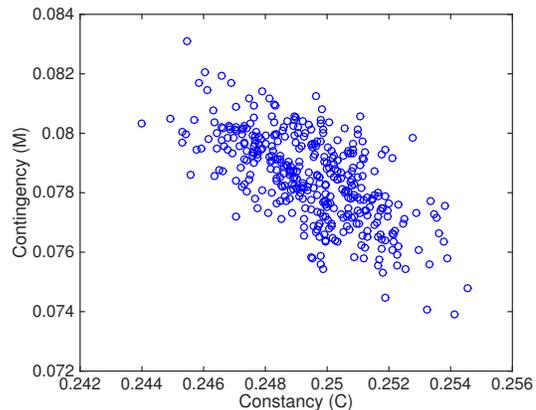 which minimizes the individual P value to conduct the trade-off. Fig. 5 shows the value of the critical point and its corresponding number of consumption level. It is obvious that for all the 378 users in the data set, the critical value resides in a narrow range of [0.324,0.333], while the corresponding number of consumption level is in the range of [92,108]. Thus, we use 100 electricity consumption levels for the subsequent analysis.

It is worth noting that Fig. 6 further illustrates that for all the 378 residential loads, all of their three predictability related metrics form a small cluster. It is the contrast with the distribution of commercial buildings that makes this observation extremely interesting, as will be discussed in Section V.

Remark: In this section, we try to trade-off between capturing more information and avoiding over-fitting. In practice, there could be more aspects to be considered. For example, fewer bins may help protect the users' privacy. Then, what is the relationship between the number of bins and the users' privacy? While a detailed discussion is beyond the scope of this paper, we believe this is a very interesting future direction.

IV. Key Parameters to Predictability

Having decided on the appropriate consumption levels, we now seek to exploit the key parameters of predictability, especially from two aspects. Does a finer granularity imply that the load profile is more difficult to predict? To what extent, the aggregation of loads implies better predictability?

A. Key Parameter - Granularity

It is natural to imagine that a finer granularity implies that the load is much harder to predict. This is almost right - indeed, to predict the electricity consumption level every six hours is a much easier task than to predict the consumption level every hour. However, it is a bit tricky to predict load profile in a much finer scale. Fig. 7(a) shows a sample scaled individual load profile with one minute resolution. It is evident that the discrete spike events are hard to predict, but the background load has a very stable pattern, which is not hard to predict at all. The load profile with 10-minute granularity (shown in Fig. 7(b)) seems much harder to predict since it looks more like a random process. Fig. 7(c) shows the load profile with 1-hour resolution, which furthers the impact of discrete events. This observation is also verified by Fig. 8, which demonstrates that the predictability first decreases and then increases with the increase of granularity. In Fig. 8, we are mainly interested in the relationship between granularity g (unit is minute) and the critical points - the minimal predictability P. Cubic fitting yields the following relationship:

$$P(g) = -0.98 \times 10^{-8}g^3 + 2.5 \times 10^{-5}g^2 \\ -0.0021g + 0.249, \tag{11}$$

This relationship cannot bear the sole responsibility for the counter-intuitive observation. As remarkable as it was, our definition has its own limitations: in our definition, we do not fully consider the intra-temporal forecasting (for example, using the electricity consumption level at time slot $t - 1$ *today* to predict the electricity consumption level at time slot t *today*). Constancy cannot fully capture this kind of forecasting methods. Instead, we mainly focus on the inter-temporal forecasting methods which only utilizes historical data (using the electricity consumption level at time slot t *in the past* to predict the electricity consumption level at time slot t *today*). We admit that utilizing intra-temporal information is a natural and powerful way to conduct forecasting, but it is beyond the capability of our proposed metrics.

B. Key Parameter - Aggregation Level

Another important parameter of predictability could be the aggregation level. Due to the law of large numbers, more aggregated load profiles should be much easier to forecast compared with the individual load profile forecasting. This is true under the condition that all the random variables are identically and independently distributed. This condition, however, is generally not true for residential load profiles. Fig. 9 shows that the predictability rapidly drops with the aggregation of more load profiles, and then remains almost as a constant.

To better understand the reason behind this observation, we plot the electricity consumption level distribution at certain time slot (10:00 am) for different level of aggregations. In Fig. 10 (a)-(c), we randomly select three single users. Each of the figures shows the electricity consumption level distributions of one of them. In Fig. 10 (d)-(f), we randomly select three
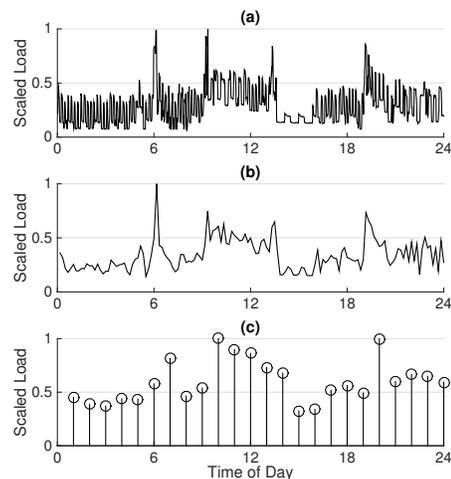
Fig. 7: Individual scaled load profile with granularity of (a) one minute, (b) 10 minutes, (c) one hour.
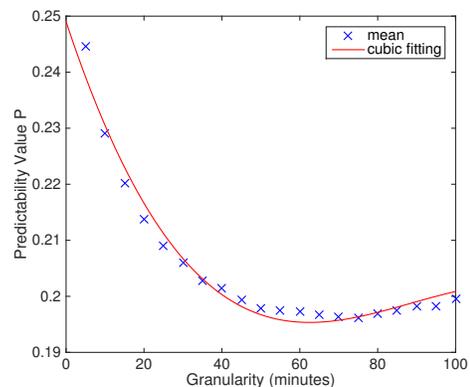
Fig. 8: Granularity v.s. predictability.

sets of 20 users. Each of the figure plots the corresponding electricity consumption level distribution. Similarly, Fig. 10 (g)-(i) show that of three sets of 200 randomly selected users, respectively. It is evident that the distribution is approaching identical Gaussian only after the aggregation of 200 users. This helps us explain our observations.

The initial drop in predictability is because the data aggregation makes the distribution less contingent on time (Fig. 9 also verifies that contingency M drops quickly with P.). After the aggregation of 200 users data, we start noticing the distribution is approaching identical distributions. Without the support of
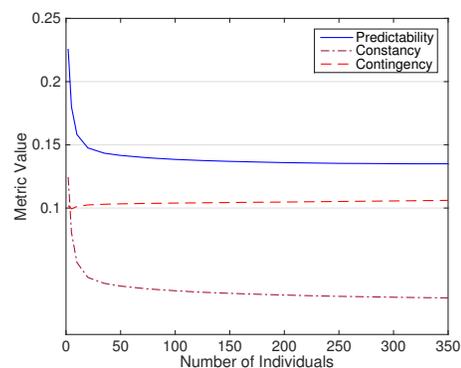
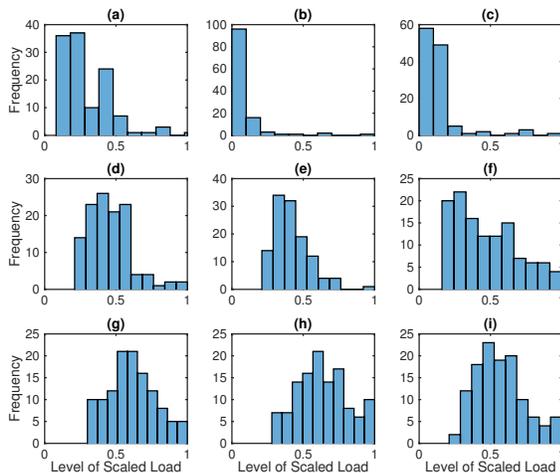**Fig. 9:** Aggregation level v.s. predictability.

Fig. 10: Electricity consumption level distributions at 10:00 am of (a)-(c) single user, (d)-(f) aggregated 20 randomly selected users, (g)-(i) aggregated 200 randomly selected users.

Fig. 11: Residential loads v.s. commercial buildings.

more data, we cannot conduct a reliable independence test. We can only conjecture that with the aggregation of more than 1,000 users, predictability is likely to start increasing again and the law of large numbers is likely to take effect.

V. RESIDENTIAL LOADS V.S. COMMERCIAL BUILDINGS

We want to demonstrate how to utilize our proposed metrics by comparing the predictability of residential loads and commercial building profiles. We use the commercial building data set provided by OpenEI [3], which contains more than 6,000 buildings' data with 1 hour resolution spanning a year. We randomly select 600 buildings' data from the data set. The red dots in Fig. 11 show the contingency and constancy distribution of the 600 selected buildings. The highly self clustered blue dots in Fig. 11 show the contingency and constancy distribution of the 378 residential loads in the Pecan Street data set. Note that the sum of contingency and constancy is predictability. We conclude that compared with residential load profiles, the load profiles of commercial buildings are often more predictable. Their predictability is contributed by various levels of constancy and contingency, almost spanning half of the figure. On the other hand, though residential load profiles are harder to predict, they have almost identical constancy and contingency levels.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate the predictability of electric load profiles. Then, we identify two key parameters of predictability - granularity and data aggregation. With the help of predictability's two separable components (constancy and contingency), we explain some counter-intuitive impacts of the two parameters on predictability.

As we mention in the paper, much remains unclear. We plan to include the privacy concerns and use a more complicated model to fully capture the intra-temporal properties in the future work. This work can also be extended in many other ways. For example, we can further strengthen our theoretical discussion by considering the unbiased entropy [15]. A
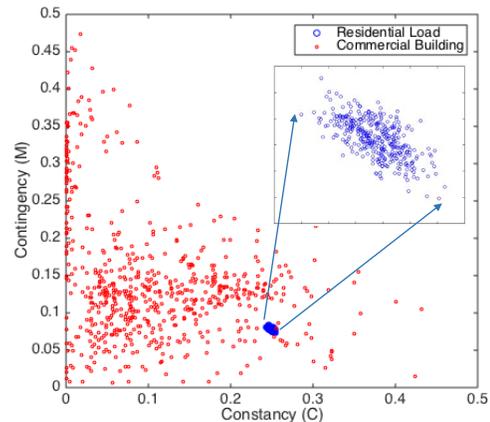
rigorous significance testing is needed to draw more decisive conclusions. Conducting the analysis on a significantly larger data set will help us verify our conjecture. Based on a well-defined predictability metric, how to systematically compare different forecasting methods? Is there a better way to classify various loads to support a better power system operation? They are all very interesting directions to go.

REFERENCES

[1] B. Hodge, D. Lew, and M. Milligan, "Short-term load forecasting error distributions and implications for renewable integration studies," in *Proceedings of IEEE green technologies conference, Denver*, 2013.

[2] PECAN STREET INC., "Pecan Street Data." [Online]. Available: http://www.pecanstreet.org

[3] US Department of Energy, "Open Energy Information." [Online]. Available: http://en.openei.org

[4] K. Lee, Y. Cha, and J. Park, "Short-term load forecasting using an artificial neural network," *Power Systems, IEEE Transactions on*, vol. 7, no. 1, pp. 124–132, 1992.

[5] M. Kandil, S. M. El-Debeiky, and N. Hasanien, "Long-term load forecasting for fast developing utility using a knowledge-based expert system," *Power Systems, IEEE Transactions on*, vol. 17, no. 2, pp. 491–496, 2002.

[6] T. Hong and S. Fan, "Probabilistic electric load forecasting: a tutorial review," *International Journal of Forecasting*, 2016.

[7] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1030–1081, 2014.

[8] Y. Ji, R. J. Thomas, and L. Tong, "Probabilistic forecast of real-time LMP via multiparametric programming," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2015, pp. 2549–2556.

[9] R. H. Brown, S. R. Vitullo, G. F. Corliss, M. Adya, P. E. Kaefer, and R. J. Povinelli, "Detrending daily natural gas consumption series to improve short-term forecasts," in *Power and Energy Society General Meeting, 2015 IEEE*. IEEE, 2015, pp. 1–5.

[10] C. F. Keller, "Climate, modeling, and predictability," *Physica D: Nonlinear Phenomena*, vol. 133, no. 1, pp. 296–308, 1999.

[11] L. Menzly, T. Santos, and P. Veronesi, "Understanding predictability," *Journal of Political Economy*, vol. 112, no. 1, pp. 1–47, 2004.

[12] R. Sevlian and R. Rajagopal, "Short term electricity load forecasting on varying levels of aggregation," *working paper*, 2014.

[13] R. K. Colwell, "Predictability, constancy, and contingency of periodic phenomena," *Ecology*, pp. 1148–1153, 1974.

[14] F. Jelinek, *Probabilistic information theory: discrete and memoryless models*. McGraw-Hill, 1968.

[15] V. Petkov, R. Rajagopal, and K. Obraczka, "Characterizing per-application network traffic using entropy," *ACM Trans. Model. Comput. Simul.*, vol. 23, no. 2, pp. 14:1–14:25, May 2013.