View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Ghent University Academic Bibliography

Exploiting Road Traffic Data for Very Short Term Load Forecasting in Smart Grids

Juan Aparicio[1,3], Justinian Rosca[1]
[1]*Siemens Corporation,*
Corporate Technology
Princeton, NJ, USA
{juan.aparicio, justinian.rosca}@siemens.com

Markus Mediger[2], Alexander Essl[2],
Klaus Arzig[2]
[2]*Siemens AG,*
Humboldtstr. 59,Nuremberg, Germany
{markus.mediger, alexander.essl,
klaus.arzig}@siemens.com

Chris Develder[3]
[3]*IBCN group, Dept. of Information Technology*
Ghent University (UGent)
Ghent, Belgium.
chris.develder@intec.ugent.be

***Abstract*— If accurate short term prediction of electricity consumption is available, the Smart Grid infrastructure can rapidly and reliably react to changing conditions. The economic importance of accurate predictions justifies research for more complex forecasting algorithms. This paper proposes road traffic data as a new input dimension that can help improve very short term load forecasting. We explore the dependencies between power demand and road traffic data and evaluate the predictive power of the added dimension compared with other common features, such as historical load and temperature profiles.**

***Index Terms*—load forecasting; power demand; regression analysis; smart grid; traffic data**

I. Introduction

Load forecasting applications are extremely important for any energy management system, to increase the energy distribution efficiency, minimize financial risks and enable demand response strategies, e.g., through advance pricing schemes. Furthermore, any error in the power demand prediction can have a significant economic impact [1]. Consequently, load forecasting is a well researched area addressed in a large number of publications every year.

Energy consumption flow is the relative change in the electrical consumption over a geographical region. For instance, during working hours, electrical consumption is mostly concentrated around industrial and office locations; whereas during other times, residential locations have higher levels of energy consumption [2]. Thus, positive energy flows take place during commuting times from industrial and office locations to residential areas. Should energy consumption flows be known a priori, the energy generation, transmission and distribution systems can shift loads proactively, without causing instability in the system, plan resources accordingly and offer accurate price predictions. For that reason, correlating the movement of vehicles (and inherently considering the mobility of people) with electrical usage seems conceptually valuable for improving power consumption predictions. This could be very useful particularly in the case of Very Short Term Load Forecasting (VSTLF), with the forecasting period ranging from minutes to several hours in advance of present time. In such cases, traffic data can be used to fine tune day-ahead load forecasts.

In this context, our paper investigates if and how readily available traffic information, supplied by an Intelligent Transportation System, can be used to enhance very short term load forecasting, by exploring the dependencies between traffic flows and power consumption in the same geographical area. We propose the simultaneous analysis of multiple sources of information and data about the smart grid network, the intelligent transportation system (e.g., geographic information systems, movement of people/vehicles), and other sources (e.g., demographics, public events and schedules, etc.) in order to correlate these multiple sources of scheduled and real-time information to predict energy flow dynamics in a smart grid.

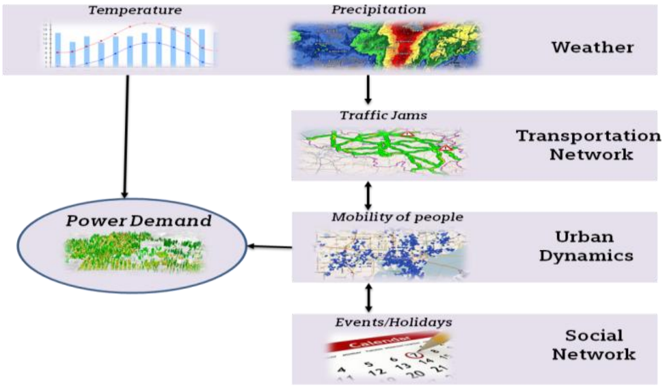

Figure 1. Causal network with factors affecting Power Demand

Our ultimate goal is to narrow down the uncertainty about the factors that drive power demand and take into account unexpected, direct or indirect, events outside the electrical network (e.g., a fast evolving storm or a traffic jam), which can significantly influence human mobility and consequently power demand. Figure 1 illustrates the targeted overall system. In particular, this paper explores the potential predictive value of transportation data.

The rest of the paper is organized as follows. Section II outlines related work in this area. Section III presents the different datasets we used. Section IV analyzes the statistical dependencies between traffic data and power demand. Section V shows the experimental results of our proposed novel predictor to enhance a sample load forecasting algorithm. Section VI ends with conclusions and future work.

II. RELATED WORK

There exist a wide variety of forecasting algorithms employing statistical techniques or artificial intelligence principles. Outstanding examples from the literature are based on iterative reweighted least-squares [3], autoregressive integrated moving average (ARIMA) [4], fuzzy logic [5], ARMAX models and genetic algorithms [6], wavelet echo state networks [7], expert systems [8] and support vector regression (SVR) [9]. Algorithms get more sophisticated (for example, Aung et al. [9] claim a 98.4% accuracy of predicting the peak load of a given day), but the number of attributes and features applied in the load prediction remain almost the same: historical load and temperature, day of the week, season, whether it is a holiday or not, any social events (strikes, sport events, etc.) and forecasted temperature.

As pointed out in [10], similar days in terms of the known variables mentioned above, may present very different patterns in the electric load. Thus, unknown factors remain that influence the power demand. It is well known in machine learning that more independent features that correlate well with the class improve the learning [11]. Hence, any previously mentioned algorithm could benefit from an additional variable that captures a new dimension correlated with power consumption, e.g., road traffic data. Due to its increasing importance in traffic management and traveler information systems [12], nowadays multiple companies offer such traffic data in real time. Thus, traffic information can be used as sensory data to measure human dynamics and extract useful information in independent dimensions. Integration into, e.g., power grid management/control systems can be seen as part of the combination of physical infrastructures and technologies in the paradigm of Smart Cities [13].

In order to reduce the "unknown factors" that influence power consumption, we propose road traffic as a new feature to improve prediction. It can take into account unexpected (non-recurrent) events, such as a fast evolving storm or an accident that might create an unexpected load profile, which could not be explained by the previously mentioned features. However, whereas the relation between weather and load is well understood (e.g., [14] [15]), to our knowledge, traffic has never been used to improve load forecasting.

III. DATA SETS

Our study exploits three different datasets from different sources to analyze the interdependencies of power, traffic and weather data in the same region and over the same period of time. In this section, we provide an overview of the different datasets that have been the basis of our studies, and formulate the mathematical representation of all our variables.

The power consumption data used in this study is part of the Flemish project LINEAR (Local Intelligent Networks and Energy Active Regions [16]). It deals with hundreds of houses distributed across the region of Flanders, Belgium. As part of the project, power consumption readings from every participating house are collected every 15 minutes. We will focus on the year 2011 for the experiments in this paper. Figure 2 gives a visual overview of the aggregated day profiles of 200 houses over the entire year, denoted by *L(t,d)*.

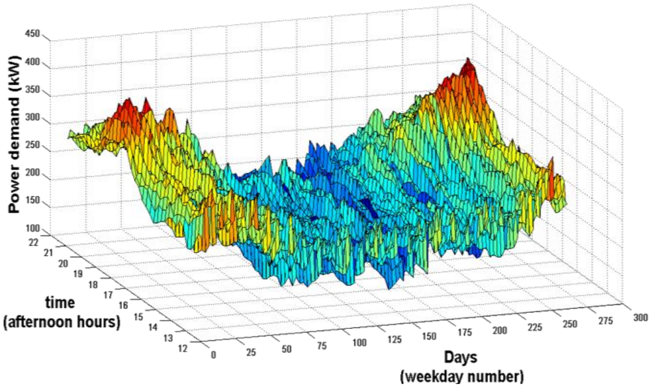

Figure 2. L (t,d): load in kW per 15min interval over 260 weekdays. Summer days occupy the central strip of the graph and they present less power consumption.

The traffic dataset was obtained from the Belgian company Be-Mobile. It was provided in the form of traffic jam length for the main roads in Belgium, with an accuracy of 50m. Calculations are based on live gps-position tracking of both professional drivers and users of personal traffic information devices, resulting in a fleet of more than 200.000 vehicles. Figure 3 gives a visual overview of the aggregated traffic jam length day profiles for the year 2011, denoted by *J(t,d)*.

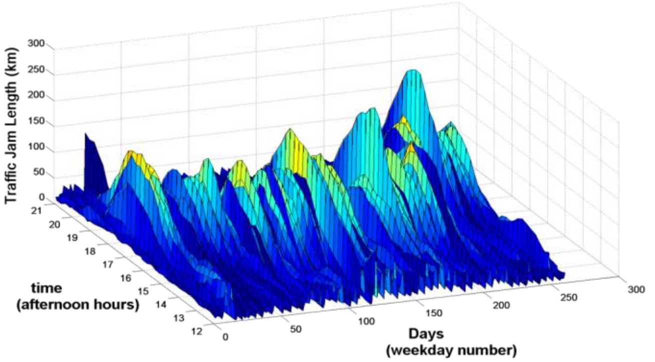

Figure 3. J(t,d): cumulative traffic jam length in kilometers over 260 weekdays, as in Figure2. No seasonal patterns are obvious.

The meteorological data used in this study was provided by the National Weather Institute of Belgium (KMI). It consists on temperature and precipitation data with a 15 minute interval for a representative city in the Flemish region. In this paper we use *T(t,d)* for temperature and *R(t,d)* for rain/precipitation data.

It is important to note that the geographical region of interest in all our datasets is Flanders. Temperatures vary from around 3°C in the winter to 17°C in the summer. Thus, heating in the winter triggers power consumption, but we don't observe a contribution of AC equipment, since this is not widespread in Flanders. In addition, the load profiles belong to residential homes. Consequently, only afternoon data is relevant to our studies (afternoon traffic influences when the users arrive home). Furthermore, we will focus our studies on weekdays where the commute work-home can be identified in the traffic profiles.

In this work we use the notation presented in TABLE I to represent the different dimensional and measured variables:

TABLE I. Variables

variable	definition	Value/unit
N	Number of samples per day	All days: 96 (one every 15min) Afternoon: 37 (from 12:00 to 21:00)
t	Sample#	12:00 < t <21:00
D	Total number of days	Total:365 days Weekdays:260
d	Day#	1 ≤ d ≤260
Δt	Shift in traffic data for accounting for delayed dependencies	0< Δt ≤8 (2hours)
δt	Time in advance targeted for prediction	15min≤δt≤2hours
L(t,d)	Power Consumption – Load profiles	unit: kW
J(t,d)	Traffic Jam Length profiles	unit: meter
T(t,d)	Temperature profiles	unit: degree Celsius
R(t,d)	Precipitation profiles	unit: centimeter

IV. Data Analysis

In this section, we analyze the causal dependence of variables of interest using correlation metrics in order to determine whether traffic could be used as a predictor or not, and under which circumstances. Our goal is to demonstrate that traffic jam length (J) has a measurable influence on power consumption, i.e., load (L). First, we will experiment with correlations at different times of the day (t) and various shifts in the traffic data (Δt). Second, the effect of various environment conditions – rain and traffic jams – will be analyzed. And finally, we will extend our calculations to the relationship between load and temperature (T); which gives us a metric to compare both influences.

Predictability is strongly connected to correlation. When two series are highly correlated, a predictive relationship can be extracted and exploited in practice. Consequently, if traffic is correlated with load, it could be used as a predictor in load forecasting and improve the accuracy of the forecast. The correlation degree between two variables can be measured by different statistical metrics: Pearson's correlation coefficient, Spearman's correlation coefficient, Mutual Information, etc. We have chosen the Pearson's correlation coefficient as it is the most common and widely used in the literature. Pearson's correlation coefficient ranges from –1.00 (negative correlation) to +1.00 (positive correlation) and is calculated by dividing the covariance of two variables by the product of their standard deviations. In our case, for a fixed $t \in [12{:}00, 21{:}00]$ with d as running variable:

$$\rho_{L,J}(t) = \frac{\mathrm{cov}(L(t,\bullet), J(t-\Delta t,\bullet))}{\sigma_L \sigma_J} = \frac{E[(L(t,\bullet)-\mu_L)(J(t-\Delta t,\bullet)-\mu_L)]}{\sigma_L \sigma_J}\ ; \qquad (1)$$

In our study, we will discretize the levels of correlation more strictly compared to [15]:

- | ρ | ≤ 0.1 represents no correlation.
- 0.1 < | ρ | ≤ 0.3 represents small correlation.
- 0.3 < | ρ | ≤ 0.6 represents medium correlation.
- 0.6 < | ρ | ≤ 1.0 represents strong correlation

Compared to temperature, the effects of traffic may be much more delayed in time, e.g., traffic at 4pm may influence load at 6pm. Figure 4 shows the correlation coefficient for load and traffic. The different curves represent different shifts in travel data, i.e. they correspond to the correlation between $L(t)$ and $J(t - \Delta t)$ for Δt = 15 min, 30 min, 1h and 2h.

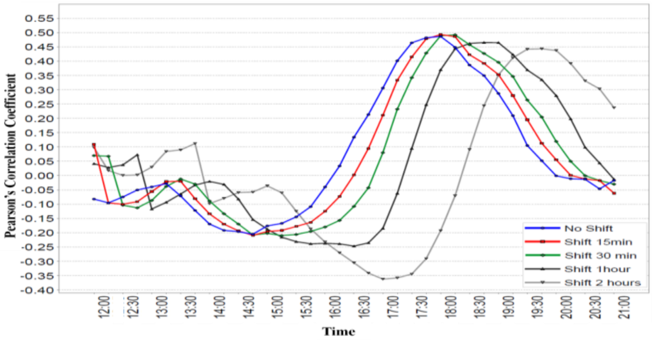

Figure 4. The correlation coefficient between traffic and load with different offsets shows medium correlation at relevant afternoon hours.

One can see that correlation varies along the day, being stronger from 17:00 until 20:00. At the peak, the correlation is around 0.5, which can be considered as medium correlation. In addition, at each time t, the highest correlation coefficient occurs for a different traffic offset (Δt); e.g., at 17:00 the correlation coefficient ρ is maximum for the curve Δt = 0, while at 19:30 the load correlates better with traffic from two hours ago (Δt = 2h).

Mobility patterns are reflected in the way curves in Figure 4 change from negative to positive correlation. From one day to another, when traffic is high during work-home commute hours, it slows down people and makes them arrive home later; therefore homes start consuming energy later. Before 16:00, if traffic is high, load is low, thus negative correlation. However, after 16:00, more traffic implies a higher consumption, hence positive correlation.

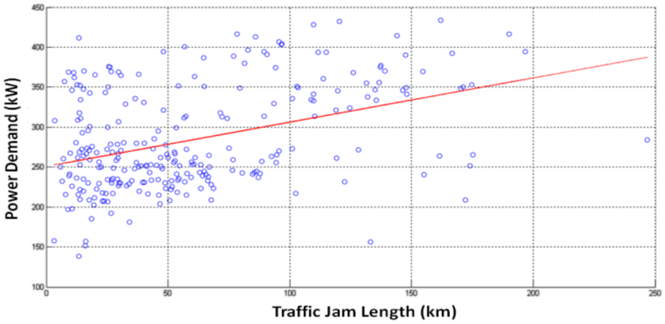

Figure 5. Scatterplot of traffic jam length vs. power demand for t = 5:15pm and Δt =15 min, where $\rho_{J,L}$=0.4. Most of the scatterplot points are located at low traffic positions.

Like many popular statistics, the Pearson's coefficient is not robust [17]. As a result, its value can be ambiguous if outliers are present [18]. An inspection of the scatter plot in Figure 5 (traffic and power for t = 17:00, where p = 0.5), will help us understand better the correlation values calculated in

Figure 4. Note that most of the points are grouped at low traffic positions, below 50km.

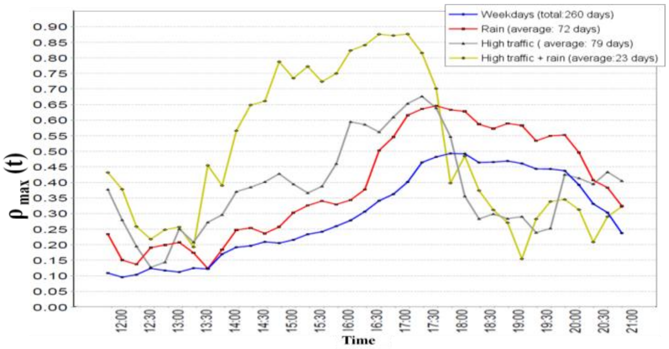

Figure 6. Absolute value of the correlation coefficient between traffic and load considering a traffic offset up to 2hours. The various lines area calculated for days with different traffic and precipitation conditions. The correlation coefficient reaches its maximum in rainy days with high traffic.

Our traffic datasets aggregate traffic information for a large region and the residential houses are distributed across the area. The effects of an accident or an abnormal traffic jam in a specific road segment may be obscured in the overall data. In spite of the larger region aggregation, rain shows an effect on the entire network, slowing down the traffic in the entire region. Days with combined heavy rain and heavy traffic are the most relevant in this analysis. Indeed, as we can see in Figure 6, correlation between traffic and load dramatically increases for the rainy days.

Figure 6 compares the correlation for different types of days. At every time *t*, it represents the maximum absolute correlation coefficient; across the time shifts of interest (very short term), as defined below.

$$\rho_{\max}(t) = \max\left\{\left|\rho_{L,J}(t-\Delta t)\right|\right\};\ for\ 0 \le \Delta t \le 2h \tag{2}$$

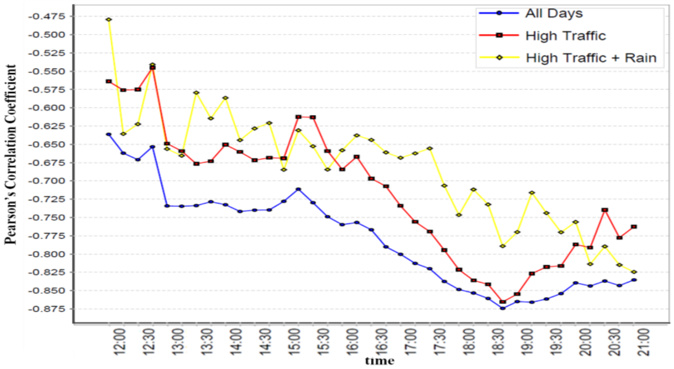

Figure 7. Correlation between temperature and load for days with different traffic and precipitation conditions. The correlation coefficient gets weaker in high traffic and rain scenarios.

At every time *t*, a day is considered to suffer heavy traffic if the aggregated traffic from *t-2h* until *t* is at least ¼ above the average.

Correlation coefficients are not absolute metrics, i.e., a correlation of 0.5 may be significant in some cases and insufficient in others. It is therefore important to compare the correlation between load and traffic with that of load and other variables. Figure 7 shows the correlation of temperature and load for all days, high traffic days and days with high traffic and rain. As we can see, the dependency gets weaker under certain environmental conditions (heavy rain + heavy traffic).

V. Load Forecasting

As we have seen in Section IV, there are several indications that support our hypothesis that traffic has a measurable influence on load, particularly in certain conditions such as rain and heavy traffic. However, our ultimate goal is to demonstrate that traffic data can be exploited for better load forecasting. In this section, we evaluate predictive contributions from the traffic variable. For illustrating this, we have designed four different pattern matching based prediction algorithms.

Our evaluation is for weekdays only (260 days), of which we have used the 80% for training (208 days) and 20% for testing (52 days). The load forecasting algorithm for time *t* gives a prediction for load at a future time *t*+ *δt*. Intuitively, it looks for patterns in the training data similar to the pattern in the present data, and then aggregates load from the similar days. Formally this is done as follows:

$$\hat{L}(t+\delta t) = L(t)\cdot\frac{\tilde{L}(t+\delta t)}{\tilde{L}(t)} \tag{3}$$

$$\tilde{L}(t) = \sum_{i=1}^{k} w_i \cdot L_i(t) \tag{4}$$

where $\hat{L}$ is the predicted load, *L* is the load at the present time, $\tilde{L}$ is the aggregated load from similar days, *k* is the number of similar days, W_i = *1/d* and *d* is the Euclidian distance between the measured present point and similar points in the feature space.

The main difference between the algorithms lies in the feature spaces used in the calculations:

1. Alg_L: Load features
2. Alg_L_T: Load and temperature features
3. Alg_L_T_L: Load, temperature and traffic features
4. Alg_L_T_L_R: Load, temperature, traffic and high traffic + rain features.

The load and traffic features are computed for the present time and a time offset equal to Δ*t=15min, 30min,1hour and 2 hours.* The high traffic + rain feature is a Boolean dimension, where the number of rainy and high traffic days (R+T days) is computed at every time *t* with a window of two hours. After the features are selected, the feature vectors are normalized using the second norm. The second norm of a vector X is calculated as:

$$\|X\| = \sqrt{\max eigenvalueof\ \ X^{H} X} \tag{5}$$

TABLE III shows the relative prediction error (average and standard deviation), of the previously described algorithms; over all the testing days. For each testing day, the relative error is calculated as follows:

$$relative\ error = \frac{\left|actualLoad - calculatedLoad\right|}{actualLoad} \qquad (6)$$

As we can see in TABLE II, using traffic as an additional feature adds value to the prediction. This is true for the times and environment situations discussed along the paper. The results show the best performance when considering rain and traffic days separately. These numbers may vary using more sophisticated algorithms, but the intrinsic benefit of using traffic as a forecasting feature would remain if carefully chosen.

TABLE II. Relative error in % (μ) and standard deviation (σ) for the load forecasting algorithms with time window for prediction *δt= 1hour* and prediction times *t + δt =15:00, 16:00, 17:00 and 18:00*. Every experiment have been run 100 times for assuring statistical significance.

Algorithm	**15:00**	**16:00**	**17:00**	**18:00**
	μ(σ)	***μ (σ)***	***μ(σ)***	***μ(σ)***
Alg_L	4.03 (3.43)	3.95 (3.48)	5.17 (3.84)	5.18 (4.14)
Alg_L_T	4.17 (3.33)	3.84 (3.55)	4.77 (3.67)	4.68 (3.74)
Alg_L_T_J	4.27 (3.38)	3.91 (3.54)	4.76 (3.5)	5.06 (4.25)
Alg_L_T_J_R	3.71 (2.44)	2.86 (2.43)	3.59 (2.96)	3.16 (2.16)

VI. Conclusions and future work

In this paper, we proposed road traffic data as additional input to improve short term load prediction. We analyzed the dependence between traffic and load and evaluated prediction algorithms exploiting the added feature on real datasets gathered within the region of Flanders, Belgium. We demonstrated the correlation between road traffic data in the form of traffic jam length, and power demand, and compared this with correlations between power demand and temperature. Results indicate a significant correlation at relevant hours for rainy and high traffic days. These findings have been applied in a novel load forecasting algorithm with positive results.

Linking mobility with electrical grid data is becoming increasingly important due to the growing green transportation initiatives around the world. These initiatives promote zero-emissions technologies, such as e-Highways or electric vehicle charging. Future work will focus on applying and extending our approach to new datasets with different characteristics, e.g., AC consumption patterns, big and small cities, etc. In addition, a natural extension of our study is to analyze a network with a noteworthy penetration of electric vehicles, where the electrical load is significant and varies depending on when and where cars recharge.

Acknowledgment

The authors would like to acknowledge contributors of data that made this study possible. The datasets were kindly provided by Be-Mobile, the National Weather Institute of Belgium (KMI) and the Flemish Government through the LINEAR [16] project (IWT/090800) organized by the Institute for Science and Technology (IWT).

References

[1] M. Ortega-Vazquez and D. Kirschen, "Economic impact assessment of load forecast errors considering the cost of interruptions," *IEEE Power Engineering Society General Meeting(2006),* p. 8 pp, 2006.

[2] A. Phdungsilp, "Energy Analysis for sustainable Mega-cities," Licentiate Thesis at School of Industrial Engineering and Management - Department of Energy Technology - Royal Institute of Technology (KTH) , Stockholm, Sweden, 2006.

[3] G. Mbamalu and M. El-Hawary, "Load forecasting via suboptimal seasonal autoregressive models and iteratively reweighted least squares estimation," *IEEE Transactions on Power Systems,* vol. 8, no. 1, pp. 343-348, 1993.

[4] X. Jin, Y. Dong, J. Wu and J. Wang, "An Improved Combined Forecasting Method for Electric Power Load Based on Autoregressive Integrated Moving Average Model," *International Conference of Information Science and Management Engineering,* vol. 2, pp. 476-480, 2010.

[5] M. Ganjavi, C. Lucas and M. Javidi, "Short term load forecasting using fuzzy neural network modified by the similarity and subsethood measures," *Journal of Intelligent & Fuzzy Systems,* vol. 7, no. 4, p. 347, 1999.

[6] B. Wang, N.-l. Tai, H.-q. Zhai, J. Ye, J.-d. Zhu and L.-b. Qi, "A new ARMAX model based on evolutionary algorithm and particle swarm optimization for short-term load forecasting," *Electric Power Systems Research,* vol. 78, no. 10, pp. 1679-1685, 2008.

[7] A. Deihimi, O. Orang and H. Showkati, "Short-term electric load and temperature forecasting using wavelet echo state networks with neural reconstruction," *Energy,* vol. 57, pp. 382-401, 2013.

[8] J. Liu, C. Niu, X. Liu, W. Tan and J. Li, "Short-term load forecasting system based on decision tree and expert system," *Journal of Computational Information Systems,* vol. 4, no. 5, pp. 2057-2062, 2008.

[9] M. T. J. W. A. S. S. H. Z. Aung, "Towards Accurate Electricity Load Forecasting in Smart Grids," *Proc. 4th International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA'12),* pp. 51-57, 2012.

[10] M. Lopez Garcia, S. Valero, C. Senabre and A. Gabaldon Marin, "Short-Term Predictability of Load Series: Characterization of Load Data Bases," *IEEE Transactions on Power Systems,* vol. 28, no. 3, pp. 2466-2474, 2013.

[11] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM,* vol. 55, no. 10, pp. 78-87, 2012.

[12] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp and H. Scholten, "Data from telecommunication networks for incident management: An exploratory review on transport safety and security," *Transport Policy,* vol. 28, no. (Special Issue on Transportation Pricing Policies), pp. 86-102, 2013.

[13] D. Bartlett, W. Harthoorn, J. Hogan, M. Kehoe and R. Schloss, "Enabling integrated city operations," *IBM Journal of Research and Development,* vol. 55, pp. 15:1,15:10, 2011.

[14] H. A. Dryar, "The Effect of Weather on the System Load," *Transactions of the American Institute of Electrical Engineers,* vol. 63, no. 12, pp. 1006-1013, 1944.

[15] L. Hernandez, C. Baladron, J. M. Aguiar, L. Calavia, B. Carro, A. Sanchez-Esguevillas, D. J. Cook, D. Chinarro and J. Gomez, "A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework," *Sensors (14248220),* vol. 12, no. 9, pp. 11571-11591, 2012.

[16] C. D. J. D. J. D. R. B. E. Peeters, "LINEAR: towards a Breakthrough of Smart Grids in Flanders," *Proc. 2nd Int. Conf. Innovation for Sustainable Production (i-SUP 2010), Bruges, Belgium,* vol. 3, pp. 3-6, 18-21 Apr. 2010.

[17] D. M. V. Marona R., Robust Statistics - Theory and Methods, Wiley, 2006.

[18] R. J. K. S. Devlin, "Robust estimation and outlier detection with correlation coefficients," *Biometrika,* vol. 62, no. 3, pp. 531-535, 1965.