Urban electric load forecasting with mobile phone location data

Selvarajoo, Stefan; Schläpfer, Markus; Tan, Rui

2019

Selvarajoo, S., Schläpfer, M., & Tan, R. (2018). Urban electric load forecasting with mobile phone location data. Proceedings of the 2018 Asian Conference on Energy, Power and Transportation Electrification (ACEPT), 18403376-. doi:10.1109/accept.2018.8610757

<https://hdl.handle.net/10356/145170>

<https://doi.org/10.1109/ACEPT.2018.8610757>

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/ACEPT.2018.8610757>

Downloaded on 26 Apr 2023 13:19:46 SGT

Urban Electric Load Forecasting with Mobile Phone Location Data

Stefan Selvarajoo
School of Computer Science and Engineering
Nanyang Technological University
Singapore
stefan003@e.ntu.edu.sg

Rui Tan
School of Computer Science and Engineering
Nanyang Technological University
Singapore
tanrui@ntu.edu.sg

Markus Schläpfer
Future Cities Laboratory, Singapore-ETH Centre
ETH Zurich
Singapore
schlaepfer@arch.ethz.ch

Abstract – In recent years, electrical load forecasting has received continuous research efforts aiming to improve the short-term prediction accuracy of local energy demands. However, current methods are not able to take explicitly into account the dynamic spatial population distribution over the course of a day, which is particularly relevant in dense urban areas. In this paper, we harness society-wide mobile phone data to map the time-varying population distribution in the Trentino region, Italy, and to use these insights for a novel electrical load forecasting method. Our results demonstrate that the integration of aggregated mobile phone data yields compelling forecast models.

Keywords – *Data analytics, Electrical load forecasting, Power system management, Spatiotemporal analytics*

I. INTRODUCTION

Forecasting electrical loads is a focal point globally to raise energy efficiency levels. Traditionally, electric load forecasting methods have been based on historical demand data, for instance covering several seasons with a half-hourly resolution. Recent approaches include fuzzy logic, vector models and other advanced methods to improve the quality of the forecasts, leaving potential adopters inundated with a plethora of different options [1].

This paper seeks to find an alternative approach, while building on the strengths of past methods, by using new types of data that allow for the emulation of the population-dependent demand – specifically mobile phone data. Mobile phone records have proven to accurately quantify the flows of people and the resulting dynamic population distribution over a region, as devices are connected constantly to mobile phone networks. Motivated by a recent study that used mobile phone data for the planning of new electricity infrastructure in developing countries [2], this paper aims to harness mobile phone data in the ‘smart city’ context, by modelling and predicting the electricity consumption in the Trentino region, Italy.

II. DATA DESCRIPTION

The data used in this paper consist of mobile phone data, electric load data and census data.

A. Mobile phone data

The mobile phone data used in this work has been made publicly available in the course of a big data analysis challenge organized by Telecom Italia. A detailed description can be found in [3], while the data can be downloaded from the Dandelion API [4]. The API offers data for up to 2 months of aggregated records from calls, Short Message Service (SMS) and Internet connections (cellular data) with a time resolution of 10 minutes. The data are aggregated into a total of 6,575 grid cells covering the Trentino region, which ensures proper data anonymization.

B. Electric load data

The electric load data were also provided by the Telecom Italia big data challenge [3] [4]. The data consist of line information and corresponding ampere values with a time resolution of 10 minutes.

C. Census data

We used census population information at the municipality level (administrative level 8) [5]. In total there are 217 municipalities within the Trentino region as of 2014. However, due to irregularities within the dataset, only 149 municipalities could be successfully mapped. Their geographic boundaries were acquired as shapefiles from Open Street Map [6].

III. DERIVATION OF THE DYNAMIC POPULATION

A. Mapping cellular data

To quantify the dynamic spatial population distribution, we adopted the following procedure. For each individual grid cell i

Fig 1: Mobile phone data of Internet activity at 4am

Fig 2: Linear Regression of Census vs Cell phone Activity

Fig 3: Histogram of r^2 values over 149 municipalities in Trentino

and each hour, we normalized the mobile phone activity (calls, SMS, Internet connection), $A_i(t)$, by the total activity over all grid cells, $a_i(t) = A_i(t) / \sum_{i=1}^n A_i(t)$, where n is the total number of grid cells and a_i the normalised activity value. Mapping data in this way on an hourly basis reflects the actual movement of people within a region. For example, considering only the morning times between 6am and 7am would represent commuters travelling to the city centre where a convergence of people is seen in our data. Similarly, 11am to 12pm showed movements during lunch time, while 3pm to 5pm showed a gradual divergence of people leaving their workplaces. In the evenings, 8pm to 10pm the data showed people engaging in evening events. Importantly, the time between 3am and 4am corresponds to the highest probability that people stayed at home, which mirrors the census-based population distribution, see Fig. 1.

B. Distilling a dynamic population equation

To establish the functional relation between the mobile phone activity and the actual number of people present, we first established a ‘baseline proxy’. More precisely, we used the cellular data as a proxy for the presence of people because this type of mobile phone activity is less sensitive to people being asleep during night-hours (compared to calls and SMS). We directly compared the cellular data at 4am with the census population as ground truth, assuming that the majority of people are at home resting during this hour. Thereby, to derive a dynamic population equation, we followed the approach proposed in [7] [8] and performed a linear regression on the log-transformed data as shown in Fig. 2 (hourly data for 25 November 2013, 3am-4am was used). The resulting equation is

$$P_{D,i} = 0.55A_i^{0.85}, \quad (1)$$

where $P_{D,i}$ represents the Dynamic Population and A_i represents the cellular data activity in grid cell i . Equation (1) corroborates the results presented in [8] [9] in that the relationship between census data and mobile phone activity has a super-linear effect.

IV. POWER FLOWS VERSUS DYNAMIC POPULATION

Fig. 3 shows the histogram values of r^2 values from the regression analysis of power flow values over dynamic population in the Trentino region. The number of municipalities where the r^2 values are greater than 0.5 is 70. To exemplify the effects of this region’s behaviour, the following sections will focus on the power flows of two municipalities – Trento and Avio. This was done to show that dynamic population would work efficiently in areas with a high correlation, but also to point out the reasons why dynamic population would fail in some cases.

A. Good-fit correlation model

Trento, being the dense city centre (Fig. 5), provides a good fit model for associating dynamic population and power flows. Each data point in Fig. 4 represents an hourly period. Notably, its r^2 value is high at 0.77. It is expected that the city centre is a reliable model upon which power flows are correlated using the derived baseline model. Since large swathes of people, commuters, would travel to the city centre for business, the convergence of the mobility pattern was clearly enforced. Additionally, Trento is not only a proverbial business district, but also a residential and industrial area which further explains why the relationship between the power flow values and the dynamic population is high.

Fig 4: Power Flow vs Dynamic Population in Trento

Fig 6: Power Flow vs Dynamic Population in Avio

Fig 5: Geographic location of Trento

Fig 7: Geographic location of Avio

B. Poor-fit correlation model

Conversely, Avio (Fig. 7), is located on the southern tip of Trentino. Fig. 6 displays a negative correlation between power flows against dynamic population with a low r^2 value. This could be due to the specific geographic area – many vehicles are passing by the municipality on the highway to neighbouring regions and therefore the cell phone data activity could factor in these short spurs of transient mobility patterns. Therefore, it cannot contribute to a model to base the power flows on.

Comparing Fig. 4 with Fig. 6, it is clear that dynamic population has to be tailored specifically for regions with a disproportionately high population as in the case of Trento.

V. POWER FLOW FORECASTING

As the municipality of Trento shows a high correlation between power flows and dynamic population, the forecast

models will be based on this municipality accounting for up to one week of historical power flow values and dynamic population.

A. Overview of power flows and dynamic population

As depicted in Fig. 8, there is a noticeable lag time between power flows and dynamic population. Together with the rising peaks of both dynamic population and power flows, it is evident that there is indeed a relationship where dynamic population was able to compliment the forecasting values [10]. This was done by adding dynamic population as an external variable, or regressor, into the following ARIMA^{1,2} models.

B. Historical power-driven forecasts

Fig. 9 displays the 24-hour historical power flow model of the Trento municipality. This ARIMA model represents the non-seasonal part of the model. The seasonal part is not detected as

¹ ARIMA values were obtained using the auto.arima function in forecast package of R, and later verified using the ACF and PACF.

² Inherent design of the forecast package in R causes a gap between last historical value and first predicted value.

Trento 6-day observation from 23 November to 29 November

Fig 8: Trento 6-day observation of historical power values and dynamic population from 23 November to 29 November

Forecast with ARIMA(2,1,0) without dynamic population

Fig 9: Forecast of Power Flow without Dynamic Population on 23 November 2013 with a 5-hour forecast

Forecast from ARIMA(0,2,0) 10am to 1pm(10 minute interval) without dynamic population

Fig 10: Power Flow historical values on the 23 November 2013 (10-minute interval) from 7am to 10am with 2-hour forecast

this would require the availability of a longer period of historical data.

Additionally, short-term forecasting was performed which is particularly relevant if the available data covers only a few hours. Fig. 10 shows the past 3-hour data on the 23 November 2013 from 7am to 10am indicated from observations 1 to 19. These measurements are gathered in their raw form of 10-minute intervals, without aggregating them hourly. Observations from data point 20 onward represent forecast values from 10am to 12pm. This was done to visualise how feeding almost immediate data using past 3 hours affects the forecast model.

C. Forecast combining historical power and dynamic population variables

Fig. 11 shows how adding the dynamic population as an external regressor improves our forecast model. Similar to Fig. 9, the historical values were aggregated hourly on the 23 November 2013. Note the reduction of the confidence intervals.

Similar to Fig. 10, Fig.12 shows the past 3-hour historical data from 7am to 10am at 10-minute intervals with the dynamic population incorporated. It is evident that the extended forecast model predicts closer to the actual power values.

Forecast with ARIMA(2,0,0) with dynamic population

Fig 11: Forecast of Power Flow and Dynamic Population on 23 November 2013 with a 5-hour forecast

Forecast from ARIMA(2,0,0) 10am to 1pm(10 minute interval) with dynamic population

Fig 12: Forecast of Power Flow and Dynamic Population on 23 November 2013 (10-minute interval) from 7am to 10am with a 2-hour forecast

Error Measures ~ Forecast Models	Mean Error	Root Mean Square Error	Mean Average Error
Fig.9	36.63	4363.30	3349.82
Fig.10	-29.44	202.87	155.55
Fig.11	23.44	3159.86	2673.41
Fig.12	7.82	180.02	155.04

Fig 13: Table measuring the forecast accuracy values

D. Interpreting the accuracy of forecast models

The Mean Average Error (MAE) was suggested as the recommend accuracy measure [11]. Fig. 13 shows that the forecast models with dynamic population do reduce the MAE across Fig. 9 through Fig. 12. The Root Mean Square Error and Mean Average Errors have a large scale of difference due to the greater amount of training data used for hourly forecasting compared to only 3-hours of past values.

The Mean Error for Fig. 10 is negative indicating that the forecast model does not fit well to the training data. Additionally, the forecast model may not have enough data or may have used a too high polynomial approximation. Given that time series data can be noisy, the underlying model may not account for some variables and negative error has to account for the variance in the signal.

Fig. 10 did not include dynamic population while Fig. 12 produced a lower MAE which suggests that adding dynamic population as an external regressor improves forecast accuracy.

Since the correlation, as shown in Fig. 4, is high (0.77), it was likely that there was dependence between the variables. In other words, there may be hidden traits within historical power values such as weather or seasonality factors which affect any external variable, like dynamic population, to cause a drastic change within the forecast model. This could also mean that there is multicollinearity [12] between any predictor variables. To detect against and prevent this, applying Variance Inflation Factors (VIF) would be crucial. However, this paper does not

delve into the specifics of VIF, but rather its main intention is to integrate the use of cell phone data records into historical power values as a means to improve forecast accuracy. The behavior of the MAE supports the idea that adding dynamic population does improve the forecast accuracy overall.

VI. CONCLUSIONS

A. Limitations of data set

The electric load data are subject to the following limitation. A given power-grid line typically intersects over different grid cells while providing the Ampere values flowing through the line at each interval. Therefore, when aggregating power values running through the line, values may be duplicated across different grid cells and even propagated to other municipalities. However, the understanding is that since this duplication was spread, the error is assumed to be within the known limits of this line and electricity data set.

B. Limitations of the approach

As mentioned above, a highway close to a municipality could distort the estimation of the dynamic population and thus limit to application of our approach to such locations. Moreover, regions with large industrial activity are expected to show lower correlations between the dynamic population and the electricity demand, so that the consideration of mobile phone data might not lead to significant improvements of the forecast. It thus will be important to more systematically explore the applicability of our approach to regions with various land uses.

C. Secondary factors influencing mobility patterns

As with all forecasting models, weather patterns [13] greatly impact the variance of the model. Other factors such as the recency, holiday and weekend effect will also shape the result. These secondary factors will be potentially used as regressors for future models.

D. Closing thoughts

We have shown that the dynamic population approach provides valuable insights into the time-varying energy consumption of city centers with large human activity. For those locations – and if granular mobile phone data sets are provided that cover longer periods of time – the dynamic population can be used to improve the accuracy of electric load forecasting.

ACKNOWLEDGMENT

M.S. acknowledges the Future Cities Laboratory at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074016) under its Campus for Research Excellence and Technological Enterprise Programme.

REFERENCES

- [1] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, no. 32, pp. 914-938, 2016.
- [2] E. A. Martinez-Cesena, P. Mancarella, M. Ndiaye and M. Schlöpfer, "Using mobile phone data for electricity infrastructure planning," *arXiv preprint arXiv:1504.03899*, 2015.
- [3] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrioni, F. Antonelli, A. Vespignani, A. Pentland and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trento," *Scientific Data*, p. 15, 2015.
- [4] "Dandelion API," SpazioDati, 01 11 2013. [Online]. Available: <https://dandelion.eu/datamine/open-big-data/>.
- [5] "Italian National Institute of Statistics," [Online]. Available: <http://www.istat.it/en/>. [Accessed 3 February 2018].
- [6] M. Raifer, "Overpass Turbo," Overpass API, 3 February 2018. [Online]. Available: <https://overpass-turbo.eu/>.
- [7] A. Bogomolov, B. Lepri, R. Larcher, F. Antonelli, F. Pianesi and A. Pentland, "Energy consumption prediction using people dynamics derived from cellular network data," *EPJ Data Science*, p. 15, 2016.
- [8] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel and A. J. Tatem, "Dynamic population mapping using mobile phone data," *PNAS*, vol. 111, no. 45, November 11, 2014.
- [9] M. Schlöpfer, L. M. A. Bettencourt, S. Grauwin, M. Raschke, R. Claxton, Z. Smoreda, G. B. West and C. Ratti, "The scaling of human interactions with city size," *Journal of the Royal Society Interface*, vol. 11, no. 98, p. 20130789, 2014.
- [10] R. Nau, "Stationarity and differencing of time series data," Duke University, [Online]. Available: <https://people.duke.edu/~rnau/411diff.htm>. [Accessed 23 May 2018].
- [11] R. J. Hyndman, "Measuring forecast accuracy," 31 March 2014. [Online]. Available: <https://robjhyndman.com/papers/forecast-accuracy.pdf>. [Accessed 24 March 2018].
- [12] G. Kemalbay and O. B. Korkmazoglu, "Effects of Multicollinearity on Electricity Consumption Forecasting using Partial Least Squares Regression," *Procedia - Social and Behavioral Sciences*, vol. 62, no. 1877-0428, pp. 1150-1154, 2012.
- [13] J. Keirstead, M. Jennings and A. Sivakumar, "A review of urban energy system models: Approaches, challenges and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3847-3866, 2012.