Electricity Demand and Population Dynamics Prediction from Mobile Phone Metadata

Brian Wheatman[2], Alejandro Noriega[1](✉), and Alex Pentland[1]

[1] Media Laboratories, MIT, Cambridge, MA, USA
noriega@mit.edu
[2] Computer Science and Electrical Engineering, MIT, Cambridge, MA, USA

Abstract. Energy efficiency is a key challenge for building modern sustainable societies. World's energy consumption is expected to grow annually by 1.6 %, increasing pressure for utilities and governments to fulfill demand and raising significant challenges in generation, distribution, and storage of electricity. In this context, accurate predictions and understanding of population dynamics and their relation to electricity demand dynamics is of high relevance.

We introduce a simple machine learning (ML) method for day-ahead predictions of hourly energy consumption, based on population and electricity demand dynamics. We use anonymized mobile phone records (CDRs) and historical energy records from a small European country. CDRs are large-scale data that is collected passively and on a regular basis by mobile phone carriers, including time and location of calls and text messages, as well as phones' countries of origin. We show that simple support vector machine (SVM) autoregressive models are capable of baseline energy demand predictions with accuracies below 3 % percentage error and active population predictions below 10 % percentage error. Moreover, we show that population dynamics from mobile phone records contain information additional to that of electricity demand records, which can be exploited to improve prediction performance. Finally, we illustrate how the joint analysis of population and electricity dynamics elicits insights into the relation between population and electricity demand segments, allowing for potential demand management interventions and policies beyond reactive supply-side operations.

1 Introduction

In today's developing world, efficient energy procurement is a key challenge for building sustainable societies. The world's energy consumption is expected to grow annually by 1.6 %, increasing pressure for utilities and governments to fulfill demand and raising significant challenges in generation, distribution, and storage of electricity. In this context, accurate predictions of electricity demand are of salient relevance for supply-side operations by allowing efficient use of the installed capacity for generation, distribution and storage, as well as efficient electricity purchasing and trading. Moreover, reliable electricity demand predictions can enable the incorporation of low-carbon technologies into electricity grids [9].

K.S. Xu et al. (Eds.): SBP-BRiMS 2016, LNCS 9708, pp. 196–205, 2016.
DOI: 10.1007/978-3-319-39931-7_19

Recent research has developed various methodologies for electricity demand prediction. Most prediction methodologies involve the use of diverse datasets, such as regional weather forecasts [15], calendar data, building construction materials [10], and building occupancy rates estimated from WiFi connections [11]. These data sources are often only partially available over large regions and are subject to human and institutional boundaries as well as error in their generation. Moreover, only a few of these – such as research on the use of WiFi data that estimates dynamic building occupancy rates – capture the human dynamics element underlying energy consumption.

Anonymous mobile phone records – or CDRs (Call Detail Records) – have become one of the most salient sources of information that elicit large-scale patterns of human activity. CDRs contain metadata on the social and mobility patterns of users and are generated by mobile network infrastructure on a regular basis. In recent years, researchers have developed applications of CDRs in domains relevant and diverse as crime prediction [3], population modeling for disaster response in earthquakes and floodings [1,7], modeling of epidemic outbreaks [6,14], inferring local socio-economic statistics in both the developed and developing worlds [5,12], and urban transportation systems development [2]. In addition, handsets and airtime are becoming cheaper, leading to ubiquitous mobile phone penetration, which by 2013 approached 90 % in developing countries and 96 % globally [13].

Today, few studies have explored the intersection of electricity demand and population dynamics elicited from large-scale mobile phone records datasets (CDRs). The novel intersection of these two perspectives on our built systems can prove valuable in (1) increasing performance of electricity demand predictive models, useful for efficient supply-side management of generation, distribution and storage and (2) uncovering insights on population to grid dynamic relationships, which can allow for policies that go beyond reacting to demand into shaping it.

In this work, we jointly examine population dynamics extracted from anonymous CDRs and electricity demand dynamics from hourly electric grid records. The datasets encompass all call, texts and data connections and all electricity consumptions for a small European country over an overlapping period of nine months. Section 2 describes the datasets used.

Section 3 builds the baseline purely autoregressive models – those in which feature variables are exclusively previous realizations of the predicted variable – for (1) prediction of the daily amount of active population in the country, segmented by country of origin and (2) prediction of hourly electricity demand, segmented by region within the country[1]. We show that these basic benchmark models are capable of predicting hourly energy demand at percentage errors below 3 % and daily population activity at percentage errors below 10 % (Tables 1 and 2). Section 4 explores the joint information carried by mobile phone and energy records. We show that population dynamics extracted from CDRs significantly

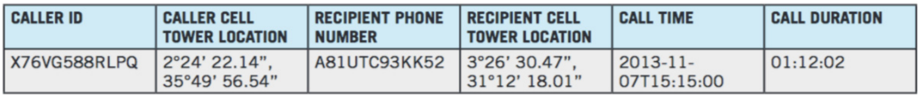
[1] There is a large amount of tourism in the country. Over the time analyzed, on roughly one in four people connecting to a cell tower were not from the local country.

correlate with the errors yielded by the energy autoregressive models (Table 3) and that these correlations can be exploited towards energy prediction performance improvements. Finally, Sect. 5 illustrates how the joint analysis of population and electricity dynamics can elicit insights on the relation between population and electricity demand segments (Table 4), potentially allowing for demand management interventions and policies beyond reactive supply-side operations.

2 Datasets Description

2.1 Mobile Phone Records

Mobile phone records (CDRs) consist of metadata about call, short message service (SMS) and data communications, such as location and time of call, but provide no information on the content of the communication. Figure 1 shows an example of the call detail record of a phone call.



CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24' 22.14", 35°49' 56.54"	A81UTC93KK52	3°26' 30.47", 31°12' 18.01"	2013-11-07T15:15:00	01:12:02

Fig. 1. Example of a CDR record

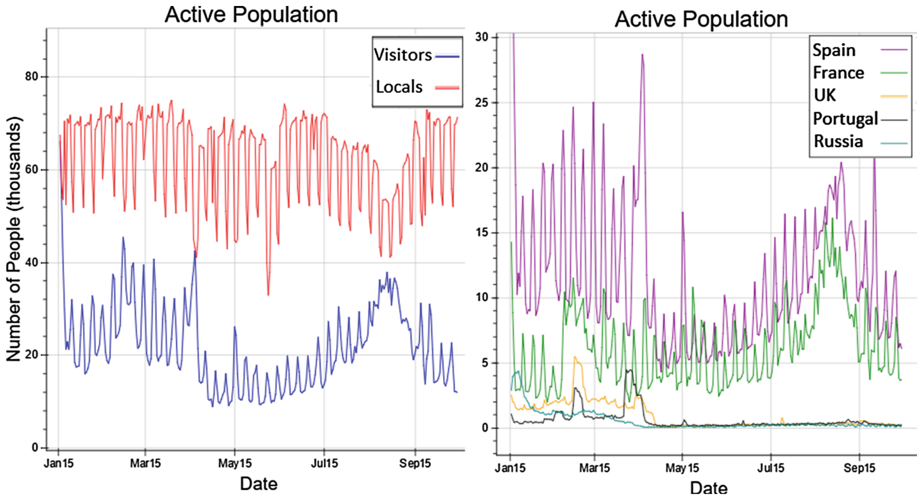
Relevant characteristics of CDRs are (1) the caller and receiver identities are pseudonymized, i.e., names and phone numbers are replaced by anonymous codes and (2) the geographic location of towers used for each communication provide an approximation of users' location. Additionally, CDRs may contain mobile country codes (MCC) and type allocation codes (TAC), which encode the home country of the mobile subscription and the mobile phone model respectively.

The CDR dataset used in this study comprises metadata, including MCC and TAC codes, of all calls, SMS, and data communications of a small European country over a period of nine months in 2015. From raw CDRs, we compute daily active population, defined as the number of users that engaged in at least one communication within the country on a given day. Figure 2 shows population dynamics for locals and visitors over the period studied[2]. Segmentation by country of origin is germane as the tourism industry plays a central role in the country's economy.

2.2 Energy Records

Energy records used in this study are registries of all electric current flowing through the national grid. This and the CDRs dataset are available for an overlapping period of nine months in 2015. Temporal resolution for aggregate current volumes is hourly. Geographic segmentation is naturally defined by four main high voltage lines. Figure 3 shows the daily, weekly and seasonal dynamics of energy demand, segmented by region.

[2] Shown are countries with highest amount of visitors out of more than 50 countries.



(a) (b)

Fig. 2. (a) Active population of locals vs. visitors. (b) Active population of top five countries of visitors.

3 Baseline Autoregressive Models

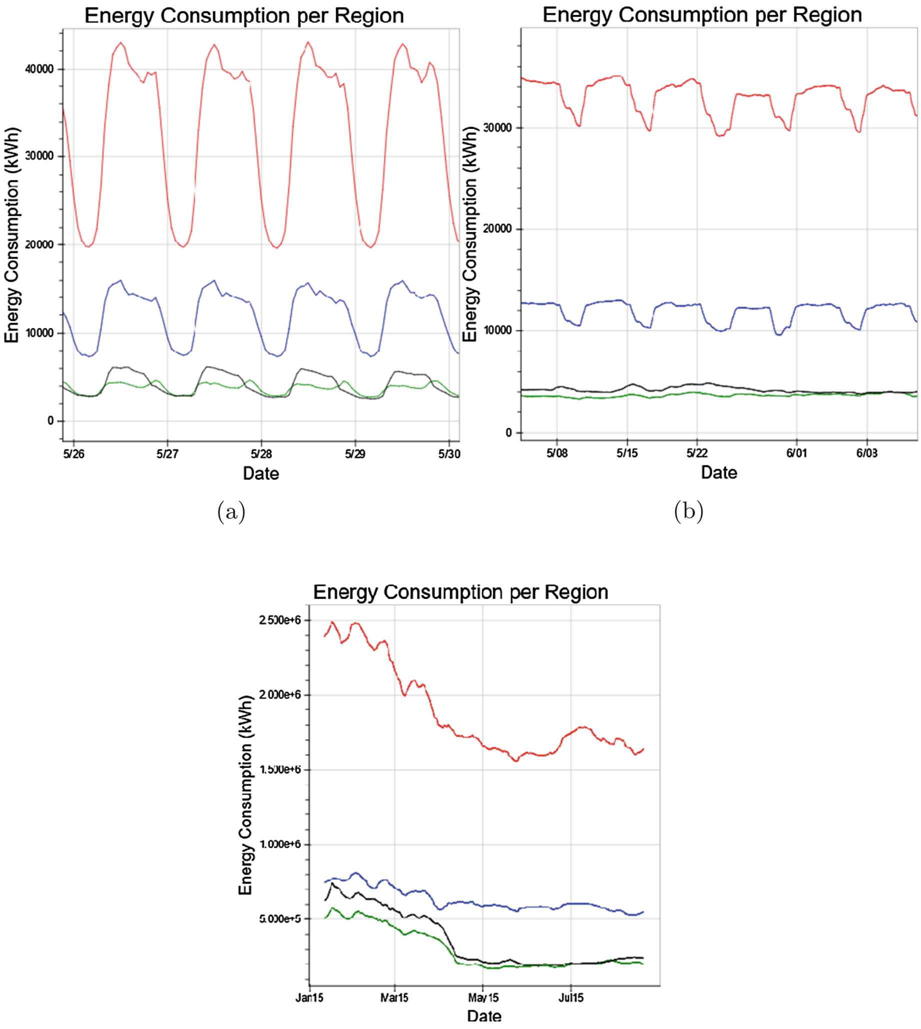
We build baseline purely autoregressive models – i.e., models in which feature variables are exclusively previous realizations of the predicted variable [4] – for (1) prediction of the daily amount of active population in the country, segmented by country of origin and (2) prediction of hourly electricity demand, segmented by region within the country. All prediction tasks are solved for a horizon (h) 24 h in the future (also called day-ahead predictions).

These models are meant as modern baselines for more sophisticated prediction methodologies in terms of information inputs and statistical learning methods. We use standard statistical and machine learning (ML) methodologies:

- Feature vectors used are composed of autoregressive values of the 14 days prior to the prediction, which allows us to capture daily and weekly patterns shown in Fig. 3.
- Standard ML methodology is used for training, cross-validating, model selection, and testing[3].
- Regression models used are a linear Lasso and a support vector machine (SVM) regression, with a radial basis function (RBF) kernel [8].

We assess predictive performance in terms of percentage errors (PE), defined in Eq. 1 and normalized mean square errors (NMSE) defined in Eq. 2. PEs provide an intuitive interpretation, which is easily translatable to value metrics such

[3] We train on 150 days and test on subsequent 30 days. We optimize regularization parameters on a sequence of sequential 180 day blocks and assess prediction on a final set of 30 unseen days.



(a)

(b)

(c)

Fig. 3. (a) Daily cycle of energy demand. (b) Weekly cycle of energy demand. (c) Seasonal pattern of energy demand.

as trading cost and energy waste. Complementarily, NMSEs provide a natural benchmark for predictive models, where NMSE < 1 entails that a model performs better than the sample mean of the predicted variable [4]. Finally, we benchmark energy predictions against those yielded by the commercial predictive tool currently used by the national utility company, developed by a strong player on the European market for energy forecasts.

$$\mathrm{PE} = \frac{1}{n} \sum_{\forall i} \frac{y_i - \hat{y}_i}{y_i} \tag{1}$$

$$\mathrm{NMSE} = \frac{\sum_{\forall i} (y_i - \hat{y}_i)^2}{\sum_{\forall i} (y_i - \mu)^2} \tag{2}$$

3.1 Population Dynamics

We use an autoregressive linear Lasso for population dynamics prediction. This standard model uses a square loss function and an $L1$ norm, performing thus regularization and variable selection, as specified by Eq. 3.

$$\min_{A \in \mathbf{R}^{15}} \left\{ \frac{1}{N} \Big(y_i - a_0 - \sum_{t=-15}^{-1} a_t y_{it} \Big)^2 + \lambda \|A\|_1 \right\} \tag{3}$$

Table 1 shows that the baseline autoregressive Lasso can predict active population at percentage errors near 5 % for Total and Locals population segments. It also shows that the smaller the group, the harder it is to predict with pure autoregressive models, yielding higher PEs for the Spanish and French population predictions. All three population segments had NMSE < 1 where, for example, the Locals predictor yielded errors of only 34 % of errors yielded by the basic benchmark of using y's sample mean as a constant predictor.

Table 1. Baseline autoregression predictions of active population per country of origin

Population	Percentage error	Normalized mean square error
Total	3.61 %	0.270
Locals	5.85 %	0.336
Spain	11.81 %	0.687
France	23.89 %	0.653

3.2 Energy Demand

We predict hourly electricity demand for each region using an analogous autoregressive approach. Here we implement a linear Lasso and a SVM with RBF kernel as regression models.

Table 2 shows that baseline autoregressive models are able to predict electricity demand at percentage errors of 2.3 % for total demand and around 3 % for regional demands. All models' performances were one or two orders of magnitude better than benchmark performance of y's sample mean as predictor (NMSE $\ll 1$). Moreover, the baseline autoregressive models yielded comparable results to the commercial tool currently used by the country's national utility and performed substantially better for Full Country and Region 3 predictions.

Table 2. Baseline autoregression predictions of energy demand per region

Energy	Commercial tool PE	Lasso PE	Lasso NMSE	SVM PE	SVM NMSE
Full Country	2.80 %	2.29 %	0.0141	2.32 %	0.0117
Region 1	2.69 %	3.39 %	0.0302	3.06 %	0.0333
Region 2	2.68 %	2.80 %	0.0192	3.17 %	0.0178
Region 3	6.41 %	3.48 %	0.0774	3.45 %	0.0755
Region 4	No Data	3.26 %	0.0205	3.26 %	0.0177

4 Energy Demand Predictions Using Population Dynamics

In this section, we explore the joint information carried by mobile phone and energy records.

We evaluate the correlation of the segmented population predictions of Sect. 3.1 with the errors yielded by the energy prediction autoregressive SVM model in Sect. 3.2. We show that population dynamics extracted from CDRs significantly correlate with the errors yielded by the energy autoregressive models and that these correlations can be exploited towards energy prediction performance improvements.

Table 3 shows that three out of four electricity demand regions present significant correlations between the errors of their autoregressive models and segmented population predictions, based on mobile phone data.

The additional information contained in mobile phone data can be leveraged towards improved prediction performance. To illustrate this application, we implemented a standard SVM regression model where the predicted variable is the hourly energy prediction error from Sect. 3.2's autoregression models and features are population predictions from Sect. 3.1's autoregression models. This simple sequential approach was able to reduce percentage errors of Sect. 3.2's baseline models in 4.2 % and 20.5 % for Region 1 and Region 4 respectively. No improvement was achieved for Region 2 and Region 3.

5 Relating Segments of Population and Energy Demand

Lastly, we illustrate in this section how the joint analysis of population and electricity dynamics can elicit insights on the relation between population and electricity demand segments. The motivation for this analysis is that such relations can, in addition to improving predictive power, potentially allow for demand management interventions and policies beyond reactive supply-side operations.

We solve simple linear regressions of electricity demand for each region against segmented population activity, as shown by Eq. 4, where explanatory variables x_{ij} denote active population of segment j for day i. Coefficients of these regressions elicit relationships between energy and population segments.

Table 3. Correlations between electricity demand prediction errors and segmented population predictions

Region 1	Correlation coefficient	P-Value
Locals	0.070	0.308
Spain	−0.213	0.0015***
France	−0.170	0.0118**
UK	−0.0516	0.447
Region 2	Correlation coefficient	P-Value
Locals	0.105	0.123
Spain	0.060	0.378
France	−0.0081	0.904
UK	−0.010	0.881
Region 3	Correlation coefficient	P-Value
Locals	−0.086	0.206
Spain	0.169	0.018**
France	0.083	0.223
UK	0.016	0.816
Region 4	Correlation coefficient	P-Value
Locals	−0.139	0.0385**
Spain	0.116	0.0862*
France	0.154	0.0227**
UK	−0.0288	0.671

* 90 %, ** 95 %, and *** 99 % confidence levels.

These coefficients capture the effect that a singular increase in the active population of a country has on the energy consumption of a region, also known as the unique or *ceteris paribus* effect.

$$\min_{\beta} \sum_{\forall i} \epsilon_i \qquad \text{S.T. } y_i = \beta_0 + \sum_{\forall j} \beta_j x_{ij} + \epsilon_i \tag{4}$$

Table 4 shows how different energy regions are affected by population segments. For example, we see how a unit increase in UK visitors tends to have a larger impact on Region 2's electricity demand than a unit increase of Spanish visitors does, and that the region most affected by an increase in visitors of any nationality is Region 2, the capital. Future research paths may explore the network of relationships among more specific energy and population segments.

6 Concluding Remarks

We proposed standard Lasso and SVM autoregressive models as basic benchmarks for segmented predictions of large-scale population and electricity demand

Table 4. Population segments on electricity segments regression coefficients

Coefficients	Locals	Spain	France	UK
Full Country	0.385**	0.567**	−0.644	6.98**
Region 1	0.0710**	0.0644*	0.0146	0.436*
Region 2	0.197**	0.240**	−0.244	3.04**
Region 3	0.0495**	0.118**	−0.164*	1.54**
Region 4	0.0682**	0.145**	−0.2509*	1.96**

* 95 % confidence level
** 99 % confidence level

dynamics. We showed that these baseline models can yield accuracies below 3 % and 10 % percentage errors for electricity demand and active population predictions respectively, and that segmented predictions are possible at similar accuracy levels. Moreover, we showed that population dynamics extracted from CDRs contain information additional to that of electricity dynamics, and that this information can be leveraged toward improved accuracy of electricity demand predictions. Finally, we illustrate how the joint analysis of these datasets can elicit the underlying network of relations among population and electricity segments.

Future work may develop more sophisticated models that incorporate richer information such as weather, calendar and building construction data and implement more sophisticated feature engineering, feature selection, and regression models. Pure autoregressive models proposed here are meant as modern benchmarks for more sophisticated models and commercial tools.

A promising research path forward consists of using finer segmentation in both predictive and explanatory models. Possible paths include disaggregated segmentation of the electric grid per town, industry or building complex and additional population segmentations such as social and spatial behavior characterizations and disposable income proxies inferred from users' mobile phone models. Finally, as pointed out in Sect. 5, further eliciting the underlying relations among energy and population segments can, in addition to improving predictive power, allow for demand management interventions and policies beyond reactive supply-side operations.

References

1. Bengtsson, L., Lu, X., Thorson, A., Garfield, R., Von Schreeb, J.: Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med **8**(8), e1001083 (2011)
2. Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M.L.: AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part III. LNCS, vol. 8190, pp. 663–666. Springer, Heidelberg (2013)

3. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 427–434. ACM (2014)
4. Bontempi, G.: Machine Learning Strategies for Time Series Prediction, Hammamet (2013)
5. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)
6. Frias-Martinez, E., Williamson, G., Frias-Martinez, V.: An agent-based model of epidemic spread using human mobility and social network information. In: 3rd International Conference on Social Computing, SocialCom (2011)
7. Global Pulse, U.N.: Using mobile phone activity for disaster management during floods. Technical report, UN Global Pulse (2013)
8. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2013)
9. Kalogirou, S.A.: Artificial neural networks in renewable energy systems applications: a review. Renew. Sustain. Energy Rev. **5**(4), 373–401 (2001)
10. Kolter, Z., Ferreira, J.: A large-scale study on predicting and contextualizing building energy usage. In: Proceedings of the Conference on Artificial Intelligence (AAAI) (2011)
11. Martani, C., Lee, D., Robinson, P., Britter, R., Ratti, C.: Enernet: Studying the dynamic relationship between building occupancy and energy consumption. Energ. Build. **47**, 584–591 (2012)
12. Smith, C., Mashhadi, A., Capra, L.: Ubiquitous sensing for mapping poverty in developing countries. In: Proceedings of the Third Conference on the Analysis of Mobile Phone Datasets (2013)
13. United Nations, I.T.U.: Facts and figures. Technical report, UN International Communications Union (2013)
14. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O.: Quantifying the impact of human mobility on malaria. Science **338**(6104), 267–270 (2012)
15. Zhao, Hx, Magoulès, F.: A review on the prediction of building energy consumption. Renew. Sustain. Energy Rev. **16**(6), 3586–3592 (2012)