Bogomolov et al. *EPJ Data Science* (2016) 5:13
DOI 10.1140/epjds/s13688-016-0075-3

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Energy consumption prediction using people dynamics derived from cellular network data

Andrey Bogomolov[1*], Bruno Lepri[2], Roberto Larcher[3], Fabrizio Antonelli[3], Fabio Pianesi[2] and Alex Pentland[4]

[*]Correspondence: andrey.bogomolov@unitn.it
[1]University of Trento, via Sommarive, Trento, Italy
Full list of author information is available at the end of the article

Abstract

Energy efficiency is a key challenge for building sustainable societies. Due to growing populations, increasing incomes and the industrialization of developing countries, the world primary energy consumption is expected to increase annually by 1.6%. This scenario raises issues related to the increasing scarcity of natural resources, the accelerating pollution of the environment, and the looming threat of global climate change.

In this paper we introduce a new and original approach to predict next week energy consumption based on human dynamics analysis derived out of the anonymized and aggregated telecom data, which is processed from GSM network call data records (CDRs). We introduce an original problem statement, analyze regularities of the source data, provide insight on the original feature extraction method and discuss peculiarities of the regression models applicable for this big data problem.

The proposed solution could act on energy producers/distributors as an essential aid to smart meters data for making better decisions in reducing total primary energy consumption by limiting energy production when the demand is not predicted, reducing energy distribution costs by efficient buy-side planning in time and providing insights for peak load planning in geographic space.

Keywords: energy consumption prediction; mobile phone data; human dynamics; machine learning

1 Introduction

Energy efficiency is a key challenge for building sustainable societies. Due to growing populations, increasing incomes and the industrialization of developing countries, the world primary energy consumption is expected to increase annually by 1.6%. This scenario raises issues related to the increasing scarcity of natural resources, the accelerating pollution of the environment, and the looming threat of global climate change.

In order to improve the efficiency of the supply systems and thus to reduce the amount of energy consumption, a critical step is to understand energy needs at relatively high spatial and temporal resolution. An accurate prediction of energy demands could provide useful information to make decisions on energy generation and purchase. Furthermore, an accurate prediction would have a significant impact on preventing overloading and allowing an efficient energy storage. Hence, several computational works have started developing machine learning models to predict the energy consumption of residential and commercial buildings using features such as weather and energy bills [1]. For example, Kolter and Ferreira [2] used monthly electricity and gas bills and buildings' characteristics to model energy consumption. Other studies investigated the relationship between the human occupancy of buildings and the consumption patterns, using WiFi connections as a proxy for human occupancy [3].

Nowadays, the almost universal adoption of mobile phones is generating an enormous amount of data about human behaviors with a breadth and depth that was previously inconceivable [4]. In 2013, there was 6.8 billion of mobile phone subscribers worldwide, with millions of new subscribers every day,[a] and several studies have shown that the mobile phone data, specifically the Call Detail Records (CDRs) needed by the mobile phone operators for billing purposes, can be exploited to model individuals' mobility patterns [5–7] and to map the distribution of the population in space and time [8]. Not surprisingly, a couple of works have proposed to use mobile phone data for the design and the planning of energy systems and infrastructures [9, 10]. However, with the exception of a very recent work by [11] using 'Data for Development' (D4D) data from Senegal [12], no quantitative studies have investigated the potential of mobile phone data to understand energy consumption.

In the current paper, we propose and evaluate the usage of anonymized and aggregated people dynamics features, derived from the mobile phone network activity, to predict energy consumption. Specifically, we target two different tasks of paramount importance to increase the efficiency of energy producers and distributors and to meet consumers' peak demands: (i) predicting the *daily average energy consumption* and (ii) predicting the *peak daily energy consumption*. It is worth to notice that none of the anonymized and aggregated people dynamics features can be traced back to make inferences about individuals and hence there are minimal - if any - privacy concerns.

To validate our approach we use mobile phone records from a territory in the Northern Italy, the province of Trentino. The data, released for the Telecom Italia Big Data Challenge 2014, were collected from November 1, 2013 to December 31, 2013 [13].

Our results prove that people dynamics, extracted from aggregated and anonymized mobile phone data, are good proxies for modeling energy consumption.

2 Datasets description

In this section we introduce the datasets that have been used to evaluate our approach: (i) an *energy consumption dataset* and (ii) a *mobile phone records dataset*. The datasets were collected from November 1, 2013 to December 31, 2013 over a territory of 6,000 square kilometers in the Northern Italy, the province of Trento. The datasets contain 50 thousand records for energy consumption and 600 million data records concerning telecommunication events respectively [13]. The two datasets have also the same spatio-temporal aggregation. The temporal aggregation is of ten minute intervals, while the spatial one results by partitioning the territory using a regular square grid. Each square of the grid measures approximately 1 square kilometer. In our paper, we refer to this grid as the *partitioning grid*.

2.1 Energy consumption dataset

The *energy consumption* dataset is provided by the local energy company, SET, that manages almost the entire electrical network over the Trentino territory. SET uses around 180 primary (medium voltage) distribution lines to bring energy from the national grid (high voltage) to Trentino's consumers. To ensure the privacy of SET's customers, their locations and the geometry of the 180 primary distribution lines is not explicitly exposed.

Consequently, the *Customer site dataset* shows the number of customer sites of each power line per grid square, while the *Line measurement dataset* indicates the amount of flowing energy through the lines at time t. Customer sites provide energy to different types of customers (e.g. houses, condominiums, business activities, industries etc.), which require different amount of electricity. For privacy reasons this information is hidden, meaning that in the dataset the energy flowing is uniformly distributed among the various types of customers.

In Figure 1 we show the process done by the organizers of the Telecom Italia Big Data Challenge 2014 to transform the original dataset to the one we had access. In the first layer there is the exact position of each customer site (e.g. some of them are industries, others are small houses) and the precise geometry of each line. In the second layer we lose the exact geometries of customer sites and power lines. However, this information is summarized in the *Customer site dataset* where for each square grid the number of customer sites is recorded along with the information about the power line they are connected to. In the third layer we know how the customer sites of a power line are distributed over the grid and the energy flowing through each power-line (from the *Line measurement dataset*). It is then possible to distribute the energy flowing through a powerline p over the grid in or-

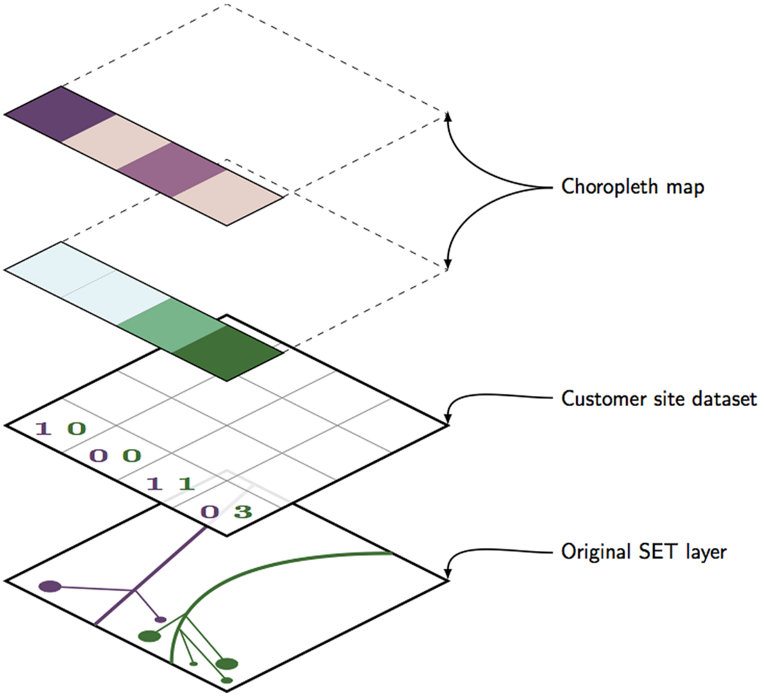

Figure 1 The SET customers are spatially aggregated into the grid squares and the energy consumption is uniformly divided among the customers, hiding their different type (e.g. houses, condominiums, business activities, industries).

der to build a choropleth map of the energy consumption in each partitioning grid square (last layer in Figure 1).

In sum, the structure of the *Customer site dataset* is the following:

- *Square id*: identification string of a given square of the partitioning grid;
- *Line id*: identification string of the distribution power line, which is grouped with the partitioning grid square;
- *Number of customer sites*: number of customer sites present in a given square of the partitioning grid, connected to the grid powerline (*Line id*).

Instead, the *Line measurement dataset* is composed by:

- *Line id*: identification string of the distribution power line;
- *Timestamp*: timestamp relative to the instant when the measurement of the current passing through the power line is done. Date in the format `YYYY-MM-DD HH24:MI`;
- *Value*: the ampere value of the current passing through a given powerline (*Line id*) at a given *Timestamp*. This quantity is positive if the direction of the current goes from the national grid into the local line, negative otherwise.

2.2 Call Detail Records

The *Call Details Records* dataset contains anonymized and aggregated incoming and outgoing calls, received and sent SMSs, and Internet connection events, generated from November 1, 2013 to December 31, 2013 by the cellular network of Telecom Italia Mobile, the largest mobile operator in Italy with 34% of the entire market share.

The dataset is composed by three sub-datasets: (i) the *Telecommunications Activity* dataset providing the activity of Trentino, showing all the mentioned telecommunication events which took place within this area. The data provides information of Telecom Italia's customers interacting with the network and of other people using it on roaming. For each square of the *partitioning grid* the dataset provides every ten minutes the activity in terms of sent and received SMSs, issued calls, received calls and Internet traffic related events. The information is aggregated using the country code, which has a different semantic for each kind of activity (e.g. the country of the person receiving/sending the message, the country of the person receiving/issuing the call, the country of the person connected to Internet), (ii) the *Telecommunications - Square to Counties* dataset providing the level of interaction between each square of the *partitioning grid* and the national counties. The level of interaction between a square A and a county B is given as a pair of decimal numbers. The first number is proportional to the number of calls issued from the square A to the county B, the second one is proportional to the number of calls from the county B to the square A. The *temporal aggregation* is done in timeslots of ten minutes, and (iii) the *Telecommunications - Square to Square* dataset providing information regarding the directional interaction strength between each pair of squares of the *partitioning grid*. The directional interaction strength between the square A and the square B is proportional to the number of calls issued from the square A to the square B. Again, the *temporal aggregation* is done in timeslots of ten minutes.

3 Methodology

We formulate the problem of predicting the *electric energy consumption* of a given geographical area as a nonlinear regression task. More specifically, we deal with two different prediction tasks: (i) *average daily energy consumption* and (ii) *peak daily energy consumption*. Each task is solved for the next 7 days interval for each electric line ID. This setting is justified by the economic and managerial value of the expected output - it is easy to plan energy supply for the next week, given we have the predicted energy consumption demand.

Electric energy consumption is measured in W · h (Watt × Hour). In terms of electromagnetism, one Watt is the rate at which work is done when one Ampere (A) of current flows through an electrical potential difference of one Volt (V). Assuming that electrical potential, measured in V is standardized in Trentino province (thus is equal for all line IDs) and given the same timeframes for the analysis, the electric energy consumption prediction task reduces to predicting electric current measured in Ampere per each time frame per each line ID. The values in Ampere of the current passing through the given power line are given by the electric energy distribution company.

Forecasting model training and prediction is done for daily intervals in high order Hilbert space, derived from the anonymized and aggregated mobile network activity in Trentino. The features which are extracted from the source data characterize *diversity*, *regularity* and *general human dynamics* in each small part of the Trentino territory spatially separated by square grid.

In sum, the proposed technical solution includes the following main steps:

1. An highly parallelized feature extraction algorithm, which characterizes *diversity*, *regularity* and *general human dynamics*, derived from telecommunication data and aggregated by the square grid areas, including innovative second-order features in time and frequency domains;
2. A feature selection algorithm (32 features for the final models are selected out of >3,000 features), thus reducing the computational complexity of the model;
3. A non-linear regression modeling and prediction based on ensemble of decision trees, which are bootstrapped and aggregated;
4. A model generalization strategy, as opposed to data overfitting, including strict separation of the test set from the training set (the test set is the next week after the training set with the dependent variables taken with 7-days shift to the future), random splits, bootstrapping and bagging techniques.

In the next subsections we provide further details of the experimental setup we followed (preliminary data analysis, feature extraction, feature selection, and model building).

3.1 Preliminary data analysis

As preliminary analysis we performed a spectrogram of the temporal current line (see Figure 2) in order to visually justify the feature extraction approach described in the next section. In Figure 2, the horizontal axis represent days (the temporal scaling was done starting from the 100 milliseconds initial resolution), while the vertical axis represents frequency. The amplitude of a particular frequency at a given time is given by the intensity and the color of each point in the plot.

As expected, we found that the response variable - the measure of the amount of electric current passing a point in an electric circuit per unit of time for each power line - has a number of cyclic characteristics and trends. Cyclic characteristic of a time series in data analysis is called *seasonality* - a property of a signal, experiencing regular changes, which recur every observed time frame, e.g. daily, weekly, yearly. We found predictable changes of the pattern in a response variable time series, that repeat over daily and weekly periods. Interestingly, these temporal regularities were characteristics of different locations of the

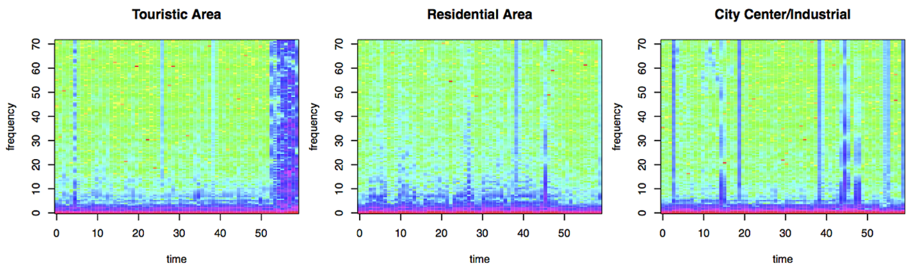

Figure 2 Spectral characteristics of typical energy consumption response variables. The units reported in the temporal *x*-axis are days (the temporal scaling was done starting from the 100 milliseconds initial resolution), while the vertical axis represents frequency. The amplitude of a particular frequency at a given time is given by the intensity and the color of each point in the plot.

Province of Trento. Hence, we were able to identify three possible clusters roughly corresponding to (i) the residential areas, (ii) the touristic areas and (iii) the city center areas and/or industrial areas (see Figure 2).

Specifically, we separated the energy consumption signal into three major components: daily seasonality, trend and a remainder component applying seasonal-trend decomposition procedure based on loess [14]. An interesting result for each power consumption cluster type is presented in Figures 3, 4 and 5.

As shown in Figure 3, the typical energy consumption behavior of a residential area shows uneven seasonality on weekly scale, varying seasonality during day and night, variable consumption during the weekdays, low consumption during holidays, and low noise of measurements.

Turning our attention to the typical energy consumption behavior of a touristic area (e.g. Cavalese, a small village and very famous ski resort in Fiemme Valley), we observed uneven seasonality on a weekly scale, varying seasonality during day and night, variable consumption during the weekdays, upward sloping trend toward holidays, abnormally high load during holidays, and noisy values, which are probable effects of solar energy production. In particular, the significant increase in energy consumption during weekends and holidays is justified in Northern Italy, where a large amount of people leaves the major cities to reach mountain touristic locations.

Interestingly, no significant differences were found for city center areas and industrial areas. They both show stationary seasonality on weekly scale, stationary seasonality during day/night, stable consumption during the weekdays, low consumption during weekends, and low noise of measurements.

The discovered seasonalities were explicitly coded into the feature space by using a number for the hour of the day and a number for the weekday for each data source being processed.

3.2 Feature extraction

To solve the problem of computational complexity due to the huge amount of data samples (>600 millions of Call Data Records) we moved from the time domain of communication patterns to the frequency domain, applying the Fast Fourier Transform algorithm to each group of daily time series. Also we found that only a small set of harmonics in Fourier domain explains the response variable variance for each type of first-order feature space time series, which reduces the computational complexity by a number of orders. The usage

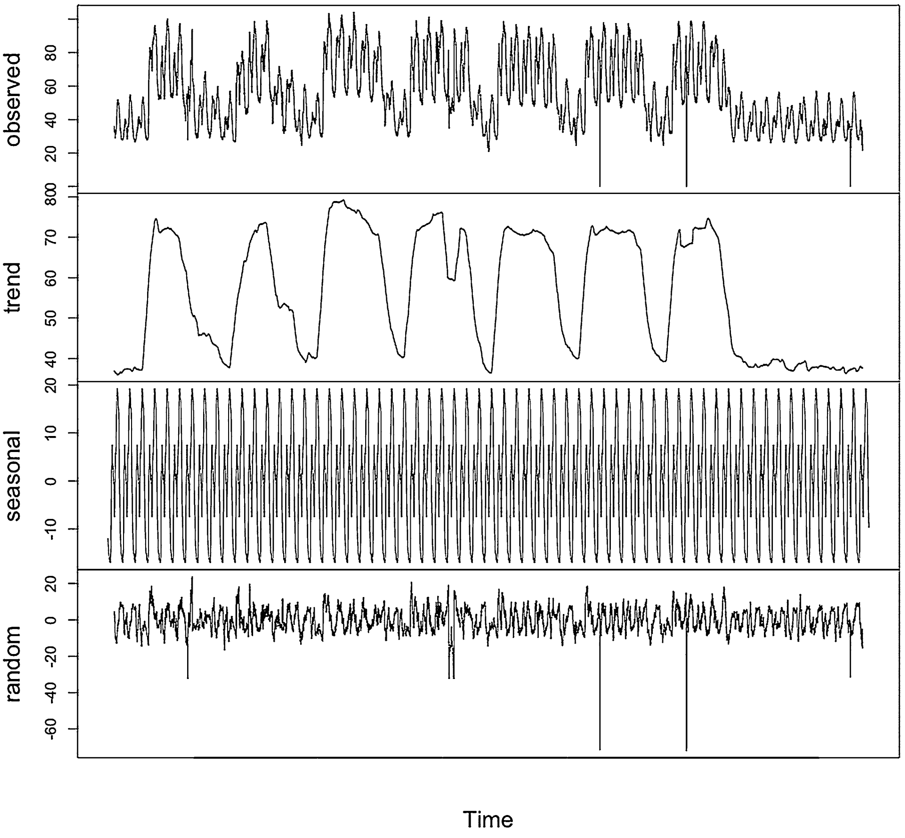

Figure 3 Energy consumption time series decomposition for a residential area. The *x*-axis shows the temporally sampled electric current flowing through the distribution lines recorded every 10 minutes. The *y*-axis reports the electric current passing through a given power line measured in Ampère at a given time. Monday is the first day of the week.

of a limited number of harmonics in Fourier domain is a known method of compression, which is frequently applied in the field of digital signal processing [15]. For example, some lossy image and sound compression methods employ discrete Fourier transforms. In our experiments, we used from 16 to 64 Fourier coefficients, which are enough to represent the temporal properties of the communication data.

Diversity and *regularity* have been shown to be important in the characterization of different facets of human behavior and, in particular, the concept of entropy has been applied to assess the predictability of mobility [6] and spending patterns [16, 17], the socio-economic characteristics [18] and the crime levels [19, 20] of cities and some individual traits such as personality [21]. Hence, for each variable from CDRs we computed the mathematical functions, which characterize the distributions and measure the information theoretic and statistical properties of such variables, e.g. mean, median, standard deviation, min and max values and Shannon entropy.

In order to be able to also account for temporal relationships, the same computations as above were repeated on sliding windows of variable length (1-hour, 4-hour and 1 day), producing *second-order features* that capture spatio-temporal relationships, thus preserving useful source data properties.

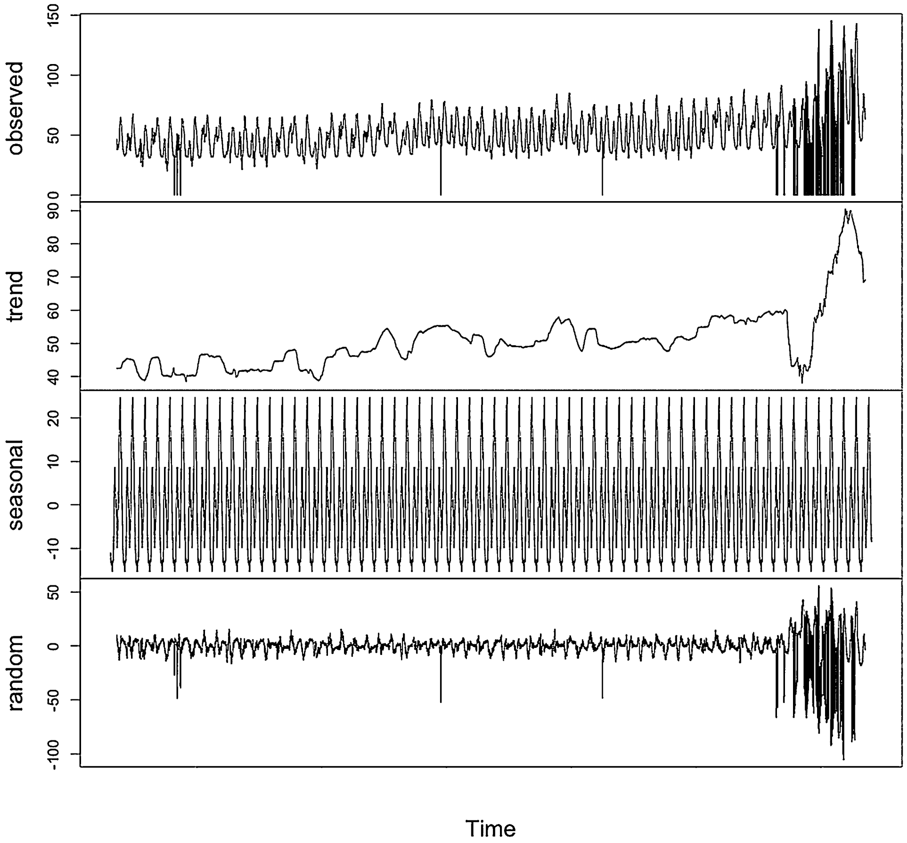

Figure 4 Energy consumption time series decomposition for a touristic area. The *x*-axis shows the temporally sampled electric current flowing through the distribution lines recorded every 10 minutes. The *y*-axis reports the electric current passing through a given power line measured in Ampère at a given time. Monday is the first day of the week.

It is worth noticing that in the computation of the distributions' properties in frequency domain we do not limit the higher-order functions to metrics with an intuitive explanation from physics. For example, the 'variance of real numbers part of Fourier transform of area codes' represents, for each spatial square, a measure of diversity of the area codes of telecommunication activity.

3.3 Feature selection

In order to reduce model complexity and enhance generalization properties by reducing the risk of overfitting [22], a feature selection step was performed before the model building. The feature selection was done on a reduced sample of the training data, which was one week long. The metric used to rank the features was the total decrease in node impurities, which is the impurity measure of a decision tree node derived from relative entropy metric [23]. This choice was motivated because it outperformed other metrics such as mutual information, information gain, and chi-square statistic [24, 25]. We reduced the feature space only to 32 dimensions for each of the two models without loosing much accuracy. The 32 dimensions were chosen because the addition of other features increased the computational complexity without improving significantly the performance. The fi-

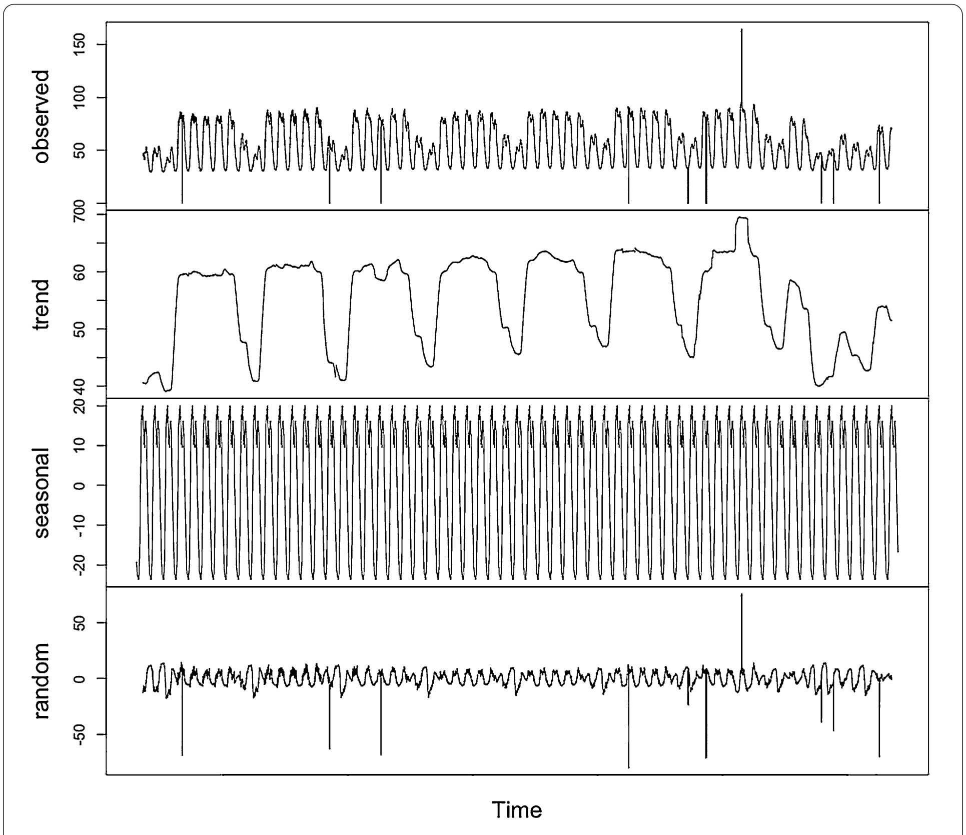

Figure 5 Energy consumption time series decomposition for a city center/industrial. The *x*-axis shows the temporally sampled electric current flowing through the distribution lines recorded every 10 minutes. The *y*-axis reports the electric current passing through a given power line measured in Ampère at a given time. Monday is the first day of the week.

nal feature sets are provided in Tables 3 and 4. In these tables the mean decrease in node impurity is presented in non-normalized form.

3.4 Model building

We formulated two separate problems - (1) predicting *mean daily consumption* and (2) predicting *peak daily consumption*. To this end, we built 2 regression models. For each of the regression models and for each sample we have a scalar outcome variable, $Y \in \mathbb{R}$ and a vector of explanatory variables in the selected feature space $\vec{X} \in \mathbb{R}^d$.

Our goal was to estimate the regression function

$$\bar{r}(x) = \mathbb{E}[Y|\vec{X} = x] \tag{1}$$

for any $x \in$ the space $\mathbb{R}$, by generating decision trees at random and combining them to form the aggregated regression estimate

$$\bar{r}(\vec{X}, \Omega_n) = \mathbb{E}_\Theta\big(r_n(\vec{X}, \Theta, \Omega_n)\big), \tag{2}$$

where $\mathbb{E}_\Theta$ is the regression expectation with respect to a random parameter which is conditional on the vector $\vec{X}$ and data set Ω_n.

A random forest is a collection of tree predictors, such as $\mathrm{RF}(\vec{x}, \vec{T}, \Omega_k)$, $k = 1, 2, \ldots, K$, where the Ω_k are random vectors. The random forest prediction for our regression problem is an unweighted average over the forest. The keys to convergence and superior metrics are low correlation and low bias of the model. Hence, in order to keep bias low the trees are grown to their maximum depth. At the same time, to keep correlation low when trees are grown we use randomization, such as each tree is grown on a bootstrap sample of the training set and the number of predictors in each specified tree is much smaller than the total number of total variables in the training set. At each node, variables are selected at random out of all variables, and the split is fitted as the best split on this subset of variables.

The choice of Leo Breiman's Random Forest algorithm [26] is justified because it yields one of the best performances among ensemble models and it is still very simple and not dependent on multiple hyperparameters optimization, which is a good way to demonstrate the properties of the functional relationships we are modeling. Random Forest approach is also known obtaining excellent performances in terms of accuracy and scaling up due to the ability of parallelizing tree growth, to the ability of handling thousands of variables, to the robustness for badly unbalanced data, and finally to the ability of providing internal unbiased estimates of the error as trees are added to the ensemble.

Specifically, Random Forest consists of a collection of randomized primary regression trees $r_n(\vec{x}, \Theta_m, \Omega_n)$, $m \geq 1$, where $\Theta_1, \Theta_2, \ldots$ are outputs of a randomizing variable Θ. These random trees are combined to form the aggregated regression estimate $\bar{r}(\vec{X}, \Omega_n)$. At each node, a coordinate of $\vec{X} = (X_1, \ldots, X_d)$ is selected, with the jth feature having a probability $p_{nj} \in (0, 1)$ of being selected. At each node, once the coordinate is selected, the split is done at the midpoint of the chosen side. The splits are traversed to the terminal node (leaf), minimizing the mean squared error.

Thus, our model regression estimate is:

$$\bar{r}(\vec{X}) = \mathbb{E}_\Theta\bigl(r_n(\vec{X}, \Theta)\bigr). \tag{3}$$

For each model we used a random selection of features to split each node, growing binary trees and averaging them [27], which has *computationally efficient* outstanding properties exploited by machine learning community. We also took advantage of the well-known performance improvements that are obtained by growing an ensemble of trees and averaging. Random vectors were sampled before the growth of each tree in the ensemble, and a random selection without replacement was performed [28].

4 Experimental results and discussion

The main outcome of our approach is an ensemble machine learning algorithm that predicts energy consumption as a non-linear time series regression problem on a daily scale for each electrical line id.

Several metrics to compare our models with baselines are provided in Tables 1 and 2. In particular, we report the Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Relative Squared Error (RSE), the Relative Absolute Error (RAE), and R^2.

Turning our attention to the MSE, the prediction performance for *daily average energy consumption* for the next 7 days prediction interval is 2.43 times better than the base-

Table 1 Mean daily consumption model metrics: Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Relative Squared Error (RSE), the Relative Absolute Error (RAE), R^2

Metric	Baseline	Model
MAE	20.8468	12.3683
RAE	98.9169	58.6869
MSE	790.6041	325.2679
RMSE	28.1177	18.0352
RSE	100.9551	41.5346
R^2	−0.0096	0.5847

Table 2 Peak daily consumption model metrics: Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), the Relative Squared Error (RSE), the Relative Absolute Error (RAE), R^2

Metric	Baseline	Model
MAE	186.1440	17.3112
RAE	621.7292	57.8201
MSE	36,062.7851	601.7531
RMSE	189.9020	24.5307
RSE	2,551.8602	42.5810
R^2	−24.5186	0.5742

line, MSE = 325.2679 compared to the baseline MSE = 790.6041 (training set arithmetic mean).

Interestingly, the prediction performance for *daily peak energy consumption* for the next 7 days prediction interval is 59.93 times better than the baseline (MSE = 601.7531 vs baseline MSE = 36,062.7851, which is the training set maximum value). The choice of this baseline is based on existing practice of energy companies to meet the maximum energy demand they experience in the past.

As shown in Tables 1 and 2, we got a negative R^2 metric for our non-linear regression problem baseline. Usually, R^2 is defined as the proportion of variance explained by the regression model fit. If the fit is actually worse than just fitting a horizontal line, then R^2 could be negative.

Tables 3 and 4 show the feature space used for the two final prediction models. The most powerful feature for both the prediction tasks is the number of consumers per electric power line - a feature from the energy consumption dataset. This feature provides a static characterization of a specific geographical area. Our machine learning algorithm uses this feature to build different energy consumption models for each range of consumers and power lines on each square of the partitioning grid. Then, the algorithm combines a number of these models into an ensemble model leveraging the decision tree regression properties. In sum, the number of consumers per power line is an efficient way to connect spatio-temporal human dynamics characteristics, detected by telecommunication data, with the static property of the geographic area.

As shown in Tables 3 and 4, the other relevant predictors describe human mobility patterns in a geographical space, which are found to be a good proxy for predicting daily electric energy consumption.

In sum, we found that together with static properties of the places, i.e. number of consumers per grid powerline, the spatial and temporal distribution properties of the mobile network data, such as Internet communications activity, are good predictors of near-term

Table 3 Mean daily energy consumption features

Rank	Feature	Decrease in MSE	Decrease in node impurity
1	Number of consumers per grid powerline	19.64	69,109.19
2	Variance of real numbers part of Fourier transform of area codes	4.43	6,534.18
3	Variance of real numbers part of Fourier transform of outgoing SMS activity	3.89	6,811.37
4	Variance of calling direction area codes	3.56	7,964.44
5	Variance of entropy of outgoing call activity in time domain	3.47	11,568.62
6	Entropy of first harmonic of outgoing calls	3.47	3,671.82
7	Entropy of Internet activity summed in time domain	3.21	2,750.14
8	Kurtosis of entropy distribution of outgoing calls	3.14	1,472.87
9	Variance of skewness of temporal distribution of outgoing calls	3.10	2,111.95
10	Entropy of fundamental frequency (first harmonic) of Internet activity	3.03	1,668.04
11	Standard deviation of frequencies distribution skewness of incoming SMS	3.03	2,848.24
12	Entropy of sum in time domain of outgoing calls	2.98	3,651.77
13	Variance of the kurtosis of outgoing calls in time domain	2.98	2,136.39
14	Kurtosis of time entropy of mobile Internet activity	2.96	4,409.30
15	Median of outgoing calls variance in frequency domain	2.92	1,317.35
16	Sum of 4 harmonic of incoming calls	2.88	5,035.26
17	Sum of outgoing calls skewness in frequency domain	2.88	825.13
18	Median of outgoing SMS temporal distribution kurtosis	2.88	1,260.23
19	Kurtosis of outgoing calls 32 harmonic	2.81	947.15
20	Median of 11 harmonic of Internet activity	2.72	185.14
21	Variance of 29 harmonic of outgoing SMS	2.71	1,243.93
22	Variance of outgoing SMS fundamental frequency	2.69	5,418.83
23	Sum of intertemporal mean of outgoing SMS	2.67	182.93
24	Sum of calling direction area codes	2.63	4,075.28
25	Kurtosis of skewness of Internet activity frequencies	2.62	2,290.58
26	Median of temporal distribution kurtosis of Internet activity	2.61	1,510.62
27	Sum of 5 harmonic of outgoing calls	2.60	4,752.84
28	Sum of incoming calls temporal entropy	2.59	2,494.76
29	Variance of Internet activity temporal entropy	2.58	3,569.25
30	Sum of 7 harmonic of outgoing calls	2.58	201.58
31	Skewness of 13 harmonic of incoming calls	2.55	751.57
32	Sum of 11 harmonic of incoming calls	2.55	1,431.18

For each variable from CDRs we computed the mathematical functions, which characterize the distributions and measure the information theoretic and statistical properties of such variables, e.g. mean, median, standard deviation, min and max values and Shannon entropy. Moreover, the same computations were repeated on sliding windows of variable length (1-hour, 4-hour and 1 day), producing second-order features that capture temporal relationships.

energy consumption. Also voice calls and SMS activity add some value to the prediction metrics; specifically, the spectral statistics of these activities and the cross-temporal entropy.

We also found that among second- and higher-order statistics, skewness and kurtosis of harmonics in frequency domain and entropy of cross-temporal communication patterns are the best predictors for maximum energy consumption prediction task, which is inline with the intuition of extreme value theory [29].

The full analysis of best predictors is provided in Table 3 for the *average energy consumption prediction* task, and in Table 4 for the *maximum energy consumption prediction* task.

For a commercial application it is possible to improve prediction metrics by creating a separate model for each power line, increasing the feature space during feature selection process and adding additional information to the feature space, such as the historical energy consumption properties and the weather forecast. These multimodal data sources are out of the scope of this research result, but in fact improve the model metrics. The

Table 4 Peak daily energy consumption features. Mean daily energy consumption features

Rank	Feature	Decrease in MSE	Decrease in node impurity
1	Number of consumers per grid powerline	21.22	132,035.53
2	Entropy of temporal sum of outgoing calls	5.86	24,013.99
3	Kurtosis of temporal entropy of Internet activity	4.33	15,479.71
4	Entropy outgoing calls fundamental frequency	4.10	14,650.13
5	Sum of 4 harmonic of incoming calls	3.96	11,449.25
6	Skewness of 5 harmonic of incoming calls	3.66	9,595.91
7	Skewness of temporal entropy of Internet activity	3.51	9,498.52
8	Sum of spectral distribution skewness of outgoing calls	3.47	2,108.19
9	Sum of temporal distribution kurtosis of Internet activity	3.35	2,267.17
10	Spatial variance of spectral variance of incoming SMS	3.30	3,971.29
11	Sum of 5 harmonic of incoming calls	3.27	9,468.86
12	Sum of calling direction area codes	3.25	6,815.90
13	Spectral variance of outgoing SMS activity total in time	3.20	13,899.48
14	Spatial skewness spectral skewness distribution of Internet activity	3.19	3,512.42
15	Kurtosis of temporal mean distribution of outgoing SMS	3.17	27,704.10
16	Spectral variance of temporal median distribution of outgoing SMS	3.02	15,366.03
17	Spatial median of incoming temporal sum	2.99	2,642.34
18	Spectral variance of outgoing SMS fundamental frequency	2.96	11,624.28
19	Skewness of incoming calls temporal entropy	2.92	4,174.09
20	Sum of outgoing calls 5 harmonic	2.89	10,251.36
21	Variance of Internet activity temporal entropy	2.86	5,382.14
22	Median of calling direction area codes	2.85	3,591.92
23	Sum of outgoing calls 4 harmonic	2.84	4,440.13
24	Median of incoming calls 16 harmonic	2.74	943.98
25	Spatial standard deviation of outgoing call temporal skewness	2.73	2,502.05
26	Spatial standard deviation of outgoing call temporal kurtosis	2.72	2,888.08
27	Skewness of incoming calls 4 harmonic	2.71	8,041.38
28	Spectral distribution skewness of mean Internet activity	2.69	1,759.88
29	Spatial standard deviation of temporal Internet activity entropy	2.68	4,233.89
30	Variance of outgoing calls temporal distribution kurtosis	2.67	3,283.40
31	Kurtosis of spectral distribution skewness.	2.67	5,160.36
32	Sum of Internet activity 4 harmonic	2.65	1,695.49

For each variable from CDRs we computed the mathematical functions, which characterize the distributions and measure the information theoretic and statistical properties of such variables, e.g. mean, median, standard deviation, min and max values and Shannon entropy. Moreover, the same computations were repeated on sliding windows of variable length (1-hour, 4-hour and 1 day), producing second-order features that capture temporal relationships.

cost of this improvement is an increase in computational complexity of each model, that could be efficiently parallelized in the cloud or by an efficient use of high performance computing (HPC) infrastructures, which usually exist in telecommunication and energy companies. All the computations we propose could be done in batch mode and do not require real-time processing.

5 Implications and limitations

Our results prove that human dynamics, which can be extracted from aggregated and anonymized mobile phone data, are good proxies for modeling energy consumption. This contribution has several practical implications for the energy producers and distributors, the telecom companies and, more in general, for the whole society. For example, our results could help to optimize the economy of energy producers/distributors value chain, also acting as an efficient tool for meeting peak electrical energy demands and creating a new market for telecom data usage. Again, our results could help to reduce the total primary energy consumption and thus its ecological footprint (e.g. climate change).

Among the limitations of our approach we consider that rural areas and areas, which are not equipped with telecom equipment or having small number of telecom activity, could not be used as proxies for energy consumption prediction in a powergrid. Also, our approach uses data from a single operator, Telecom Italia, and does not characterizes the households' activities. Finally, the introduced models do not account for seasonality on a yearly scale due to the 2-months limitation of our dataset. However, given the good horizontal scaling of the learning algorithm, this latter limitation could be solved by training the model on much more data.

6 Conclusion

Looking at the amount of electric current passing through a point in an electric circuit per unit of time and for each power line, we found that it has a number of cyclic characteristics and trends. We found predictable changes that repeat over daily and weekly periods. Based on these regularities we separated all power lines into 3 clustered areas: residential, touristic and city center/industrial areas. Then, we hypothesized and proved that cellular communication patterns, which represent human dynamics in space and time, could be a good proxy for energy consumption prediction. To this end, we computed, from the anonymized and aggregated mobile network activity, a number of predictors characterizing diversity, regularity and general mobile network activity in each part of the territory spatially aggregated by the square grid.

The prediction tasks, (i) predicting the *daily average energy consumption* and (ii) predicting the *peak daily energy consumption*, are solved for the next 7 days intervals for each electric line ID and are formulated as non-linear regression tasks. We used ensemble learning methods (Random Forest) to solve the optimization problem and avoid overfitting.

To solve the problem of computational complexity of the huge amount of data samples we moved from time domain to frequency domain. We also found that only a small set of harmonics in Fourier domain explains the response variable variance for each type of first-order feature space in time series, which reduces the computational complexity by a number of orders. The state-of-the-art feature selection pipeline that we apply, reduce the feature space down to 32 dimensions without losing significant accuracy.

The obtained results prove that human dynamics, extracted from aggregated and anonymized mobile phone data, are good proxies for modeling energy consumption. This contribution could help to optimize the economy of energy producers/distributors value chain and to reduce the total primary energy consumption, meeting the people's energy needs.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Authors equally designed and performed research and wrote the paper.

Author details
[1]University of Trento, via Sommarive, Trento, Italy. [2]Fondazione Bruno Kessler, via Sommarive 18, Trento, 24105, Italy. [3]SKIL, Telecom Italia, via Sommarive 18, Trento, 24105, Italy. [4]MIT Media Lab, 75 Amherst St, Cambridge, 02139, USA.

Acknowledgements
We thank Telecom Italia SpA and SET Distribuzione SpA for providing additional data and help in preparation of this research.

Endnote

[a] http://www.itu.int

Received: 15 November 2015 Accepted: 22 March 2016 Published online: 31 March 2016

References

1. Zhao H, Magoules F (2012) A review on the prediction of building energy consumption. Renew Sustain Energy Rev 16:3586-3592
2. Kolter ZJ, Ferreira J (2011) A large-scale study on predicting and contextualizing building energy usage. In: Proceedings of the conference on artificial intelligence (AAAI), special track on computational sustainability and AI
3. Martani C, Lee D, Robinson P, Britter R, Ratti C (2012) ENERNET: studying the dynamic relationship between building occupancy and energy consumption. Energy Build 47:584-591
4. Blondel V, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4:10
5. Gonzalez M, Hidalgo C, Barabasi A (2015) Understanding individual human mobility patterns. Nature 453:779-782
6. Song C, Qu Z, Blumm N, Barabasi AL (2010) Limits of predictability in human mobility. Science 327:1018-1021
7. Kung KS, Greco K, Sobolevsky S, Ratti C (2014) Exploring universal patterns in human home-work commuting from mobile phone data. PLoS ONE 9(6):e96180
8. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci 111(45):15888-15893
9. Batty M, Axhausen K, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, Ouzounis G, Portugali Y (2012) Smart cities of the future. Eur Phys J Spec Top 214(1):481-518
10. Keirstead J, Jennings M, Sivakumar A (2012) A review of urban energy system models: approaches, challenges, and opportunities. Renew Sustain Energy Rev 16:3847-3866
11. Martinez-Cesena EA, Mancarella P, Ndiaye M, Schläpfer M (2015) Using mobile phone data for electricity infrastructure planning. arXiv:1504.03899
12. de Montjoye Y, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD (2014) D4D-Senegal: the second mobile phone data for development challenge. arXiv:1407.4885
13. Barlacchi G, De Nadai M, Larcher R, Casella A, Chitic C, Torrisi G, Antonelli F, Vespignani A, Pentland A, Lepri B (2015) A multi-source datatset of urban life in the city of Milan and the Trentino province. Sci Data 2:150055
14. Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: a seasonal-trend decomposition procedure based on loess. J Off Stat 6(1):3-73
15. Stranneby D, Walker W (2004) Digital signal processing and applications. Elsevier, Amsterdam.
16. Krumme C, Llorente A, Cebrian M, Pentland A, Moro E (2013) The predictability of consumer visitation patterns. Sci Rep 3:1645
17. Singh VK, Freeman L, Lepri B, Pentland AS (2013) Predicting spending behavior using socio-mobile features. In: 2013 international conference on social computing (SocialCom). IEEE Comput. Soc., Los Alamitos, pp 174-179
18. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029-1031
19. Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, Pentland A (2014) Once upon a crime: towards crime prediction from demographics and mobile data. In: Proc. 16th ICMI. ACM, New York, pp 427-434
20. Bogomolov A, Lepri B, Staiano J, Letouze E, Oliver N, Pianesi F, Pentland A (2015) Moves on the street: classifying crime hotspots using aggregated anonymized data on people dynamics. Big Data 3(3):148-158
21. Montjoye Y, Quoidbach J, Robic F, Pentland A (2013) Predicting personality using novel mobile phone-based metrics. In: Greenberg AM, Kennedy WG, Bos ND (eds) Social computing, behavioral-cultural modeling and prediction. Lecture notes in computer science, vol 7812. Springer, Berlin, pp 48-55
22. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157-1182
23. Singh SR, Murthy HA, Gonsalves TA (2010) Feature selection for text classification based on Gini coefficient of inequality. J Mach Learn Res 10:76-85
24. Raileanu L, Stoffel K (2004) Theoretical comparison between the Gini index and information gain criteria. Ann Math Artif Intell 41(1):77-93
25. Tuv E, Borisov A, Runger G, Torkkola K (2009) Feature selection with ensembles, artificial variables, and redundancy elimination. J Mach Learn Res 10:1341-1366
26. Breiman L (2001) Random forests. Mach Learn 45(1):5-32
27. Breiman L (1999) Random forests-random features. Technical Report 567, Department of Statistics, UC Berkeley
28. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123-140
29. Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3(1):119-131