Big Data from Cellular Networks: How to Estimate Energy Demand at real-time

Conference Paper · October 2015

DOI: 10.1109/DSAA.2015.7344881

5 authors, including:

Davide Tosi
Università degli Studi dell'Insubria
65 PUBLICATIONS **390** CITATIONS
SEE PROFILE

Mario La Rosa
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Stefano Marzorati
Vodafone
7 PUBLICATIONS **33** CITATIONS
SEE PROFILE

Giovanna Dondossola
Ricerca Sistema Energetico
52 PUBLICATIONS **252** CITATIONS
SEE PROFILE

Big Data from Cellular Networks: How to Estimate Energy Demand at real-time

Davide Tosi
Dipartimento di Scienze Teoriche e Applicate
Università degli Studi dell'Insubria
Varese, Italy
davide.tosi@uninsubria.it

Stefano Marzorati, Mario La Rosa
Vodafone Italia spa
Milano, Italy
stefano.marzorati@vodafone.com
mario.la-rosa@consultant.vodafoneomnitel.it

Giovanna Dondossola, Roberta Terruggia
Ricerca Sistema Energetico RSE
Milano, Italy
[giovanna.dondossola, roberta.terruggia]@rse-web.it

Abstract— **Efficient energy planning is a key feature for the future smart cities. The real-time optimization of the energy distribution and storage is the real added value for smart grid and cities. However, the available energy providers' infrastructures are not able to estimate and predict real-time fluctuation of the energy demand and are not scalable enough to integrate, with low cost and effort, hardware elements able to estimate energy demand in real-time.**

The solution proposed in this paper exploit heterogeneous big data sources to forecast in real-time energy demands without requiring physical interventions on the energy providers' infrastructures. The proposed approach is mainly based on the use of statistical models and cellular network big data to estimate in advance energy demand without observing the actual behaviour of the energy network.

Distributor System Operators can use these estimations to self-manage the energy demand, distribution and storage in real-time, without any user intervention.

The approach has been extensively validated in a real world case study for the Milan city, in the production infrastructure of Vodafone Italy and with all the Vodafone Mobile Users, and the quality of the probabilistic models in forecasting energy consumption is really promising.

Keywords— Energy Forecast; Big Data Analysis; Regression Models; Cellular Network Data.

I. INTRODUCTION

Energy distribution, mobility and transportation optimization, water management, and other key services are at the basis for the future smart cities where efficiency, reuse and safety have a key role to make smart cities a reality. The energy service can have a direct impact all over these key services: a down-time of the energy service can potentially cease all other functions. Hence, energy infrastructures able to guarantee a high level of reliability are mandatory in the smart city arena. The real-time optimization of the energy distribution and storage is the real added value for smart grid and cities. This requires infrastructures able to self-manage their resources and to act proactively when energy demand changes over time, hourly or daily. However, traditional energy providers' infrastructures are not scalable enough to integrate, with low cost and effort, hardware elements able to estimate energy demand in real-time [10]. To avoid invasive and very expensive interventions on these infrastructures, mathematical and statistical models can be defined and adopted to estimate and predict in real-time the energy consumption of a location or city to automatically react when demand changes suddenly and to achieve a cost-effective efficiency. Currently, several models try to correlate energy consumption and demand with workload in server systems, ambient temperature or weather conditions [1][2][4][5][6][7][8][11]. However, none of these approaches take into consideration one of the most important factors that impacts energy consumption: the user behaviour. Hence, the approach proposed in this paper takes into account several heterogeneous variables and also includes the real-time distribution of people moving or stationing in target areas of a city. It is important to highlight that the proposed approach for energy estimation has been derived in the real production environment of the Vodafone Italy (VI) Mobile Operator and by collecting and elaborating real big data coming from thousand of real mobile users in a real big city such as Milan.

The approach has been set-up in the VI premises to gather real cellular network signalling events in order to understand whether a correlation exists also between cellular events and mobile users distribution (i.e., how people move around the city) with energy consumption. People movements are derived as Origin/Destination matrixes computed by analysing the cellular network signalling events. Statistical models based on regression functions are used to describe correlations and provide policy makers and energy optimizers with a tool to estimate and then plan energy consumption and demand just observing the cellular network events and how mobile users

move in the city. These models can also be exploited and integrated into the existing software systems of energy operators, with reduced costs and without infrastructural impacts, to self-manage the energy distribution and storage at run-time, without any user intervention.

From a conceptual point-of-view, the approach described in this paper is mainly based on top of two phases where statistical models are built from the observation of historical data, and where the significant models are implemented and used to estimate and predict energy consumption when exercised with real-time data.

Linear, logarithmic, and polynomial regression functions [3] have been adopted to find statistical significant models able to describe cell events versus energy consumption correlations. To this end, we used Vodafone network events and mobile users distribution as independent variable, while energy consumption as dependent variable as for the case of univariate models (i.e., models that describe the correlation just between one dependent and one independent variable.)

We also considered other independent variables such as heat indexes (as functions of air temperature and humidity), visibility, cloudiness and ultra-violet radiations index to understand whether multivariate models (i.e., models that describe the correlation between one dependent and several independent variables) can complement cellular network events data in providing better estimate of energy consumption.

The quality of the found model has been assessed by means of the cross-validation technique [9] to compute the absolute and relative error when estimating the energy consumption. We also applied the models to a new data set to estimate the energy consumption (for the two new days May 12 and May 13, 2015) and we compared the estimated values with the real values provided by the A2A energy distributor.

A real world case study with real energy consumption data has been adopted for deriving models and exercising them, and also for experimenting and validating the models. The models have been built over several hundred of data points and observations (confirming from a statistical point-of-view the validity of the models) and for the case of twelve energy stations (i.e., facilities for the generation of low/medium voltage electric power.)

Several relevant statistical models (both univariate and also multivariate) have been found thus suggesting that the observation of cellular network events can be used to estimate very well the energy consumption in real-time.

The validation phase confirmed also the quality of the estimations of the detected models.

Of course, this paper does not try to find a general model that describes in a general way the energy consumption all over the world, but it aims to define a new approach and an architecture based on big data to simply derive statistical models able to estimate energy demands.

The paper is structured as follows: Section II sketches the phases to identify and use statistical models to estimate data consumption. Section III describes the process and the results of the training activity we conducted to derive statistical models. Section IV summarizes the results of the validation phase of the models. We conclude and draw future work in Section V.

II. THE APPROACH IN GENERAL

This new passive approach (i.e., network signalling is silently collected from the VI cellular network) starts from the assumption that statistical models can describe the correlations between real energy demanding and consumption with cellular network events collected by dedicated probes installed over Telco operators.

Our approach works on the idea that it is possible to estimate in real-time how much energy has to be used in a certain area of a city starting from the calculating of how many people are dynamically located in the target area in a certain moment. We derive the knowledge of people behaviour (i.e., how they move, how they are distributed, how long they station in a target area) by aggregating in real-time data coming from the cellular network (i.e., mobile data, voice and SMS events.)

The approach is based on four main phases, starting from (1) the off-line detection of correlations and the definition of statistical models that describe the behaviour of the two sources of data: cellular network data collected by the VI probes, and energy demanding and consumption collected by other external sources such as feeds coming from energy providers, (2) the real-time collection of cellular network data, (3) the execution of the statistical models against the real-time data collected by the probes to estimate the current energy consumption, and finally (4) the computation and graphical representation of a set of indicators that describes the current energy demanding.

The graphical output of our approach is shown in Fig. 1. The representation shows a set of areas for Milan (to which a set of buildings are associated) with a 3D representation of the level of energy demand for each area of the city. The height of each coloured bar is proportional to the energy demand estimated by our approach. Of course, it is possible to have a clear estimation of the energy consumption (in Ampère) by clicking on each bar.

The personnel responsible for network management at Distribution System Operators (DSO) can exploit our approach and the graphical output of Fig. 1 to prepare the necessary arrangements in order to avoid excessive load (or even overload) conditions for some portions of the network. Today, no energy estimation systems are available to do that. A load estimation in real time as the one presented in the paper could be useful to the DSO to complete some sketchy information that already come from the field.

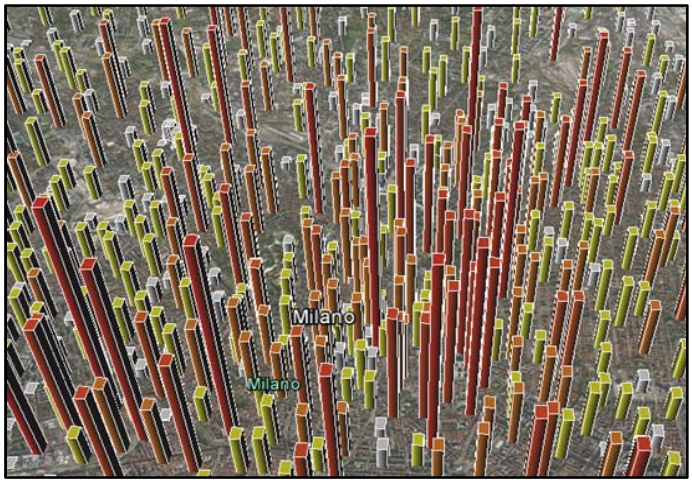

Fig. 1. Graphical output of the Energy Demand service.

It is important to note that from an infrastructural point-of-view, the cost for telco operators to setup and update their infrastructure in order to implement our approach is near to 0, because we draw our approach from the idea to collect and elaborate network events that are already available on network interfaces, which are normally used by Telco operators for network management purposes.

It is also important to highlight that the statistical models we discuss in this paper, their experimentation, and their validation have been conducted in the real production environment of VI and by collecting and elaborating real big data coming from thousand of real mobile users in a real big city such as Milan.

III. Deriving Statistical Models

Statistical models are generated off-line starting from the analysis of historical data coming from different sources to find potential correlations existing between these different sources of data. Hence, this section starts describing how these different sources of data are managed from an infrastructural point-of-view (for example, to filter and store big data coming from the VI cellular network) and then how they are statistically elaborated to find correlations models.

A. Data collection process

The first step to derive models able to estimate the energy demand forecast starts from the collection of all the data sources needed for the correlation analysis. Specifically:

1. The cellular network data to be considered as independent variable in the statistical analysis;
2. The energy consumption data to be considered as dependent variable in the statistical analysis;
3. Additional climatic data such as humidity, temperature, UV indexes to be considered as additional independent variable in the statistical analysis.

As already anticipated, the collection of these data is fundamental both to develop off-line the models but also to exercise on-line the detected models in order to preliminary assess the quality of the models and then to estimate on-line the energy demanding. In our infrastructure (see Fig. 2), a probe is able to monitor and capture the events exchanged between the mobile devices and the VI antennas inside target metropolitan areas (aka network cells). Every cell generates thousands of events per minute, collected by each probe from cellular network operator's equipment without affecting network performance. Once received, data are gathered in a proper server, in which they are elaborated, filtered and saved permanently in a dedicated DB.

The probe sniffs in real-time the events generated by the A Interface and IU-CS Interface of the 2G (GSM) and 3G (UMTS) cellular network of Vodafone on-top of the Base Transceiver Station (BTS) and Node B station, which communicate with the mobile handset and relays the information (voice, data, SMS, etc.) to the Base Station Controller (BSC) and the Radio Network Controller (RNC). Cellular conversations (voice, data and SMS) are managed and distributed by the Mobile Switching Center (MSC) throughout and outside the cellular operator's network. A dedicated server, named TAMS (Troubleshooting and Monitoring System) collects the data coming from the probe and assembles it in the correct format for transmitting it to the main infrastructure, where data are elaborated to derive the dynamic distribution of the SIM cards in the area monitored by the probes.

The events generated by the probe and stored in the DB are elaborated to filter all the cellular events that are not relevant for computing grid demand predictions or to focus the analysis on events correlated to particular type of users that can be significant for grid demand. For example, filtering events related to mobile users which are not directly generating a grid demand (i.e., people moving by foot or by bike) or focusing on events related to electric vehicles can provide an aggregated view of where and when these users are moving in order to predict the related localized grid demand.

In details, the event types that the probe is able to monitor are the following[1] (in parentheses, the id associated to each event):

- Unknown (0): the detected event does not fall into any of the admitted events;
- CM Service Request (1): the message originated by the mobile station to the network to request the assignment

[1] Those events are parameters generated from the network and described in the production of Technical Specifications for a 3rd Generation Mobile System based on the evolved GSM core networks, 3GPP is a collaboration between groups of telecommunications associations, known as the Organizational Partners.

of a service (such as a mobile originating call or a packet mode connection establishment);

- Common ID (2): the procedure to inform the RNC about the permanent Mobile Equipment Identity (i.e. IMSI) of a user;
- Paging Response (3): the Paging procedure is used to locate a mobile station to which a connection shall be established. Upon receipt of a Paging Request message the addressed mobile station initiates the immediate assignment procedure and responds to the network by sending a Paging Response;
- Location Updating Accept (4): the Location Updating procedure happens during the movement of the Mobile Equipment in order to enable the network to keep track of the subscriber. The procedure starts with a Location Updating Request originated by the Mobile Equipment asking the network to update its current location area. When the procedure complete, a Location Updating Accept is sent from the network to the ME;
- TMSI Reallocation Complete (5): the TMSI Reallocation procedure starts to provide identity confidentiality at least at each change of a location area. When the procedure complete, a TMSI Reallocation Complete message is sent by the network to the ME;
- HO Request (6): the Handover procedure starts when the ME is moving away from the area covered by one cell and entering the area covered by another cell in order to avoid call termination. In this case, the network is able to transfer an ongoing call or data session from one channel connected to the core network to another. The procedure starts with a HO Request, Relocation commands, a HO Complete and ends with a HO Performed message;
- HO Complete (7): see HO Request;
- Relocation Command (8): see HO Request;
- Location Updating Request (9): see Location Updating Accept;
- HO Performed (10): see HO Request;
- Location Report (11): the Location procedure is used to update the UMTS Cell ID from which the ME is consuming radio resources.

The events the probe is able to monitor are then processed and filtered by the TAMS server in order to provide files to the Infrastructure Traffic Sensors where data are elaborated to estimate SIM cards distribution. Each file contains one entry for event, described as below:

- Timestamp of the network event;
- SAC/CI: The Service Area Code and the Cell Identity are unique numbers used to identify each radio cell in a Location Area;
- LAC: The Location Area Code is the unique number used to identify a Location Area of the cellular radio network;
- O-IMSI, obfuscated IMSI derived by the IMSI through a cyphering algorithm in order to maintain the privacy of the customer related to the collected network event collected. The O-IMSI can be kept alive for a period configurable in order to be able to follow the movement of the SIM cards for the same period

in the form:

```
1297868695,3,51504,6360,3E6AF3303D865D23,1
```

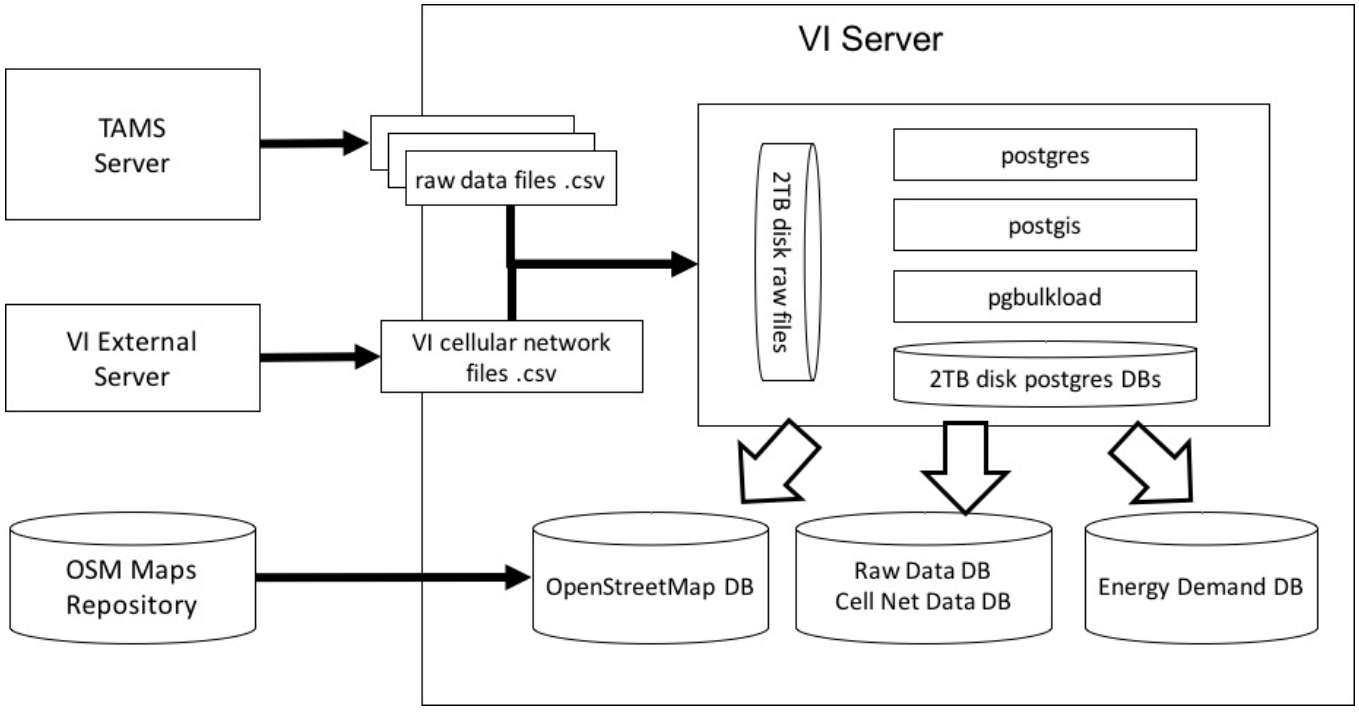

Fig. 2. VI Infrastructure for the Energy Demand estimation.

where the first element of each entry is the timestamp associated to the event, followed by the event types, the LAC, the SAC/CI, the IMSI (obfuscated by means of encryption algorithms) that generated/received the event, and the channel descriptor (UTRAN or GERAN). All these events and information are then elaborated and filtered by the TAMS server in order to provide .csv files to the Infrastructure Traffic Sensors (depicted in Fig. 2.) Each file contains one entry for each event monitored by the probe. The files are pushed to the Infrastructure Traffic Sensors every 5 seconds via a secure channel (SFTP). The data coming from the probe are stored in a proper DB (Energy Demand DB), populated with new data every 5 seconds. As depicted in Fig. 2, we can see the three input data flows:

- From TAMS server, .csv files containing network events are pushed into the VI server, stored in a 2TB archive, and uploaded to RAW_DATA DB via the *pgbulkload* Postgres utility
- From VI external server, .shp files containing the network cellular configuration are provided and uploaded to the PostgreSQL/PostGIS Cellular Net data DB via the *pgbulkload* and the *shp2pgsql* Postgres utilities;
- OpenStreetMap (OSM) geographic data extractions, including Milan map and road segments, loaded in the local Postgres/postgis OSM DB.

The EnergyDemand DB is located in the same PostgreSQL DB Cluster, on VI server.

OpenStreetMap data are needed to create graphical outputs, which are updated regularly in real-time every 15mins.)

Additional sources of data can be stored into the EnergyDemand DB. In effect, to perform a correct analysis of energy demand, the only data referred to cell net events and people density are not enough. As mentioned above, to understand how grid demand will change and in which direction, the observation of the actual trend of energy demand covers a pivotal role. For our analysis, we are using consumption data from A2A, one of the DSO (Distribution System Operator) for the city of Milan. These data concern a limited sector inside the central metropolitan area, split between ten sections covering a group of addresses. Every section is related to a certain substation, which provide energy for the buildings belonging to that section (see Fig. 3 for a geospatial representation of the considered stations for the Brera Milan area); For the rest of this paper we will refer to these buildings simply as Stations (St), distinguishing each one by a primal number. The DSO gives us the real amperage of each station every fifteen minutes for about two weeks for month, in order to show us a pretty wide look to their behaviour.

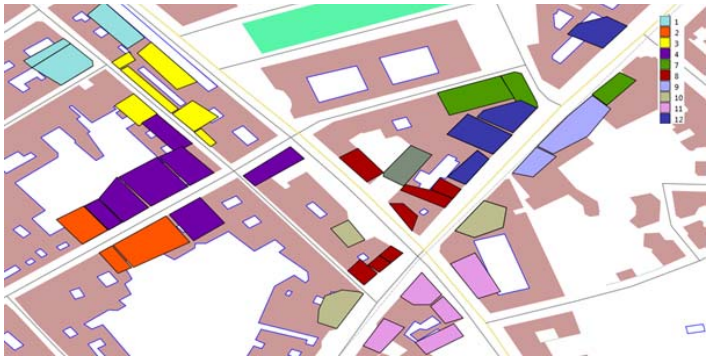

Fig. 3. Geospatial representation of the analysed Stations.

In addition to that, we consider necessary for an even more complete study to include also climatic information to better pinpoint the behalf of consumers according to the meteorological changes, to acquire a more precise estimate of their future request of energy. Power requests are influenced by the climatic situation around. Let's imagine a very cloudy day, warmed by strong hot weather, or a sunny week with each day characterized by low temperature all day long. The estimation of energy demand in these scenarios implies that our analysis must consider how many people will be forced to use more energy to warm (or cool down) house although the number of daylight hours or cloud coverage. We have therefore to collect a proper set of information regarding meteorological situation [13], to add them in the prediction model we'll describe below. We decided to introduce some supplemental parameters, such as temperature indexes, in particular heat index and *Summer simmer index* (a new index calculated as a function of air temperature and relative humidity; for Celsius degrees above 22 it pictures closer the complaint caused by heat because it points out the temperature it would feel like in a dry environment such as a desert), cloudiness and UV indexes. As we do for cellular events distribution, we can correlate these parameters with the power consumption, so that we can find the fittest to satisfy our analysis.

B. Data elaboration process

Let us assume to have two events, no matter in which cell (i.e., area of the city) they occur. An event, in our vision, is described as follow:

- IMSI (*Who* cause the event)
- Cell-id (*Where* the event occurs)
- Event timestamp (*When* the event occurs)

This triple identifies univocally a certain event, so we could pinpoint all the possible routes for a SIM card, sampling the data for every x seconds, with x configurable as small (or big) as we need for our purposes.

Sampling for the same IMSI and for a given ΔT period the consecutive events, it will provide the pattern of the location updated of the given IMSI for the observation period ΔT. This process has to be performed for every IMSI collected from the

mobile network. The resulting patterns are cumulated in a sort of 4D matrix, named Origin Destination Tables - OD table - where every element represents the number of IMSI that moved from a given Origin cell-id to a Destination cell-id at a certain time T for a certain duration ΔT.

In this paper we focus on what lie on the diagonal of this matrix, the events occurred where origin and destination cells correspond, that is the ones related to people remained in a particular place during a ΔT. Data related to these stationary events can tell us how many people in a particular area, therefore a density index to be applied to the city we decide to analyse. Moreover this can be further filtered focusing on particular SIM cards categories, such as the ones used by electric vehicles.

It has to be mentioned that all the location update have to be filtered from fake movements, such as for example when cell breathing happens. The size of the area covered by a 3G cell changes depending on the number of users attached to the cell. This change in size is called "breathing" because the size of the cell increases or decreases depending on the number of users, as consequence also neighbouring cells changes their geographical size. When a cell "breaths", some SIM change the cell it is attached to, if the position of the SIM is based on the location of the cell, this may result in an apparent change of the SIM location, which instead did not change its real position.

On the other side, we now have to include power consumption data in this schema, geolocating them inside the city map to match with cellular network cells.

Jointly with the DSO which provide us these data and RSE, the stations to analyse has been chosen in a limited central area of Milan, which peculiarities are common to many other zones, in order to simplify the analysis as well as to create a model useful to later describe the rest of the city.

Knowing that every station is related to a group of addresses, it is needed to picture them and understand how much of each cell they occupy. We know, for example, that the cell *a* houses a number *x* of events and cell *b* a number *y* of events, and the station *k* is constituted by three building, lying on *a* and *b* and occupying them respectively for z_a and z_b percent. Because in our environment the events in every cell appear as equally distributed, we can assume that in station *k* occurs the sum of the z_a percent of the events *x* and the z_b percent of the events *y*.

About the climatic parameters, our assumption is simpler: in this intermediate phase we consider the central area of Milan, whose dimension is not as wide as requiring more than one set of different meteorological indexes. So at a particular ΔT, the whole group of available stations in this study is affected by the same heat/UV index and it is covered by the same cloudy (or not) sky.

C. Building the statistical models

Starting from the aim of correlating the two main sources of data available (i.e., cellular network events versus energy consumptions), we tried to find statistical significant model to describe this correspondence using linear, logarithmic and polynomial regression functions. To this end, we used univariate models (i.e., models that describe the correlation just between one dependent and one independent variable) where: network events and mobile users distributions are considered as the independent variable of the model, while energy consumption data are considered as the dependent variable.

For the univariate regression analysis, we adopted the following formulae that can describe the correlation between independent and dependent variables *X* and *Y*, respectively:

(1) Linear model: $Y = m_1 x_1 + b$
(2) Logarithmic: $Y = m_1 \ln(x_1) + b$
(3) Polynomial: $Y = m_n x^n + m_{n-1} x^{n-1} + \cdots + m_1 x + b$

where *m* is the coefficient of the independent variable *x*, *b* is the

TABLE I. DETECTED MULTIVARIATE REGRESSION MODELS FOR NINE ENERGY STATIONS BELONGING TO GROUP 1 AND GROUP 2

Group1:	**St1**	**St2**	**St3**	**St8**	**St9**
	Coefficients	Coefficients	Coefficients	Coefficients	Coefficients
y-intercept	336.9169909	326.4879293	340.6674391	292.4642471	483.333417
X 1	46.52863743	37.87690275	17.4375361	25.5611268	20.21882225
X 2	4.015032979	28.46424478	13.64309507	13.9355972	24.19782028
X 3	29.23693952	23.88582124	54.20941153	57.24517803	28.33602001
R^2	0.743607857	0.929200733	0.918831094	0.912068209	0.841569396

Group2:	**St4**	**St7**	**St11**	**St12**
	Coefficients	Coefficients	Coefficients	Coefficients
y-intercept	86.36943549	-37.86112146	15.41968592	81.26177032
X 1	5.629285792	3.539577346	1.861652035	10.02821891
X 2	15.26579264	14.94382385	18.79633636	12.06504459
X 3	21.05009437	-4.338391889	15.83461798	56.45636621
R^2	0.84584261	0.697083754	0.878887613	0.866847761

y-intercept, and *n* the degree of the function.

In general, these univariate regression functions produce the slope of a line that best fits a single set of data. For instance, suppose you are interested in projecting the appropriate price for a house in a target area based on square footage. Using a linear regression formula, you can estimate a price, based on a database of information gathered from existing houses.

As for the case of the other additional independent variables such as heat indexes, visibility, cloudiness and ultra-violet radiations index, we used multivariate models (i.e., models that describe the correlation between one dependent and several independent variables) to understand whether these additional data can complement cellular network events in providing better estimate of energy consumption. Taking the previous example, the appropriate price for a house can be projected focusing not only on square footage, but also on number of bathrooms, age, etc. Using a multiple regression formula, you can describe relationships between the price and all the characteristics of the house.

For a multivariate regression analysis, we adopted the following formula:

(4) Linear multivariate model:

$$Y = m_1 x_1 + m_2 x_2 + \cdots + m_n x_n + b$$

where m_i is the i-th coefficient of the i-th independent variable x_i and b is the y-intercept.

Including climatic parameters entails changing a few the perspective. The additional variables we suggest are heat/SSI, UV and cloudiness indexes, but to avoid using parameters that apparently could satisfy a good R^2 without having an efficient estimation model, we decide to correlate each of these indexes with the consumption variables to find the fittest variables which help us to match closely to our objective.

D. Detected Models

The models have been derived starting from the collection and observation of a significant number (from a statistical point-of-view) of network events: We examined a group of nine stations (stations with missing primal numbers were discarded because they do not have complete data) for the days 9-10-11 and 16-17-18 June 2014 (146 data points in total). We chose this timeslot because it is equally divided in two weeks, the first one characterized by true "dog days" (29° C as average temperature, up to 34° C), the consecutive week with springy temperatures with about ten Celsius degrees of average temperature less than the previous one. In this way, we have in one month two different and relevant climatic examples. Of course, the presented methodology can be extended to different timeslots and stations to derive additional models that can better describe the real energy consumption. For example, we can straightforwardly compute a model for each season of the year to better approximate the reality.

As for univariate models, we detected several relevant models for almost all the considered variables.

Starting from the analysis of the cell net events as independent variable, Fig. 4 shows the dispersion of the energy consumption (x-axis) and the cellular networks events computed in hourly time-window in the area covered by the energy stations (y-axis). The figure shows also the univariate logarithmic models that better describe the correlation between dependent and independent variables and the R^2 value, which indicates how well data fit the statistical model. As depicted in Fig. 4, all stations are described by relevant statistical models with a R^2 higher than 0.5 (the threshold we set-up for models validity.) Consider that the average R^2 for a univariate model, which exploits just the cell net data as independent variable, is equal to 0.705 with an upper value of R^2=0.869 for the station St2 and a lower value of R^2=0.505 for station St7.

Considering one-by-one the climate parameters as single independent variable, we obtained significant models for the two variables: heat/SSI and UV index, with an average R^2=0.586 and R^2=0.655, respectively. As for the sky coverage variable, the average R^2 of the found models for the 9 evaluated stations is R^2=0.156 (lower than the threshold we set-up.)

Summarizing, we can conclude that univariate models that use only the cell net events as independent variable to estimate energy consumption are more precise (e.g., based on the R^2 analysis) than the models based on climate variables.

Fig. 4 suggests the identification of two groups of curves with a similar "electrical" behaviour: the first group (Group1) relates to stations St1, St2, St3, St8, St9 and the second one (Group2) to stations St4, St7, St11, St12.

As for the case of multivariate models, we used cell net events, heat/SSI data, and UV data as independent variables at the same time. Also in this case, we detected several relevant models from a statistical point-of-view. The Sky coverage variable has been discarded due to its low correlation with energy consumption as shown by its related univariate model. Table I summarizes all the significant models for each station belonging with all the coefficients of the independent variables (where X1 refers to cell net events, X2 to heat/SSI data, and X3 to UV data.)

With a multivariate regression on each station, the R^2 of the models associated to each station spans from R^2=0.70 to R^2=0.93 with an average of R^2=0.85. Two of the most representative regression functions, belonging to Group1 and Group2, have an upper value of R^2=0.93 in the case of St2 and R^2=0.88 in the case of St11, respectively.

The set of models reported in Table I can be then used at run-time to estimate in advance the energy consumption just starting from the observation of the three selected independent variables.

IV. VALIDATION OF THE MODELS

Among the different found models, we selected the multivariate ones with the highest R^2 for their implementation and validation. Specifically, we selected the two linear multivariate models reported in Table I for stations St2 and St11:

1. $Y_{St2} = 37.87X1+28.46X2+23.88X3+326.48$
2. $Y_{St11} = 1.86X1+18.79X2+15.83X3+15.41$

We validated the models with two experiments.

In the first one, we adopted the cross-validation technique to understand the quality of our estimations, and specifically the k-fold cross-validation technique [9]. In *k*-fold cross-validation, our original sample composed of 146 data points, is randomly partitioned into *k* equal size subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated *k* times (the *folds*), with each of the *k* subsamples used exactly once as the validation data. We created 10 folds (each with 14 randomly-selected data points), and we used the Standard Error (Std.Err) to evaluate the quality of each fold and the relative average quality of our energy consumption estimation. In this experiment, the model Y_{St2} has been used to estimate the energy consumption for all the stations belonging to Group1 (St1, St3, St8, St9), while model Y_{St11} to estimate the energy consumption for all the stations belonging to Group2 (St4, St7, St12).

In the second experiment, we applied the selected models to estimate the energy consumption to two new days (May 12 and May 13, 2015) and we compared the estimated values with the real values provided by the A2A energy distributor. It is important to notice that the two selected days have peculiarities that are slightly different from the days used to generate the models. For example, the average temperature registered in May 2015 was lower than the average temperature registered in June 2014 (i.e., 22.5C° vs. 26.0 C°). In this experiment, the multivariate models YSt2 and YSt11 derived for St2 and St11 have been applied to stations St1, St2, St4, St9, St11, and St12 without regrouping the stations into Group1 and Group2. As for St3, St7, and St8 the A2A energy distributor was not able to provide us with the energy data for the May 2015 period.

The cross-validation of the two multivariate models that refer to the aforementioned R^2, confirmed the quality of the models. The two models are able to estimate the energy consumption with a relative standard error (Rel.Std.Err) of 14.86% and 18.50%, for Group1 and Group2 stations, respectively. See Table II for details on the estimations for each station. We can conclude that since the Rel.Std.Err is lower than 25% [12], the two multivariate models can be considered reliable enough for their adoption.

The results of the second experiment were also promising (see Table III): The two models are able to estimate the energy consumption with an average Rel.Std.Err of 15.03% with a lower bound of Rel.Std.Err = 6.71% for St4. Hence, the models

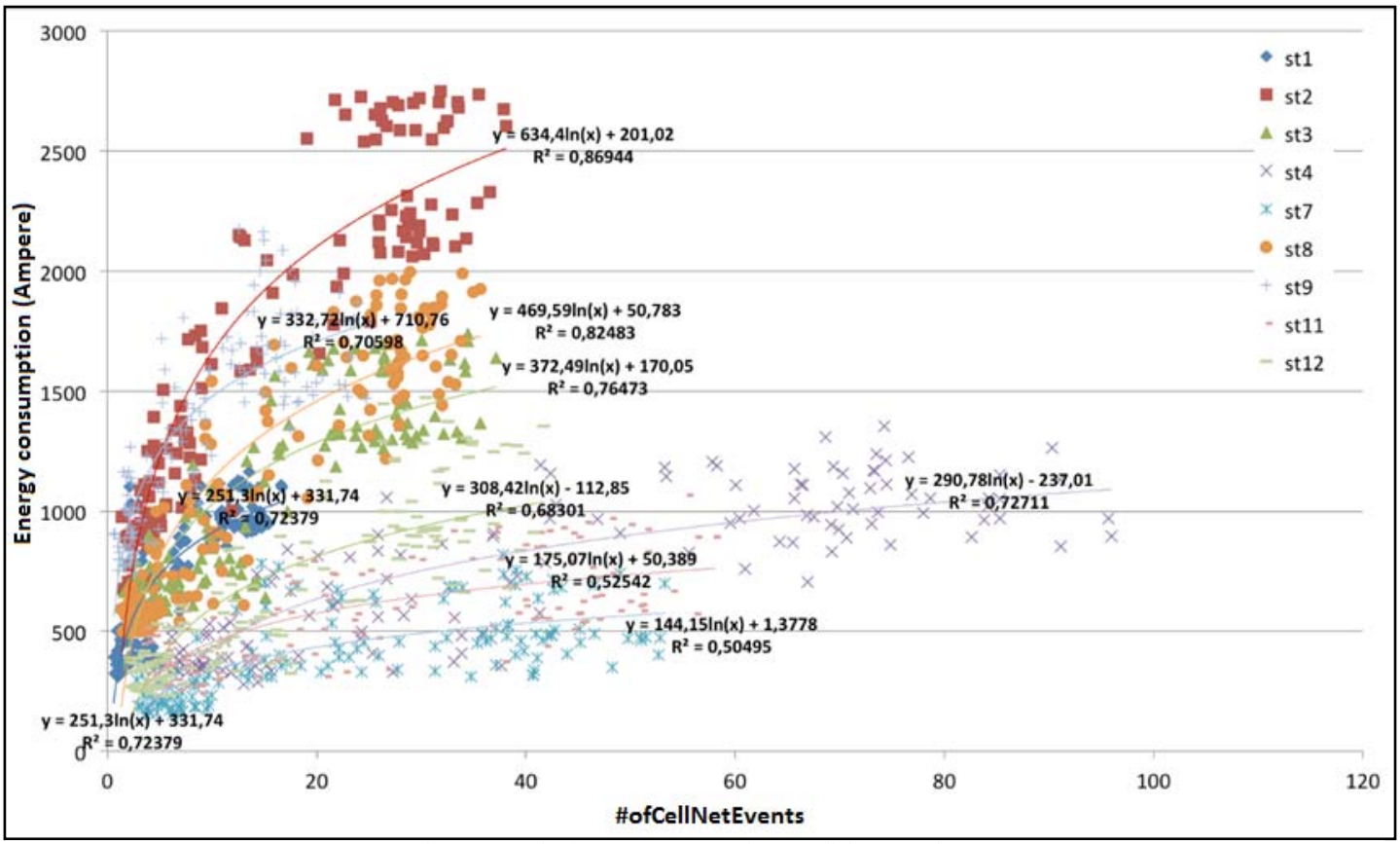

Fig. 4. Data Plot for Energy Consumption vs Cellular Network Events.

can be applied to days of the year with slightly differences.

We are conducting additional experiments to understand the applicability of the models also for periods of the year that are characterized by different temperatures and energy demand. In case the detected models are not accurate enough to describe the different periods of the year, we can simply generate new models by following our approach. For instance, we can imagine the derivation of models specifically targeted at each season of the year. To be more accurate, we can also generate models targeted at each month and then use at run-time the proper set of models with reference to the period of the year.

TABLE II. RESULTS OF THE CROSS-VALIDATION PHASE

Station Id	Real consumption (avg)	Estimated consumption (avg)	Std.Err	Rel.Std.Err (%)
1	708.34	660.04	110.36	15.58
3	1042.50	967.45	179.18	17.19
8	1131.50	972.80	163.67	14.47
9	1331.08	764.70	162.84	12.23
			Avg	**14.86**
4	728.45	586.65	98.89	13.58
7	406.40	555.78	109.82	27.02
12	665.98	545.96	99.37	14.92
			Avg	**18.50**

TABLE III. RESULTS OF THE SECOND VALIDATION EXPERIMENT

Station Id	Real consumption (avg)	Estimated consumption (avg)	Std.Err	Rel.Std.Err (%)
1	580.19	1090.51	212.20	36.57
2	1244.34	1244.34	150.78	12.11
4	477.87	271.83	32.09	6.71
9	1122.72	1191.48	217.27	19.35
11	270.72	270.72	22.41	8.27
12	512.13	270.49	36.77	7.18
			Avg	**15.03**

V. CONCLUSION AND FUTURE WORK

The adoption of the proposed approach can find its breathe in the process of estimating the energy demand. The models derived to forecast grid demand are relevant from a statistical point-of-view and can be adopted by energy providers to compute in real-time the energy consumption and demand in each area of the city. Of course, the quality of the found models can be improved by extending the training data set.

One of the further steps will focus in enclosing forecasting into different timeslots, in order to calibrate the time-window of estimation and adapt the analysis to meet correspondent requirements.

About demand forecasting, we studied an urban area where, theoretically, cell events and power consumptions are balanced. Our hypothesis is, in fact, that network event density in this zone is proportionally related to how much energy is used, because we talked about cell with high thickness of domestic buildings, which consumptions are adjusted for little amounts of people but deeply concentrated together. This hypothesis finds confirmation by the levels of statistical correlations discussed above. In addition, we overlooked the weight of the voltage phases, which contribution is to be later considered.

To enrich our analysis to a more thorough level, we should be able to forecast grid demand also in cases of non-urban areas, for example rural zones, where are located large industrial plants with very massive consumptions despite low people density, therefore less network events.

ACKNOWLEDGMENT

The research presented in this article was partially funded by the project SMARTC2NET [http://www.smartc2net.eu], sponsored by the EU in the 7th FP (grant agreement n. 318023) and by the project S-CASE [http://www.scasefp7.eu], funded by the EU in the 7th FP (grant agreement n. 610717). The authors thank A2A for the data provided.

REFERENCES

[1] A. Monticelli, Electric Power System State Estimation, Proceedings of the IEEE, vol. 88, no. 2, Feb. 2000.

[2] Geoffrey K.F. Tso, Kelvin K.W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy, Volume 32, Issue 9, Pages 1761-1768, ISSN 0360-5442. Sept. 2007

[3] MathWorld. Regression and Least Squares Fitting. Web published. Accessed: July 2014. URL: http://mathworld.wolfram.com/LeastSquaresFitting.html

[4] Coşkun Hamzaçebi, Forecasting of Turkey's net electricity energy consumption on sectoral bases, Energy Policy, Volume 35, Issue 3, Pages 2009-2016, ISSN 0301-4215. March 2007.

[5] Seligman, C., Kriss, M., Darley, J. M., Fazio, R. H., Becker, L. J., & Pryor, J. B.. Predicting Summer Energy Consumption from Homeowners' Attitudes1. *Journal of Applied Social Psychology*, *9*(1), 70-90. 1979.

[6] Saab, S., Badr, E., & Nasr, G. Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon. *Energy*, *26*(1), 1-14. 2001.

[7] Fumo, N., Mago, P., & Luck, R. Methodology to estimate building energy consumption using EnergyPlus Benchmark Models. *Energy and Buildings*, *42*(12), 2331-2337. 2010.

[8] Harris, C.; Cahill, V., "Exploiting user behaviour for context-aware power management," *Wireless And Mobile Computing, Networking And Communications (WiMob'2005), IEEE International Conference on* , vol.4, no., pp.122,130 Vol. 4, 22-24 Aug. 2005.

[9] S. Geisser. Predictive Inference. New York, NY: Chapman and Hall. ISBN 0-412-03471-9. 1993.

[10] Subhash, B.; Rajagopal, V., "Overview of smart metering system in Smart Grid scenario," *Power and Energy Systems Conference: Towards Sustainable Energy, 2014* , vol., no., pp.1,6, 13-15 March 2014.

[11] Seunghyun Park; Hanjoo Kim; Hichan Moon; Jun Heo; Sungroh Yoon, "Concurrent simulation platform for energy-aware smart metering systems," *Consumer Electronics, IEEE Transactions on* , vol.56, no.3, pp.1918,1926, Aug. 2010.

[12] Klein, RJ. "Healthy People 2010 criteria for data suppression". *Statistical Notes* (Hyattsville, MD: U.S. National Center for Health Statistics) (24).

[13] AccuWeather. Web published. Accessed: Sept. 2014. URL: http://www.accuweather.com/it/it/italy-weather