ENERGY CONSUMPTION AND DEMAND ESTIMATION FROM CELLULAR NETWORK DATA: A REAL WORLD CASE STUDY

Mario LA ROSA
Vodafone Omnitel – Italy
mario.la-rosa@consultant.vodafoneomnitel.it

Stefano MARZORATI
Vodafone Omnitel – Italy
stefano.marzorati@vodafone.com

Davide TOSI
Università dell'Insubria - Italy
davide.tosi@uninsubria.it

Giovanna DONDOSSOLA
RSE – Italy
giovanna.dondossola@rse-web.it

Roberta TERRUGGIA
RSE – Italy
roberta.terruggia@rse-web.it

Enrico FASCIOLO
A2A – Italy
enrico.fasciolo@a2a.eu

Stefano FRATTI
A2A – Italy
stefano.fratti@a2a.eu

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 predict real-time fluctuation of the energy demand, taking into account the new emerging energy demand behaviours in urban context with high density, fast moving people or new energy usages such as electric vehicles; moreover current energy forecast methods 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 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 probabilistic models and it exploits geolocalized cellular network traffic as independent variable to estimate energy demand without observing the actual behaviour of the energy network.

Probabilistic models have been derived by using the cellular network data coming from the mobile network production environment of Vodafone Italy and the A2A's grid in a real world case study in the city of Milan.

1. INTRODUCTION

Energy distribution, mobility and transportation optimization are at the basis for the future smart cities where efficiency, reuse and safety have a key role to make them a reality. In this context, the energy service can have a direct impact all over other services, because a down-time can potentially cease all other functions. The real-time optimization of the energy distribution and storage requires infrastructures able to self-manage their resources and to act proactively when energy demand changes over time. 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 interventions on these infrastructures, mathematical and statistical models can be defined and adopted to estimate and predict the energy consumption of a location or zone 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 historical data 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 takes into account several heterogeneous variables and also includes the real-time distribution of people moving or stationing in target areas of a city. The real-time distribution of people and citizens is estimated by observing and elaborating anonymous and aggregated flow of big data coming from the cellular network infrastructure of Vodafone Italy (VI).

The paper is structured as follows: Section 2 sketches the phases to identify and use statistical models to estimate consumption data. Section 3 describes the process and the results of the training activity we conducted to derive statistical models. We conclude and draw future work in Section 4.

2. APPROACH

This study starts from the idea that it is possible to estimate how much energy has to be used in a certain area starting off by calculating how many people are dynamically located in a 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 (e.g., mobile data, voice signalling, and SMS events.)

This approach is so based on four main phases, starting from (1) the off-line detection of correlations and the definition of statistical models from cellular network traffic, and power consumption collected by

measurements coming from the energy grid, (2) the real-time collection of geolocalized cellular network traffic, (3) the execution of the statistical models against the real-time cellular traffic 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 consumption.

3. DERIVING STATISTICAL MODELS

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 traffic 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.

3.1 Data collection process

In the network infrastructure, a probe monitors and captures events independently generated by the network, without any user intervention, inside the target metropolitan areas (aka network cells).
This probing infrastructure is primarily used for monitoring the network quality; hence this approach does not introduce further costs for the operator or energy provider.
The probing infrastructure and the events collection solution is primarily dedicated to the network quality monitoring and it can be exploited for several mobile analytics scenarios (e.g. urban traffic estimation and mobility patterns), hence it does not introduce dedicated costs to support the method described in this paper.
As shown in Figure 1 in a simplified schema, the probe sniffs in real-time the events (voice, data, SMS, etc.) generated by the A Interface and IU-CS Interface of the 2G (GSM) and 3G (UMTS) cellular network and relays the mentioned events 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 Centre (MSC) throughout 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 Infrastructure Traffic Sensors, where data are elaborated to derive the dynamic distribution of the SIM cards in the area monitored by the probes.

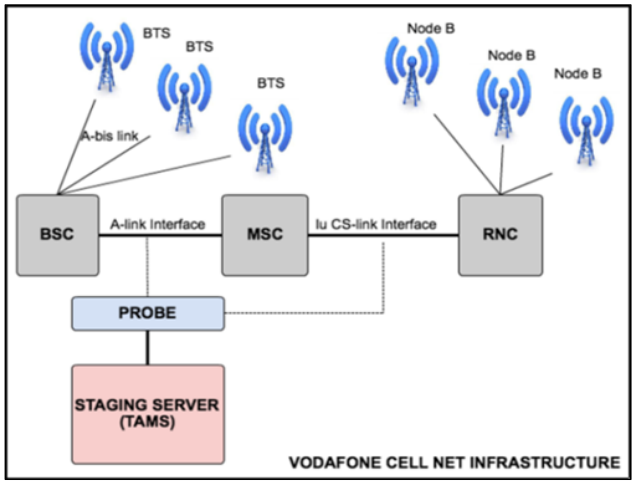

Figure 1 – Vodafone Italy Infrastructure overview

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, such as people still or moving by foot or by bike, or focusing on events related to electric vehicles, taking advantage of an aggregated view of where and when these users are moving in order to predict the related localized grid demand. Each file contains one entry for event described as follow:

- O-IMSI (*Who* cause the event)
- Cell-id (*Where* the event occurs)
- Timestamp (*When* the event occurs)

The O-IMSI, Obfuscated IMSI, is an identifier derived by the IMSI through a cyphering algorithm in order to maintain the privacy of the customer related to the collected network events. The triple (O-IMSI, Cell-id, Timestamp) 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.
Files are pushed to the Infrastructure Traffic Sensors via a secure channel (SFTP). Data we obtain from the probes described above are stored in a proper DB, populated with new data every 5 seconds.
On the other hand, to perform an exhaustive analysis of energy demand, the only data referred to event (and then people) density is 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, a main 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 power substation, which provides energy for the buildings belonging to that section; so, for the rest of this paper we will refer to these buildings simply as *substation*, distinguishing each one by a primal number. Sample DSO data gave us measure of the amperage of each substation every 15 minutes for about two weeks for 5 months, in order to show us a pretty wide look to their behaviour in the course of the year.

In addition to that, we consider necessary to 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.

3.2 Cellular network data elaboration process

Sampling for the same O-IMSI and for a given ΔT period the consecutive events, it will provide the pattern of the location updated of the given O-IMSI for the observation period ΔT. This process has to be performed for every O-IMSI collected from the mobile network. The resulting patterns are cumulated in a sort of 4D matrix, named *Origin Destination Table* - OD table - where every element represents the number of O-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.

It has to be mentioned that all the location updates have to be filtered from fake movements, such as for example when phenomena, such as cell "breathing", happens. The size of the area covered by a 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.

The substations to analyse has been chosen in a limited central area of Milan, whose 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 substation is related to a group of addresses, it is needed to identify 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 substation *k* is constituted by three building, lying on *a* and *b* and occupying them respectively for z_a and z_b percentage. Because in our environment the events in every cell appear as equally distributed, we can assume that in substation *k* occurs the sum of the z_a percentage of the events *x* and the z_b percentage of the events *y*. About the climatic parameters, our assumption is simpler: in this intermediate phase we still 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 substations in this study has the same heat/SSI, UV and cloudiness indexes.

3.3 Building the statistical models

Starting from the aim of correlating the two main sources of data available we tried to find a statistical significant model to describe this correspondence using linear, logarithmic and polynomial regression functions. To this end, we used univariate models where: network events and mobile users distributions are considered as the independent variable of the model, while energy consumption data are considered as dependent variable.

In general, univariate regression functions produce the slope of a line that best fits a single set of data. 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 to understand whether these additional data can complement cellular network events in providing better estimate of energy consumption.

Including climatic parameters entails changing the perspective a little. The additional variables we suggest are heat/SSI, UV and cloudiness indexes, but to avoid using parameters that apparently could satisfy a good coefficient of Determination, (named R^2, indicating how well data fit a statistical model) 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.

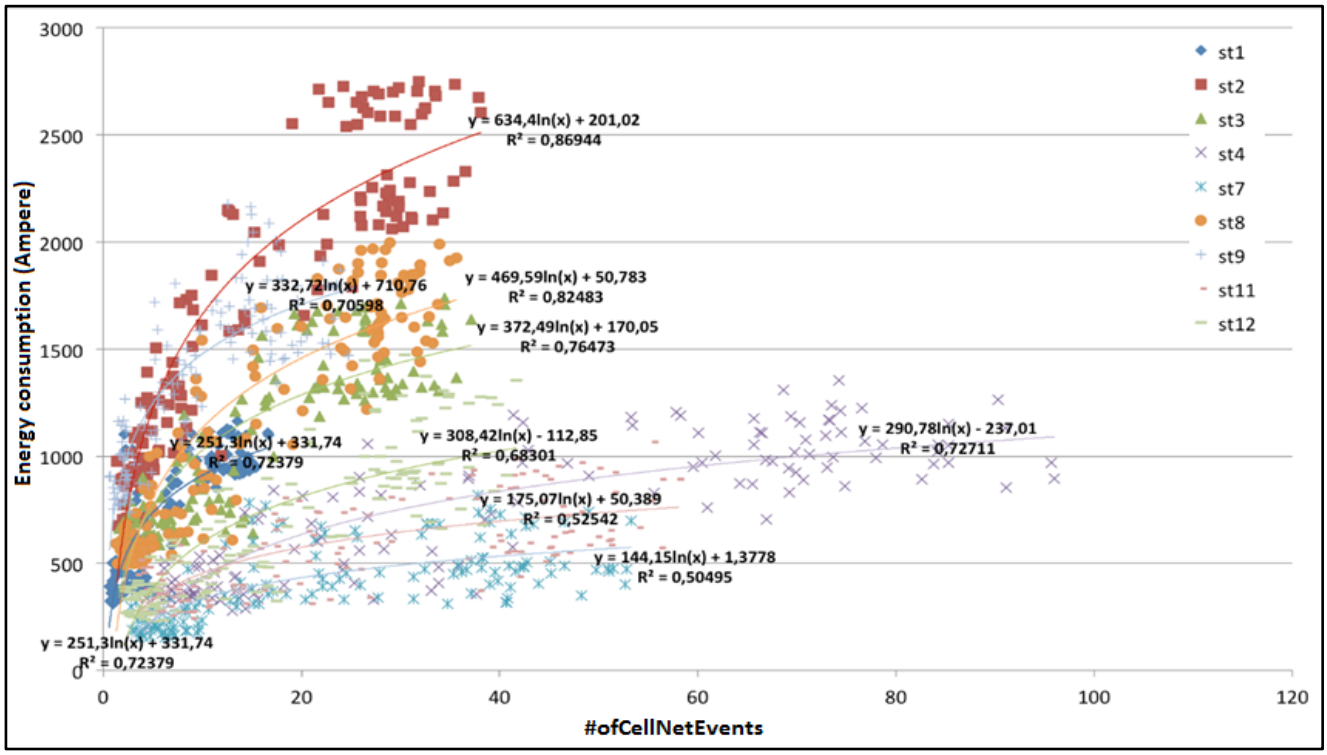

Figure 2 - Data Plot for Energy Current Consumption [A] vs. VI Cellular Network Events

3.4 Detected Models

We examined a group of 9 substations for 6 weeks from February to June, covering different seasonal conditions in terms of weather and people presence.

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, Figure 2 shows 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. All substations are described by relevant statistical models with a R^2 higher than 0.5 (the threshold we set-up for models validity.) The average R^2 for a univariate model, exploiting just the cellular network data as independent variable, is equal to 0.705 with an upper value of $R^2 = 0.869$ for the substation St2 and a lower value of $R^2 = 0.505$ for substation 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 substations is $R^2 = 0.156$ (lower than the threshold we set-up.)

Summarizing, the univariate models that use the cellular network 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.

As for the case of multivariate models, Figure 2 suggests the identification of two groups of curves with a similar "electrical" behaviour: the first group (Group1) relates to substations St1, St2, St3, St8, St9 and the second one (Group2) to substations St4, St7, St11, St12. Accordingly two models have been derived, with respectively a R^2 equal to 0.87 and 0.82 for Group1 and Group2.

The set of models 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. The execution of the statistical models against the real-time cellular traffic data showed a relative standard error equal to 16.4%. Since the Relative Standard Error is lower than 25% [12], the two multivariate models can be considered reliable enough for general adoption

4. 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 are relevant from a statistical point-of-view and prove that cellular network data are a strong indicator to forecast in real-time the energy consumption and demand in each area of the city. Of course, the quality of the models can be improved by introducing other variables affecting the energy demand and extending the training data set.

One of the further steps will focus in tuning the forecast into different timeslots, in order to calibrate the time-

window of estimation and adapt the analysis to meet correspondent requirements in terms of advance and duration validity of the demand forecast complementing the ones currently available to the Distributor System Operator that are not focused on short terms energy demand. 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, whose contribution is to be considered as future work.

To enrich our analysis to a more thorough level, the OD matrix we introduced in Section 3.2 can be further filtered focusing on particular SIM cards categories, such as the ones used by Electric Vehicles (EV), in order to foresee energy demand connected to EV stations.

Moreover, notable case studies to be investigated involve forecasting grid demand also in non-urban areas, for example rural zones, or where large industrial plants are located, with very massive consumptions despite low people density, therefore less network events.

ACKNOWLEDGMENTS

The research presented in this article was partially funded by the European project SMARTC2NET [http://www.smartc2net.eu], sponsored by the EU in the 7th FP (grant agreement n. 318023).

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 sectorial 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 Centre for Health Statistics) (24).

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