TLC POINTER – THE USE OF GEOSPATIAL DATA FOR NON-TECHNICAL LOSS DETECTION

Massimo ZERBI
Enel Global I&N NCO – ITALY
massimo.zerbi@enel.com

Paolo SANTI
MIT Senseable City Lab – USA
IIT-CNR – ITALY; *psanti@mit.edu*

Carlo RATTI
MIT Senseable City Lab – USA
ratti@mit.edu

Carlo PAPA
Enel Foundation – ITALY
carlo.papa@enel.com

Giuseppe MONTESANO
Enel Foundation – ITALY
giuseppe.montesano@enel.com

Domenico Tresoldi
Enel Global I&N NCO – ITALY
domenico.tresoldi@enel.com

ABSTRACT

Several machine-learning models specialized in providing revenue protection for power distribution companies can be found in the literature. However, traditional approaches present some limits: those models, relying solely on internal Company data, can be blind compared to some types of frequent anomalies, e.g., illegal connections, absence of consumption drops, etc. An innovative approach is the combination of proprietary data with third party data sets, preferably linked to geographical coordinates, thus representing a modeling of the territory. By integrating multiple sources of data, it is in principle possible to identify inconsistencies between activity patterns across data sets that would otherwise be impossible to identify by solely relying on proprietary data.
In this paper, we instantiate the idea sparked by Carlo Papa at Enel Foundation to combine fine-grained smart meter consumption data with cellular phone data records. The rationale for combining power consumption and cellular phone data is that both data sets have been proved to be good proxies of human activity. Hence, the identification of emergent activity patterns in the two data sets, and their spatio-temporal comparison, holds potential of substantially increasing the effectiveness of non-technical loss detection with respect to standard machine learning practices while opening the possibility to design new co-marketing & sales approach to mobile and electricity customers.

INTRODUCTION

The distribution of electricity implies a certain amount of loss of power that could be classified into technical and non-technical losses.

While technical losses in power systems occur due to energy dissipation in electrical system components such as lines, transformers, connections, measurement systems, etc., non-technical losses rise from the fact that not all of the energy delivered through the distribution network and consumed by end users can be measured or otherwise properly accounted for. NTL primarily relate to unidentified, misallocated, and inaccurate energy flows, either because the end user is unknown or the amount of energy being consumed is uncertain. NTL can also be considered as undetected load of customers unknown by the utility. The related increased losses will show on the utilities accounts, and the costs will be passed along to the customers as distribution charges according with regulation. [1]

The NTL are primarily caused by, but not limited to, the following:

- Meter tampering in order to record lower consumptions
- Bypassing meters by rigging lines from the power source
- Arranged false meter readings by bribing meter readers
- Faulty or broken meters
- Un-metered supply
- Technical and human errors in meter readings, data processing and billing

It is thus clear that while technical losses depend mostly from the grid components, non-technical losses and in particular frauds, depend mainly from variables related to the territory in which they take place, such as social, economic, geographical etc. variables.

The most effective way traditionally used in order to detect sources of NTL is to perform an inspection to the meter or the connection. Many methods for NTL identification have been studied in order to improve the detection of frauds or anomalies during inspections: the aim of these methods is to maximize the hit rate by identifying the premises in which the likelihood they could host a fraud or anomaly is the highest, thus increasing the energy recovery, and reducing the inspection associated costs.

NTL identification methods reported in the literature fall into two categories: deterministic filters (i.e. dedicated campaign) and machine learning techniques based on data mining on utility master data base. While deterministic filters imply handcrafted rules for decision making, machine learning gives computers the ability to learn from examples without being explicitly programmed. Historically, NTL detection systems were based on domain-specific rules. However, over the years, the field of artificial intelligence (AI) has become the predominant research direction of NTL detection. AI represents a

flexible and adaptable approach which is well covered in literature and allows to analyse customer profiles, their data and known irregular behaviour.

As said Machine Learning techniques applied to the utility internal master data base represent at the moment the cutting hedge solution in revenue protection issues. Nevertheless, AI based methods require high quality data in order to reach good levels of accuracy and reduce the number of false positives. At the same time it is fundamental that feedbacks on the inspections executed on field be reliable in order to avoid bias in the AI models. The main limitation that AI solutions are facing at the moment is essentially due to the fact of relying on internal company data to describe a phenomenon related, as mentioned, to the territory, i.e. to data for the most part external to the Company itself. In fact, there are several types of fraud that are hard to identify only by internal data, as the ones for which a consumption reference is not available or the ones where other variables such as supply events, alarms or patterns, are unknown or unmetered as, to give some examples:

- Direct connections to the network without any active supply contract
- Cases in which the manipulation takes place concurrently with the activation of the supply, or before the first register reading

In order to solve the aforementioned issues we have developed a vision: look at the real world, especially outside the domain of the Company, to find variables closely linked to the territory and to correlate them with the consumption of the customers and the position of the meters. The more correlated variables are placed in the model, the more the model is effective, on the very same philosophy on which features (i.e. criteria) are added in the Machine learning system described above. The ultimate goal is to build an inferential engine that uses geospatial layers. The first stage of development for this vision, however, is to determine the correlated variables, and below is described a methodological approach to determine one of these correlations: the one with the mobile TLC traffic.

It is important to point out that, before going into further details, a complete and integrated smart meter infrastructure is crucial to develop any advanced revenue protection system, in order to improve the process effectiveness with, for instance:

- Real time and simultaneous meter readings
- High availability of alarms and events
- High frequency in meters readings
- Reliable network topology information

RELATED WORK

Smart meter data, with as fine as 15*min* data consumption reading for each customer, is a very rich source of information for better understanding energy consumption patterns, and accordingly planning and operating the energy distribution network. However, the size of the data set, composed of thousands of yearly readings for each individual customer, requires the use of sophisticated data processing methods and machine learning algorithms to extract useful information.

A number of clustering methods have been recently proposed for application to smart meter data. The type of clustering approach used is mostly determined by the specific application, with typical applications being consumer profiling [2,5], energy demand prediction [3,4], and so on. An important aspect of the clustering process is related to how to deal with temporal characteristics of the energy profiles. For instance, if the goal is classifying users depending on the amount of energy used in weekend vs. weekdays, only information about the magnitude of power consumption in a large time window (e.g., a weekend) is relevant, and the clustering process is relatively more simple [5]. On the other hand, if the goal is, e.g., predicting energy demand at different times of the day, fine grained (say, hourly) temporal characteristic of the data set have to be included in the clustering process [4].

As mentioned in the Introduction, while smart meter data have been used to assist in NTL detection, even at the fine grain of spatio-temporal granularity this type of data by itself is not sufficient to enable effective detection. In fact, broadly speaking NTL detection would be empowered by a comparison between (*a*) expected energy activity in a certain area, and (*b*) the amount of energy use that is actually recorded by the smart meters. It is clear that smart meter data can be used only for (*b*), while (*a*) requires access to other sets that can be considered a proxy to human activity.

Among existing data sets, cellular phone Call Data Records (CDRs) have been extensively used to characterize human activities from different perspectives [6,7,8]. Compared to other data sets such as credit card transactions, vehicle GPS traces, and so on, CDRs have the advantage of (*i*) covering a vast majority of population, engaged in (*ii*) general, day-to-day activity. As for (*i*), we observe that the penetration rate of cellular phones approaches 100% of the active population, not only in developed country but increasingly so also in the developing world [6]. As for (*ii*), we observe that while other data sets record human traces related to a particular activity (e.g., buying goods with a credit card, or traveling on a GPS-equipped vehicle), CDRs records time and location of users engaged in general daily activities, and can thus be considered a better predictor of general human activity in an area with respect to other data sets. In particular, it has been extensively shown in the literature that CDRs allow a very accurate prediction of home and work location [6], which are the primary loci where NTL would take place.

To our best knowledge, the one reported herein is the first attempt to use smart meter data in combination with CDRs to improve the effectiveness of NTL detection.

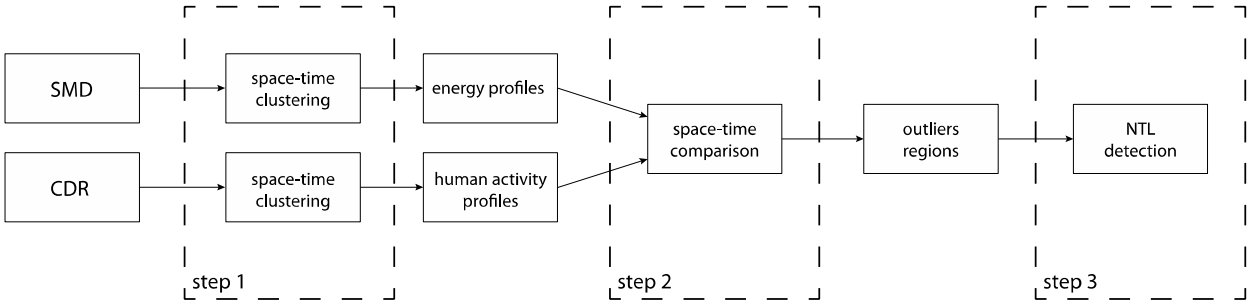

Figure 1. Methodology for NTL detection based on integration of SMD and CDR data.

PROPOSED METHODOLOGY

The main idea proposed in this paper is to combine smart meter data with CDR data to improve the effectiveness of NTL detection. The idea is motivated by the outlined limitations of NTL techniques based only on smart meter data, and by the fact that, among third party data sets, CDR have the best potential of providing a good estimation of human activity in urban environments.

The high-level description of the proposed methodology is reported in Figure 1. At step 1, spatio-temporal features of both smart meter data (SMD) and CDR are analyzed. It is important to observe that, in order to increase the potential of NTL detection, both spatial and temporal features of the recorded data should be exploited, which implies the adoption of clustering and machine learning methods for SMD that aims at detecting temporal consumption patterns, such as those reported in [4]. Similarly for CDR, where methods for extracting temporal features of human activity have been proposed, e.g., in [10].

The outcome of step 1 is a collection of spatially distributed energy profiles (from SMD) and human activity profiles (from CDR). The goal of step 2, which is the key step of the methodology, is establishing a framework for comparing the obtained profiles, with the goal of identifying *outliers*, i.e., regions of the city/area of interest where statistically significant deviations between the energy and human activity profiles are detected. The set of so identified outliners is the input to the next step of the methodology, where traditional energy loss enquiry methods (e.g., sending a crew on the field) are put in place in the identified regions to verify whether the detected discrepancy between energy and human activity behaviours are actually the result of NTLs.

Building energy and human activity profiles

The first step of the methodology consists in building profiles for both the energy and human activity data.

SMD is characterized by a temporal resolution as high as 15*min* intervals, which would allow building a very accurate temporal profile of energy use. On the other hand, for reasons related to the cost of communicating and storing data, SMD is often recorded with a lower temporal resolution. On the other hand, if the goal is detecting NTL, a minimal temporal resolution in the order of an hour is needed to enable the characterization of, e.g., daytime and night-time energy consumption patterns. In fact, it is well known that human activities have very distinct daytime and night-time profiles [10], which could be used for comparison with profiles extracted from SMD. The spatial granularity of SMD is, on the other hand, very accurate, going down to the level of the single household and/or commercial/industrial customer. However, privacy concerns, related to the fact that SMD might be shared with a third party in order to execute the methodology illustrated in Figure 1, might impose a certain level of spatial aggregation, thus reducing the spatial accuracy. This point is further elaborated in the next section.

CDR data typical comes in one of the following forms: *i*) call records; *ii*) network records; *iii*) location records. In case of *i*), a record is generated each time the user performs a phone-related activity such as incoming or outgoing call, send/receive text messages, use data connection, etc. The generated record typically contains user ID, user activity, time, and location record which is derived from the base station ID to which the user is currently connected. In case of *ii)*, a record is generated for each network connection event, such as connection to a new base station, periodic pinging of the current base station, etc. These events are routinely occurring in cellular networks, as they are needed to keep track of the current location of the user in the network. The generated record contains similar information as in *i*), where again the location information is derived from the ID of the base station to which the user is connected. Finally, case *iii*) refers to data generated when the GPS receiver on the phone is enabled, and location-based services are active. In this case, the record contains location information derived from the GPS receiver, expressed as (*lat*, *long*) coordinates.

The temporal granularity of CDR depends on the type of record stored. In case of type *i*), the recorded activity is typically bursty in nature, with spikes of intense activity followed by longer intervals of little or no activity. In case of *ii*), since what recorded is not directly related to phone usage, but more generally to the position of the phone

holder in space, we typically have a more regular and dense recording of user activity. Case *iii*) has typically the best time granularity as it is determined by the GPS receiver, that typically store records every second, or a few. In terms of spatial granularity, records of type *i*) and *ii*) have similar features, with an accuracy essentially depending on the density of base stations. An estimation of the achievable spatial accuracy can be obtained by building a Voronoi tessellation of the area of interest starting from the position of the base stations. Once the tessellation is built, the average area of the Voronoi cells gives a good understanding of the spatial accuracy that can be expected. Spatial accuracy in the order of few hundred meters is typically achieved in dense urban areas (where base station deployment is relatively denser), while an accuracy in the order of a kilometer is typical of sub-urban and rural areas. On the other hand, records of type *iii*) have much better spatial accuracy as enabled by the GPS receiver, which has a typical resolution in the order of few tens of meters.

While more desirable from both the temporal and spatial resolution viewpoint, records of type *iii*) require the user to activate the GPS receiver, which is only done by a minority of users, and often for a limited time. As a consequence of this, the user base coverage provided by type *iii*) records is typically at least an order of magnitude lower than what provided by type *i*) and *ii*) data.

Given the discussion above, the spatial and temporal granularity used to build energy and human activity profiles should be determined in a way that enables comparison of the different profiles. While both SMD and CDRs display similar temporal granularity with, e.g., the unit of one hour being very reasonable for both energy and human activity characterization, the choice of the spatial scale of analysis is less obvious. Generally speaking, the methodology reported in Figure 1 is designed to identify areas where NTL is likely to happen, to enable a narrowly focused use of more traditional NTL enquiry methods in those areas. The size of the area should be on one hand small enough to significantly scope down the search region for possible NTLs, while at the same time being large enough to enable the creation and the statistically significant comparison of the energy and human activity profiles. A reasonable choice for the spatial scale of analysis is the finer level of spatial aggregation possible for CDR data, which as discussed above can be roughly estimated as the average size of a Voronoi cell in the area of interest. So, detection areas in the order of few hundred meter squared (block level) in urban dense environments, or of one square kilometer (neighbourhood level) in sub-urban environments seem appropriate.

Comparing profiles

The second step of the methodology consists in performing a *statistically accurate* comparison of the energy and human activity profiles built in the previous step. As discussed above, energy and human activity profiles should be built using similar (ideally, the same) units for temporal and spatial analysis, so to ease a direct comparison of the two profiles. A crucial choice in this step is selecting features of the profiles that enable an effective identification of outliers. A promising candidate is looking at activity patterns occurring in what are the home/work locations of a user. The rationale is the NTL*s* are likely to occur at either residential or commercial locations, most likely overlapping with somebody's home or work location. Thus, by building human activity profiles by considering only records recorded at home or work location, it should be possible to remove noise and obtain a more accurate characterization of the human activity profiles in an area. Separate profiles should also be built for weekdays/weekend periods, as well as for night time/daytime.

Given the above discussion, the outcome of step 1 can be interpreted as giving in input to step 2, for each unit S_i of spatial analysis, a number of energy profiles $E_{i,1}$, …, $E_{i,k}$, with corresponding activity profiles $A_{i,1}$, ..,$A_{i,k}$. For instance, if we consider separate profiles for (*weekday*, *daytime*), (*weekday*, *nightime*), (*weekend*, *daytime*), (*weekend*, *nightime*), we have k=4, and four pairs of profiles can be compared for each spatial unit. Let us now assume a number synthetic metrics r_1, r_2, .. are derived from each profile, where r_j could be, e.g., the aggregate energy/activity value recorded in the profile, or the average/std deviation, and so on. For each candidate metric r_j, we select the one that displays the highest correlation between ($r_j(E_{i,h})$, $r_j(A_{i,h})$), i.e., between the same metric computed on the two corresponding profiles. The correlation is computed considering the entire area of analysis, i.e., considering all the spatial units S_i. The rationale here is that, for the proposed methodology to work, the metrics r_j extracted from the energy and human activity profiles *should be highly correlated*, so that the (relatively few) places where a high discrepancy is observed become candidate areas for NTL detection.

After the appropriate metric extracted from the profile is selected, by spanning the index i of all spatial units it is possible to build the distribution of the deviation $\Delta(r_j(E_{i,h}), r_j(A_{i,h}))$ observed between the metric for the corresponding profiles. Depending on the shape of the obtained distribution, the most appropriate outlier detection method can be selected, drawing upon the vast literature on this topic [10].

NTL detection

The final step of the methodology consists in using traditional NTL fraud detection techniques, but, instead of looking at the entire area of the city/region, looking only at the outlier spatial units obtained after step 2. By the very definition of outlier, those would account for only a very small fraction of the entire area under study, typically well below 5%. Thus, the potential of the methodology lies in the ability of significantly narrowing down the scope of application of traditional, expensive NTL detection techniques, with a corresponding reduction in cost. On the down side, NTL could occur also outside the outlier

regions identified by our proposed methodology. However, these losses are likely of lesser entity, or more spatially dispersed, than the ones occurring in the outlier regions, making the use of traditional NTL detection techniques in those area ineffective or too expensive.

FINAL CONSIDERATIONS

In this paper, we have proposed a novel methodology for NTL detection that leverage combination of energy and human activity data to significantly narrow down the area where traditional, and expensive, NTL enquiry methods are applied. The potential for cost reduction with respect to traditional methods is substantial, especially considering the fact that, by constantly monitoring energy and human activity profiles, it is in principle possible to have effective NTL detection without the need of continuously modifying the rules as in expert-system based methods.

However, in order for the proposed approach to become operational a number of technical and procedural challenges are still to be addressed. From a technical viewpoint, the proposed method builds upon the assumption that it is possible to build profiles and select metrics that display a very high correlation between energy use and human activity. While widely believed to be true, this correlation is yet to be demonstrated on real data sets. A challenge to achieve this validation is related to the difficulty of acquiring data sets from different sources that refer to the same region and time period.

From a procedural viewpoint, privacy regulations and internal data protection policies might hamper a single party to get access to both data sets. In particular, NTL should likely be done by the energy provider, which would then be in need of acquiring CDR data from a third party provider. The costs related to such acquisition may noticeably reduce the expected economic benefit of the proposed NTL detection methodology. However energy provider and mobile operators, based on the novel methodology we have defined, could find converging interest in new co-marketing & sales approach to their common customer base.

REFERENCES

[1] P. Deschamps, Y. Toravel, B.P. Swaminathan, A. Beuthel, R. Caire, D. Jeanneau, P. Mousinho, G. Pannunzio, N. Ruiz, M. Safanda, M. Zerbi, 2017, “Reduction of Technical and Non-Technical Losses in Distribution Networks", *Working Group on Losses Reduction CIRED WG CC-2015-2*, 18-23, 62-68.

[2] J. Kwac, J. Flora, R. Rajagopal, 2014, "Household energy consumption segmentation using hourly data", *IEEE Trans. on Smart Grid*, vol. 5, n. 1, 420-430.

[3] S. Haben, C. Singleton, P. Grindrod, 2016, "Analysis and clustering of residential customers energy behavioural demand using smart meter data", *IEEE Trans. on Smart Grid*, vol. 7, n. 1, 136-144.

[4] T. Teeraratkul, D. O’Neil, S. Lall, 2018, "Shape-based approach to household electric load curve clustering and prediction", *IEEE Trans. on Smart Grid*, vol. 9, n. 5, 5196-5206.

[5] T. Cerquitelli, G. Chicco, G. Di Corso, F. Ventura, G. Montesano, A. Del Pizzo, A. M. Gonzalez, E. M. Sobrino, 2018, "Discovering electricity consumption over time for residential consumers through cluster analysis", *Proceedings 14th DAS*, Suceava, Romania, doi: 10.1109/DAAS.2018.8396090.

[6] K. Kung, K. Greco, S. Sobolevsky, C. Ratti, 2014, “Exploring universal patterns in human home-work commuting from mobile phone data”, *PLOS One*, 9 (6): e96180

[7] M. Gonzalez, C. Hidalgo, A. Barabasi, 2008, “Understanding individual human mobility patterns”, Nature, vol. 453, n. 7196, pp. 779.

[8] L. Dong, S. Chen, Y. Chen, Z. Wu, C. Li, H. Wu, 2017, “Measuring economic activity in China with mobile big data”, *EPJ Data Science*, Vol. 6, n. 1, pp. 29.

[9] *https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/*

[10] S. Grauwin, S. Sobolevsky, S. Moritz, I. Godor, C. Ratti, 2014, “Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong”, *Computational Approaches for Urban Environments*, Springer, pp. 363-387.

[11] V.J. Hodge, J. Austin, 2004, “A survey of outlier detection methodologies”, *Artificial Intelligence Review*, Vol. 22, n. 2, pp. 85-126.