Smart Meter Data Analysis for Power Theft Detection

Daniel Nikolaev Nikovski[1], Zhenhua Wang[1], Alan Esenther[1], Hongbo Sun[1], Keisuke Sugiura[2], Toru Muso[2], and Kaoru Tsuru[2]

[1] Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA 02139, USA
[2] Mitsubishi Electric Corporation, 5-1-1, Ofuna, Kamakura, 247-8501, Japan

Abstract. We propose a method for power theft detection based on predictive models for technical losses in electrical distribution networks estimated entirely from data collected by smart meters in smart grids. Although the data sampling rate of smart meters is not sufficiently high to detect power theft with complete certainty, detection is still possible in a statistical decision theory sense, based on statistical models estimated from collected data sets. Even without detailed knowledge of the exact topology of the distribution network, it is possible to estimate a statistical model of the technical losses that allows indirect estimation of the non-technical losses (power theft) with high accuracy.

1 Introduction

The power grids of many countries are currently undergoing radical upgrades, and are increasingly equipped with massive sensing and communication infrastructure that can significantly improve the measurement and control capabilities of the resulting "smart" grids. This infrastructure includes devices such as phasor measurement units (PMUs) and smart power meters that are installed at many locations and/or measure data very frequently, resulting in very high-volume data streams. Using these data streams for various decision problems has opened up new opportunities for the application of data analytical algorithms and techniques.

One such decision problem of critical importance to electrical power utilities is the reliable detection of power theft [1]. It exists in practically all countries, but in some markets, for example Southeast Asia, the amount of theft can even exceed 40% [2-3]. The most common type of power theft occurs when an illegal user draws power directly from the power lines, between the distribution transformer (DT) and any electrical power meter that can measure electricity consumption. In general, the mismatch between the total energy supplied by the distribution transformer and the sum of energy consumed by all legal paying end users (EU) can be detected, and the total amount of losses in the distribution sub-network can be estimated. However, this amount includes both technical losses that are inevitable during the normal operation of the power distribution system, and non-technical losses (theft). Technical losses comprise ohmic losses in the electrical lines due to the resistance of the lines, conversion losses in any intermediate devices, leaks due to imperfect isolation, etc. Because some of these components of the technical losses depend on the amount of power being

P. Perner (Ed.): MLDM 2013, LNAI 7988, pp. 379–389, 2013.

delivered to customers, and that amount varies significantly throughout the day, week, and year, it is generally difficult to decide what part of the total loss is technical, and what part might be due to theft.

It would be possible to calculate accurately the exact amount of technical losses if all the parameters of the distribution network were known, including its connection topology, order and attachment points of all users, the line resistances between the attachment points, as well as the instantaneous power consumption by every user at any moment in time. In practice, full knowledge of these parameters is not economically feasible - a power utility would normally know which user is served by which distribution transformer from its geographic coverage plan, but the connection order and exact line resistances would not normally be known. In addition, full knowledge of the power consumption by any user at any instant in time would only be possible by installing detailed measurement equipment, such as PMUs, that performs very frequent measurements (multiple times per second). However, such an installation would be prohibitively expensive, its cost far exceeding the cost of power theft. In practice, utilities collect only infrequent, average and/or aggregated measurements, usually over a fairly long period of time - one month for traditional power meters, and 30 to 60 minutes for the new generation of smart meters that have advanced telemetry functions. The most important measurement available to power utilities is the total amount of active power consumed by a user during the measurement period, because this value is the basis on which payment by the customer is determined. Additional variables provided by some meters, for example smart meters conforming to the ANSI C12.19 standard, comprise reactive power consumed by the user (important for billing of some industrial customers), instantaneous voltage and current at the beginning and end of the measurement interval, power quality information, etc.

One popular theft detection method, implemented in most meter data management systems (MDMS), is to estimate the amount of total losses as described above, by performing energy balance between the energy supplied by the DT and the energy consumed by all metered users, and calculate the loss rate as a percentage of the total amount provided. When this loss rate exceeds a specified threshold, e.g. 3%, theft can be suspected and investigated. A disadvantage of this method is that no distinction is made between technical and non-technical losses, so when technical losses are unusually high for perfectly good and legal reasons, for example very uneven power consumption, a power theft even can be detected erroneously.

The advent of smart meters, with their much more frequent sampling intervals, has made it possible to calculate loss rates at a much finer temporal scale, and possibly detect power theft events of much shorter duration that would otherwise get lost in the much larger monthly energy aggregates measured and reported by traditional power meters. However, if power theft is not an irregular, one-time event, but is systematic and follows similar consumption patterns as those of the legal electricity users, the finer temporal scale of analysis would not improve detection significantly, because it would be computing the same loss rate, only more frequently. So, if the same loss rate calculation method is applied on smart meter data, higher accuracy of detection could not be expected for the economically more significant case of long-term systematic power theft. In order to use more productively the much larger data sets produced by smart meters, more advanced detection algorithms are needed. Section 2 proposes one

such method, Section 3 describes some experimental results using a detailed network simulator, and Section 4 concludes and proposes direction for further improvement in the accuracy and reliability of the method.

2 Power Theft Detection Based on a Technical Loss Model

We consider the problem of estimating the non-technical losses (NTL) in a branch of a distribution network consisting of a distribution transformer (DT) connected to a sub-station by means of a feeder, and a number of users connected to the secondary side of the DT in some manner (Fig. 1). The total amount of losses L_k in such a network over a particular time interval k can easily be estimated by performing the energy balance between the energy $E_{DT,k}$ supplied by the DT during this time interval and the sum of the energy E_k^i consumed by each user i, as measured by each smart meter:

$$L_k = E_{DT,k} - \sum_{i=1}^{n} E_{i,k} \quad . \tag{1}$$

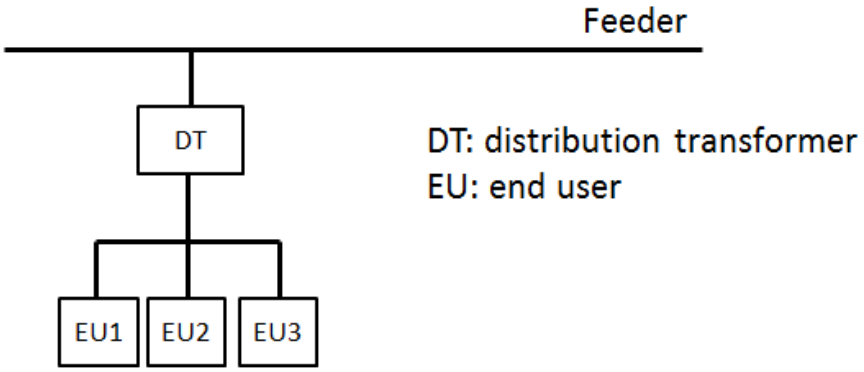

Fig. 1. Diagram of a distribution system. Power theft occurs when one of the end users (EU) is attached to the distribution transformer and draws power from it, but its consumption is not measured and paid for.

If we can estimate the amount $\hat{L}_{\mathrm{k}}^{\mathrm{TL}}$ of technical losses (TL) during the same interval, we can indirectly compute the amount $\hat{O}_k$ of NTL as $\hat{O}_k = L_{\mathrm{k}} - \hat{L}_{\mathrm{k}}^{\mathrm{TL}}$, and since NTL are usually attributed to power theft, a decision of whether to investigate can be based on this estimate. That is, our approach is to reduce the problem of estimating NTL to that of estimating TL.

One of the biggest problems in estimating the amount of TL $\hat{L}_{\mathrm{k}}^{\mathrm{TL}}$ is that, in general, the connection pattern between the end electricity users and the distribution transformer is not known. This connection pattern includes the topology and length of the power lines between the DT and EU. The most typical connection pattern is by means of a feeder connecting the secondary side of the transformer to individual users

attached at various points along the feeder. The exact location of these points is not known, and the resistances of the line between these points are not known, either.

In the absence of detailed information about the actual circuit of the branch of the distribution system, we make the simplifying assumption that it has a specific topology and connectivity pattern, shown in Fig. 2, represented as a one-line diagram [6]. We assume that each user (1101 to 1170) is connected to the secondary side of the transformer (the bus 1100) by means of an individual line. (When a user is connected to multiple phases of the distribution system, we treat each phase as a separate and independent user.) Because energy balance and loss modeling is performing independently for each phase in a multi-phase system, below we describe the model for a single phase only.

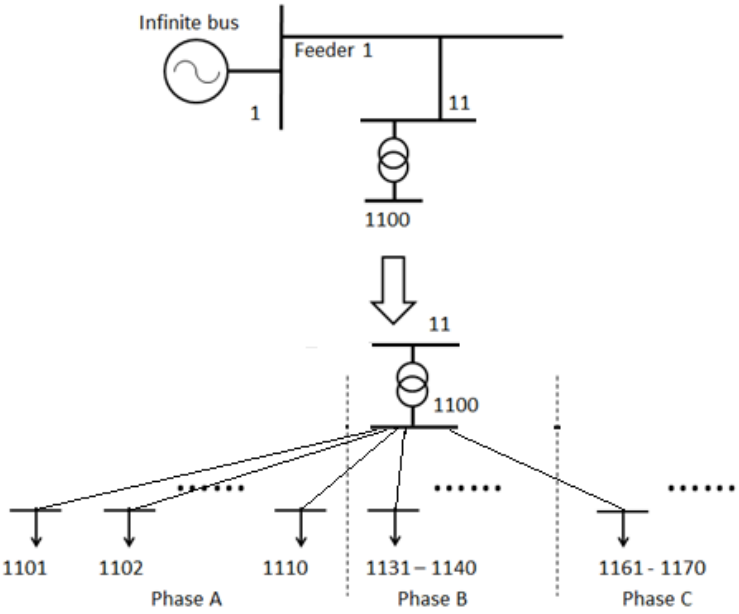

Fig. 2. Approximating circuit of a distribution system, represented as a one-line diagram. After the distribution transformer, all users are attached to the same bus 1100, through independent lines of varying resistance, one per user.

We define the following variables for the simplified circuit:

R_i The actual resistance of the line to user i

$\hat{R}_i$ The estimated resistance of the line to user i

$I_{i,k} = I_i(t_k)$ The measured instantaneous current of branch i at the end of time interval k

$L_{i,k}$ Actual technical loss of branch i during time interval k

$\hat{L}_{i,k}$ Estimated technical loss of branch i during time interval k

L_k Total loss during time interval k for all branches (users), obtained by means of power balance between the DT and all legal users

L_k^{TL} Technical loss during time interval k for all branches (users). When there is no theft, $L_k^{TL} = L_k$

l_0 Non-ohmic technical loss (time independent)

$\hat{O}_k$ The estimated non-technical loss (NTL) during time interval k

For smart meters, the measurement time interval is typically equal to 30 minutes, and for traditional meters, it could be equal to one month or longer.

The ohmic losses during measurement period k due to the resistance of the transmission line to user i are

$$L_{i,k} = \int_{t_{k-1}}^{t_k} I_i(t)^2 R_i dt \tag{2}$$

In practice, as noted above, we will know neither the actual resistance R_i of branch i, nor the instantaneous current $I_i(t)$ at all instants between times t_{k-1} and t_k. For this reason, we will make the additional simplifying assumption that the relation between current magnitude and time is piecewise linear:

$$\hat{I}_i(t) = s_{i,k}t + I_i(t_{k-1}), \tag{3}$$

where:

t time, $t_{k-1} < t < t_k$

$s_{i,k}$ slope, $s_{i,k} = \frac{I_i(t_k) - I_i(t_{k-1})}{t_k - t_{k-1}}$

We rewrite equation (2) as

$$\hat{L}_{i,k} = \frac{1}{s_{i,k}} \int_{I_i(t_{k-1})}^{I_i(t_k)} \hat{I}_i(t)^2 R_i dI_i(t) = \frac{R_i}{3s_{i,k}} [I_i(t_k)^3 - I_i(t_{k-1})^3] \tag{4}$$

The total loss then is:

$$\hat{L}_k = \sum_{i=1}^{n} \frac{I_i(t_k)^3 - I_i(t_{k-1})^3}{3s_{i,k}} R_i + l_0 \ , \tag{5}$$

where:

n The number of smart meters

The estimates of the branch resistances $\hat{R}_i$ can then be obtained by means of the least squares method, for example using Moore–Penrose pseudoinverse:

$$\hat{R} = (H^T H)^{-1} H^T L \ , \tag{6}$$

where:

$$H = \begin{bmatrix} \frac{I_1(t_2)^3 - I_1(t_1)^3}{3s_{1,2}} & \frac{I_2(t_2)^3 - I_2(t_1)^3}{3s_{2,2}} & \dots & \frac{I_n(t_2)^3 - I_n(t_1)^3}{3s_{9,2}} & 1 \\ \frac{I_1(t_3)^3 - I_1(t_2)^3}{3s_{1,3}} & \frac{I_2(t_3)^3 - I_2(t_2)^3}{3s_{2,3}} & \dots & \frac{I_n(t_3)^3 - I_n(t_2)^3}{3s_{9,3}} & 1 \\ \dots & \dots & \dots & \dots & 1 \\ \frac{I_1(t_m)^3 - I_1(t_{m-1})^3}{3s_{1,m}} & \frac{I_2(t_m)^3 - I_2(t_{m-1})^3}{3s_{2,m}} & \dots & \frac{I_n(t_m)^3 - I_n(t_{m-1})^3}{3s_{n,m}} & 1 \end{bmatrix}$$

$$L = \begin{bmatrix} L_2 \\ L_3 \\ \dots \\ L_m \end{bmatrix}$$

$$\hat{R} = \begin{bmatrix} \hat{R}_1 \\ \hat{R}_2 \\ \dots \\ \hat{R}_n \\ l_0 \end{bmatrix}$$

m the number of measurement periods

The free term l_0 represents the non-ohmic losses, that is, the losses that are not caused by and are proportional to line resistances.

In order to compute the least-squares (LS) estimate of the resistances, the system should be over-constrained, that is, the number of measurement periods m should be greater than the number of smart meters n, and the matrix H should have full rank $(n+1)$. This requirement is usually easy to satisfy.

Once the parameter vector $\hat{R}$ has been obtained, we can use it to compute the non-technical losses $\hat{O}_k$ for any future period k as

$$\hat{O}_k = L_k - \sum_{i=1}^{n} \frac{I_i(t_k)^3 - I_i(t_{k-1})^3}{3s} \hat{R}_i - l_0$$

3 Experimental Set-up

In order to verify the proposed algorithm, we conducted an experimental study in simulation, where a detailed simulator was used to calculate the state of a typical branch of a distribution network under typical loads and fairly frequently (every 10 seconds), and the resulting losses and power consumption measurements were accumulated over the much longer intervals (30 minutes) typical for the current generation of smart meters and automatic metering infrastructure. This procedure simulates the measurement process of a set of typical smart meters, and the aggregated data computed in this way was provided to the power theft analysis algorithm described above.

The test branch of the distribution system contained one user group of 30 users, such that 10 users were attached to each of the three phases. Without loss of generality, all consumption was assumed to be single-phase, and the analysis was performed on one phase only (phase A). One of the 10 users attached to phase A was assumed to be stealing power, resulting in power theft on the order of 10% of total consumption.

The time span of captured user data was 6 days (144 hours). Power theft occurred only during the last 2 days (48 hours). The first 2 days (48 hours) of data was used for estimating a predictive model for technical losses, as described above. Then, the last 4 days (96 hours) were used to verify the accuracy of prediction of technical (and, respectively, non-technical) losses. Of these 96 hours, the first half (48 hours) had no theft, and the last half (48 hours) had theft.

In order to compute the state of the network branch (represented by all voltages, currents, phase angles, and active and reactive power consumed at each node), power

flow calculation was executed every 10 seconds for the entire period of 144 hours. There were a total of 51,840 time periods for which power flow was executed.

The loading conditions for the network were specified by means of a time-varying demand profile for each user. Since actual demand profiles recorded by actual smart meters were not available, we generated them from a reasonable statistical model that assumed that the demand profile for each user had two components: a seasonal component that was the same for every user, and a random component that was different for every user. This is a reasonable model under the assumption that all users are of the same type (that is, all are commercial or all are residential), and their consumption patterns are similar, because they are driven largely by the same external factors (for example, by the air temperature in the same neighborhood that determines the demand for air conditioning services).

The seasonal component represents the average demand profile over an entire week. For our experiments, we used the actual total demand profile for the entire United Kingdom for six consecutive days in June 2012 [4]. Fig. 3 shows this seasonal profile. It is very smooth, because it is the sum of the demands of all consumers in an entire country (the UK).

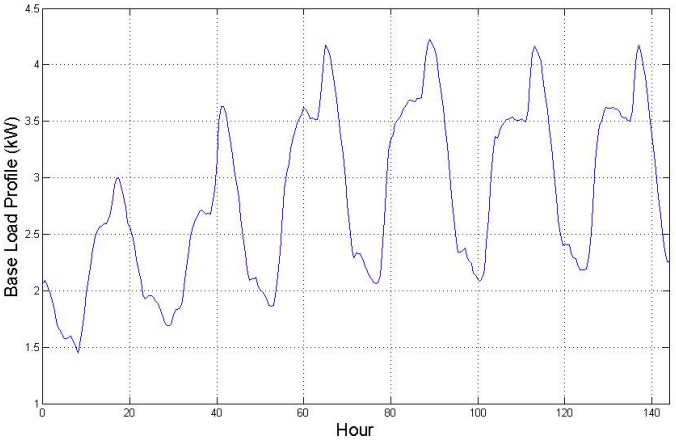

Fig. 3. Base load profile (P_{base})

The load profile for an individual user in our simulation was generated by adding a random component coming from the autoregressive (AR) process given by equation (7) to the seasonal base profile. Fig. 4 shows an example of the load profile for one user. In contrast to the seasonal component, individual load profiles are much noisier.

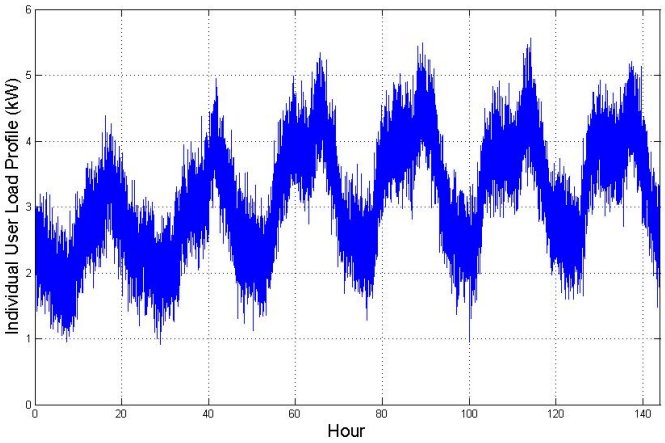

Fig. 4. Individual user load profile

$$P_i(t_k) = P_{base}(t_k) + 0.8 \cdot P_i(t_{k-1}) + 0.2 \cdot \mathcal{N}(0,1), \tag{7}$$

where:

$P_i(t)$	The load of user i at time t
$P_{base}(t)$	The base load at time t
$\mathcal{N}(0,1)$	Normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

Here, 0.8 is the autoregressive coefficient of the AR process, and 0.2 is the standard deviation of the white noise that is driving the process.

By using this stochastic process for user demand, we are ensuring that the users attached to the same transformer have similar, but not identical demands.

4 Experimental Results

4.1 Resistance Estimates

The predictive model for technical losses was estimated from the data for the first 48 hours (96 data points). The resulting estimates for the branch resistances are shown in Fig. 5. The agreement is reasonably good, and discrepancies are due to the relatively slow sampling rate of smart meters, the quadratic nature of losses, and the necessity to approximate the profile of the current during the sampling period (in our case, by a piece-wise linear curve).

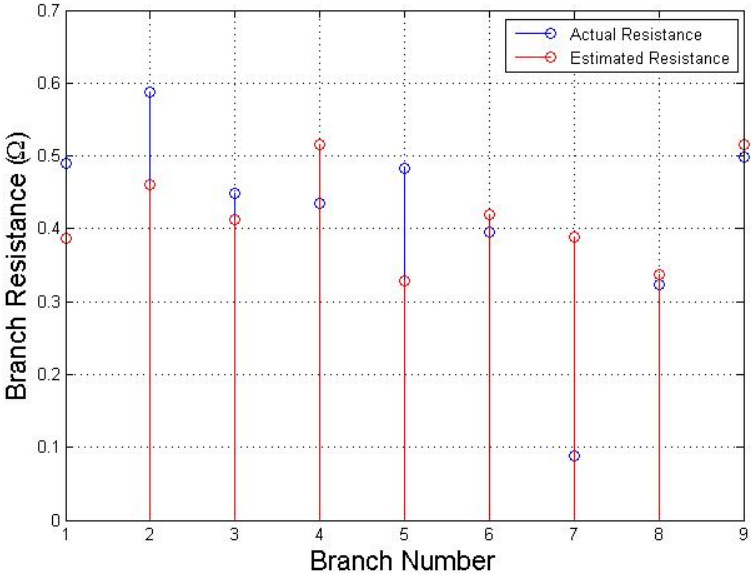

Fig. 5. Estimated and actual values for the nine branch resistances

4.2 Non-technical Loss Estimates

For every 30-minute measurement period during all three intervals of 2 days (48 hours, or 96 data points) each, we calculated the expected technical losses and subtracted them from the measured total losses to arrive at an estimate for the NTL during that period. Fig. 6 shows histograms of the NTL estimates for the three periods. The first histogram shows in red the NTL estimates from the first 48 hours (no theft). Since this dataset was used to estimate the line resistances, the shown values are in fact equal to the residuals from the least squares (LS) estimation. The implicit assumption behind the LS computation is that the residuals come from a normal (Gaussian) distribution with mean zero, and the histogram confirms that. This histogram also allows us to compute the expected variation of the NTL estimates under no-theft conditions, which we can use for determining confidence intervals and detection thresholds.

The second histogram shows in blue the NTL estimates from the second 48 hours (still no theft, but the line resistances used in producing these estimates were the ones obtained from the first data set). This histogram also shows the variation of the NTL estimates that can be expected normally, without theft, and is in agreement with the first histogram.

The third histogram shows in green the NTL estimates from the last 48 hours, when there is power theft. Visibly, the NTL values are larger than in the second case, when there was no theft. The two histograms (blue and green) do not overlap, so it is possible to completely separate the two cases, resulting in 100% accuracy of detection

for this level of theft (10%). For lower level of theft, though, the two histograms might overlap, in which case an optimal separation threshold must be determined.

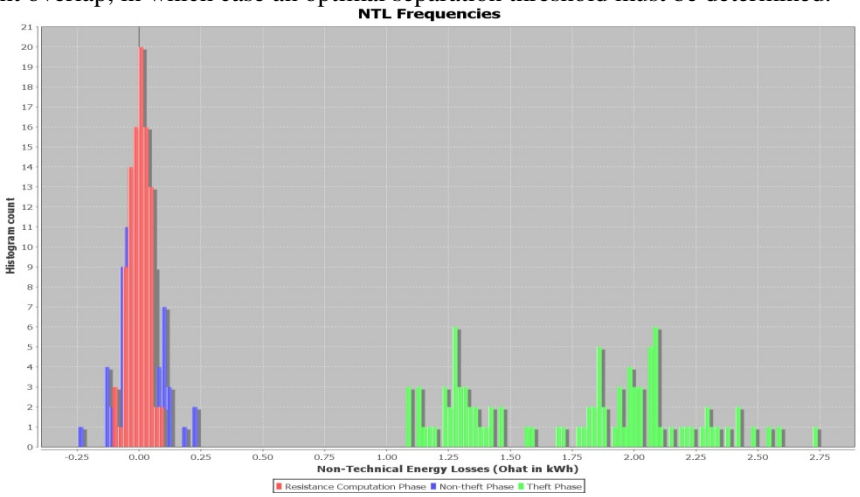

Fig. 6. Histograms for NTL estimates during three testing intervals (phases). In red, NTL estimates are shown for the interval from which the predictive model for technical losses was constructed. In blue, NTL estimates are shown for the second interval during which no theft was present. In green, NTL estimates are shown for the last interval during which theft did occur.

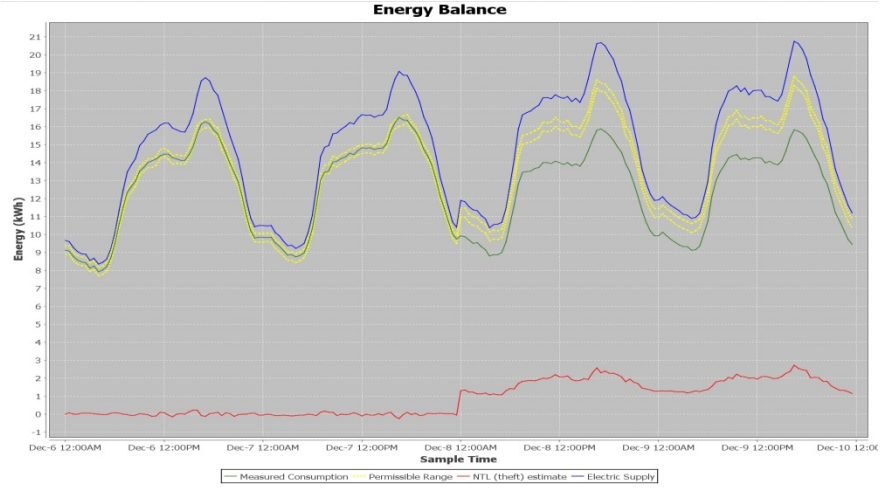

Fig. 7. A 95% confidence interval for expected total consumption (yellow lines) is computed from the measured total supply by the branch DT (blue line) and the predictive model for technical losses. When the measured total consumption (green line) goes outside of the confidence interval, theft can be suspected. The estimated amount of theft is also shown (red line).

The estimation method applied only to the last two intervals (two days of no theft followed by two days of theft) is shown in Fig. 7, as would be seen by power utility

employees during actual operation. The blue line shows the total amount of energy supplied to the group of users by the branch DT, and the green line shows the sum of the reported consumption amounts for all smart meters in the group. The two yellow lines represent a 95% confidence interval for the expected total consumption derived from the technical loss model, if no theft was happening. So, when the measured consumption (green line) is outside of the confidence interval (yellow lines), power theft can be identified. The estimated amount of theft is also shown by the red curve.

5 Conclusion

We have described a method for power theft detection based on a predictive model for technical losses in distribution networks equipped with smart meters. The predictive model is constructed entirely from data collected by smart meters. Since there are several significant sources of error and noise in the measurement process, such as infrequent measurements and unknown topology of the distribution circuit, our method relies on a statistical estimation procedure to fit a good model to the data. Experimental results in simulation showed that the resulting predictive model still allows for excellent separation between cases of theft and no theft, for power theft amounts on the order of 10% of power consumption (1 illegal unmetered user out of 10). In future work, we will further investigate the accuracy of the method for smaller amounts of theft, as well as expand the model to include other external variable factors, such as environmental temperature, rain, etc. We will also adapt it to the much more difficult case when not all users in a distribution network branch are equipped with smart meters, and some of them use the traditional kind of meters that provide power consumption readings aggregated over much longer periods (one month or more). We also plan to address other types of power theft, for example theft after the meter by a third party [3,5], again using a data analytical approach.

References

1. Depuru, S., Wang, L., Devabhaktuni, V.: Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. Energy Policy 39(2), 1007–1015 (2011)
2. Nagi, J., Yap, K.S., Nagi, F., Tiong, S.K., Koh, S.P., Ahmed, S.K.: NTL detection of electricity theft and abnormalities for large power consumers in TNB Malaysia. In: Proceedings of 2010 IEEE Student Conference on Research and Development (SCOReD 2010), Putrajaya, Malaysia, December 13-14 (2010)
3. ECI Telecom Ltd., Fighting Electricity Theft with Advanced Metering Infrastructure (March 2011) http://www.ecitele.com
4. National Grid UK, http://www.nationalgrid.com/uk/Electricity/Data/Demand+Data/
5. Nagi, J., Mohammad, A., Yap, K., Tiong, S., Ahmed, S.: Non-technical loss analysis for detection of electricity theft using support vector machines. In: Proc. 2nd IEEE Int. Power and Energy Conf., pp. 907–912 (2008)
6. McAvinew, T., Mulley, R.: Control System Documentation, ISA, p. 165 (2004)