Analytics and Optimization Techniques on Feeder Identification in Smart Grids

Larraitz Aranburu
Dep. de Matemáticas
UPV/EHU
Leioa (Bizkaia), Spain
larraitz.aranburu@ehu.eus

Aitziber Unzueta
Dep. de Matemática Aplicada
UPV/EHU
Blbao (Bizkaia), Spain
aitziber.unzueta@ehu.eus

M. Araceli Garín
Dep. de Métodos Cuantitativos
UPV/EHU
Blbao (Bizkaia), Spain
mariaaraceli.garin@ehu.eus

Juan I. Modroño
Dep. de Métodos Cuantitativos
UPV/EHU
Bilbao (Bizkaia), Spain
juan.modrono@ehu.eus

Aitor Amezua
ZIV Automation
Zamudio (Bizkaia), Spain
aitor.amezua@zivautomation.com

***Abstract*— One of the problems faced by electric power distribution system operators is to know with certainty the actual location of all their assets in order to manage properly the grid and provide the best service to their customers. In this work, we present a procedure for the identification of low voltage feeders or distribution lines in smart grids that is based on the mathematical formulation of the problem as an optimization model. In particular, we define the model with 0-1 variables (as many as meters to be identified in the different feeders) and with as many restrictions as the number of points in time that are considered. Given the large size of the problem in practice, the use of conventional optimization software becomes unfeasible. Based on this approach, and making use of the linear relaxation of the problem, some analytics over the coefficients (i.e., meter loads) and the special structure of the problem itself, we have developed an iterative procedure that allows us to recover the entire solution of the initial model in an efficient way. We have carried out a computational experience on a set of anonymized real data, obtaining results that support the efficiency of the proposed procedure.**

Keywords*— *connectivity model, integer optimization, iterative algorithm, linear relaxation, smart meter datasets.

I. Introduction

ELECTRICITY distribution system operators need to know with certainty the real location of all their assets to adequately operate the grid and provide the best possible quality of service. Particularly, revenue meters installed in clients' premises and therefore the contribution of each user with respect to the total load of each feeder or distribution line are valuable data. Mapping customers (using meters) to feeders and phases is important for several reasons, including load balancing among feeders and phases. The loads on the three phases of a transformer must be balanced for grids to be efficient, see [20]. By identifying which customer is on which phase, those can be then rearranged among the phases to balance the load.

Additionally, accurate information is needed to locate energy losses due to theft or unmetered locations in the distribution network, see [2] and [12]. Notwithstanding the aforementioned benefits, up-to-date connectivity information is required to correctly evaluate the impact of an eventual outage and to provide punctual information to customers.

[3] presents a phase identification system based on a unique signal injected into the phase line. The main criticism that has been made of these types of signal injection methods is that they require hardware equipment to receive and transmit signals to the different points of the grid, which significantly increases costs. There are also analytical techniques that make use of available data of smart meters, see [5], [1], and [19]. These works exploit the principle of conservation of the energy and use optimization approaches in a similar way to the method presented in this work. In [19], and as an extension of the two other works, the impact of correlations among customer loads on the performance of phase inference is added. In all of them, and using experiments, the same result introduced in [16] is shown, where it is proved that under certain conditions, the probability that a linear relaxation system of an integer one returns a unique integer solution is high. In particular, this probability increases when the number of measurements is at least twice as great as the number of customers. However, in some cases, this convergence can be very slow and it is necessary to increase significantly the number of time measurements, before obtaining a good approximation of the integer solution. This increase in the dimensions of the problem means that even using powerful optimization software, the time needed to reach a solution may grow exponentially.

[20] and [22] propose voltage-based techniques to infer customer phase. These methods rely on voltages measured and are based on linear regression and basic voltage drop relationships. However, not all utilities record customers' voltage measurements in short time intervals or even provide the kind of precise data needed to be able to use these models. Therefore, although these methods may be mathematically simpler, obtaining good results requires more demanding data collection and recording, as well as the use of geographical information systems.

In this work, we present a procedure for the feeders or distribution lines to customers mapping problem in smart grids that is based on the mathematical formulation of the

problem as an optimization model. It could easily be generalized to the phase identification problem. In particular, the model is defined with 0-1 variables (as many variables as meters to be identified throughout the different lines) and with as many restrictions as time measurements are available in addition to those requiring the identification of each meter on a single line. Given the large dimensions arising in virtually any real case, the use of conventional optimization software becomes unfeasible.

During the second half of the last century, a whole theory was developed around the reformulation of optimization models with 0-1 variables as linear models, due to the impossibility of solving them with conventional software tools. The idea was to try to extract as much information as possible from the structure of the problem, generating new and tightened constraints (cuts) that, added to the linear problem would achieve an equivalent but stronger reformulation of its feasible region (i.e., with the same 0-1 integer solutions). The integer solution would be obtained by solving such reformulated problem. Some references in relation to this theory are [8], [18], [10], [9], [14], [13] or [17], among many others.

Based on this approach, we develop an iterative procedure which, making use of the linear relaxation of the problem, allows to recover the integer solution in an efficient way.

The main contributions of this work can be enumerated as follows:

1) We propose a new scheme based on analytics and mathematical optimization for feeder mapping in distribution networks.

2) We develop an iterative solution procedure for grid mapping in both a noiseless and several noisy (energy losses, missing loads, incorrectly assigned measurements...) cases. This procedure requires, on the one hand, the linear relaxation of the original problem with a strategy of progressive incorporation of measurement constraints for different time periods. On the other hand,it includes a prioritisation strategy which is established to firstly identify the meters with the largest consumptions, taking into account the different solutions obtained in previous iterations, the analysis of the highest consumptions and, the progressive incorporation and implementation of a new set of valid inequalities (cuts) to the problem.

The rest of the paper is organized as follows. Section II presents the motivation, and the data sources, then introduces the mathematical formulation based on optimization, and finally introduces all the features that will be used in order to guarantee an effective resolution of the problem. Section III presents the empirical evaluation results relative to the noiseless and noisy variants. Section IV concludes and outlines future research plans.

II.DISTRIBUTION NETWORK TOPOLOGY

Electric power is generated at bulk generation power plants, utility-scale renewable generation plants and distributed energy resources. Electricity is transmitted mostly using high voltage three-phase alternating current towards electrical substations. Transformer substations transform high voltage into low voltage and, then, electricity is distributed in 3 phases, using 4 wires, usually labeled as R, S, T and N. Low voltage switchboards are used to split the output of the distribution transformer into a number of 3-phase, 4-wire feeders, usually between 4 and 8 feeders per transformer. Then, power is carried to customers' premises where load meters are installed at the connection point forming a distribution grid. Each feeder supplies power to an average of 50-200 customers, what means a range from 1 to 2,000 clients per substation, who are connected mostly to one single phase of the transformer (namely phase R, S, or T) and to the neutral conductor N. Customers with larger loads are connected simultaneously to the three phases and the neutral.

A. Data sources and information

In this work, we make use of real and synthetic smart meter datasets in order to develop an efficient procedure for feeder identification. We use a dataset of anonymized load measurements collected from customer meters connected to a set of feeders. Then, we have synthetized and simulated a network with a specific topology with the aim of evaluating the precision of the solutions obtained in terms of how exactly the pre-established network mapping becomes identified.

We have simulated a distribution network by assigning a set of actual meters to different artificial feeders or distribution lines in a pseudo random way. We have done this twice, generating two networks, one of small size with 3 lines, 14 meters (customers) and 6 hourly time measurements and a larger one, with 10 lines, 1578 meters (customers) and 215 hourly measurements. In the larger case, we have previously preproccessed the data set and cleared customers with zero or quasi-zero consumption, resulting in a reduction of the number of meters to 1351. Thus, we consider 10 power lines and 215consecutive time measurements, or points in time, forming 1351 time series of hourly frequency.

The statistical manipulation of the data has been carried out using the open source R statistical software, see [21].

B. Mathematical formulation

The mathematical modeling of the feeder identification problem requires the following sets, parameters and variables.

Let $T = \{1,\ldots,|T|\}$ be the set of time measurements; $I = \{1,\ldots,|I|\}$ the set of indices for customers or meters; and, $J = \{1,\ldots,|J|\}$, the set of indices for feeders or distribution lines to which the meters are connected. Let x_{ij} be the binary variable that shows if customer *i* is connected to line *j*.

$$x_{ij} = \begin{cases} 1 & \text{if customer } i \text{ is conected to feeder or line } j \\ 0 & \text{otherwise} \end{cases}$$

where $i \in I$ and $j \in J$. Let e_{jt} be the continuous variable that represents the errors or noise present in line *j* at time *t*. It is well known that in the distribution of power, part of it may be lost. e_{jt} arises due both to electricity losses and to certain unmetered loads on feeders such as street or traffic lights.

On the other hand, new technologies and renewable energies make possible the existence of consumers who, at any given time, can contribute with energy to the system. Therefore, we consider that the e_{jt} variable can be either positive or negative.

Let c_{it} denote the load (in KWH) of meter *i* at time *t*. Similarly, let T_{jt} denote the total measurement (possibly erroneous) of line *j* at time *t*.

The objective of this distribution line mapping is to obtain, using the optimization model below, the binary variable (x_{ij})

that assigns each customer to one of the distribution lines or feeders. Two types of constraints define the distribution network:

No consumer duplications: Each consumer is connected to a single line, so it must satisfy:

$$\sum_{j\in J} x_{ij} = 1 \quad \forall i \in I \tag{1}$$

Principle of conservation of electric power: It implies that the power supplied on each line should be approximately equal to the energy consumed by all customers connected to that line. Then, the following relationship must be satisfied:

$$\sum_{i\in I} c_{it} \cdot x_{ij} + e_{jt} = T_{jt} \quad \forall t \in T, j \in J \tag{2}$$

Here, e_{jt} compensates for the difference between the sum of the customer meter measurements and the measurement at the line they are connected to, which arise due to the errors defined above. This kind of constraints are usually known as a knapsack constraint.

Then, we can model the feeder identification problem as a mixed 0-1 optimization problem of minimizing the norm l_1 of the errors in the noisy case, or minimizing a given constant (i.e., without objective function) in the noiseless case, where the set of constraints are those given by (1) and (2). Thus, depending on whether or not we consider the errors (e_{jt}), we define two variants of the model: the noisy and the noiseless model.

The noiseless problem

In this case, the variable e_{jt} is zero and the proposed model is:

$$\begin{gathered} \min \sum_{i\in I}\sum_{j\in J} x_{ij} = \min A \\ \sum_{j\in J} x_{ij} = 1 \; \forall i \in I \\ \sum_{i\in I} c_{it} \cdot x_{ij} = T_{jt} \quad \forall j \in J, t \in T \\ x_{ij} \in \{0,1\} \quad \forall i \in I, j \in J \end{gathered} \tag{3}$$

Notice how in the objective function of (3) the sum in *I* and in *J* returns a constant, A.

The (noisy) problem with energy losses or inputs

In this case, the variable e_{jt} denotes the energy that can be lost or contributed to the line. This model is given by:

$$\begin{gathered} \min A + \sum_{t\in T}\sum_{j\in J} |e_{jt}| \\ \sum_{j\in J} x_{ij} = 1 \; \forall i \in I \\ \sum_{i\in I} c_{it} \cdot x_{ij} + e_{jt} = T_{jt} \quad \forall j \in J, t \in T \\ x_{ij} \in \{0,1\} \quad \forall i \in I, j \in J \\ e_{jt} \geq 0 \quad \forall t \in T, j \in J \end{gathered} \tag{4A}$$

Model (4A) is not a linear model as its objective function is not linear. However, we can easily make it linear by duplicating the e_{jt} variables using their positive and negative parts. That is, defining one variable for losses and another for inputs. In this way the variables e_{jt}^{+} and e_{jt}^{-}, which are both positive, are defined as:

$$e_{jt}^{+} = \begin{cases} e_{jt} \text{ if it is positive} \\ 0 \text{ otherwise} \end{cases} \qquad e_{jt}^{-} = \begin{cases} -e_{jt} \text{ if it is negative} \\ 0 \text{ otherwise} \end{cases}$$

Thus we build the equivalent model:

$$\begin{gathered} \min A + \sum_{t\in T}\sum_{j\in J}(e_{jt}^{+} + e_{jt}^{-}) \\ \sum_{j\in J} x_{ij} = 1 \; \forall i \in I \\ \sum_{i\in I} c_{it} \cdot x_{ij} + (e_{jt}^{+} - e_{jt}^{-}) = T_{jt} \quad \forall j \in J, t \in T \\ x_{ij} \in \{0,1\} \quad \forall i \in I, j \in J \\ e_{jt}^{+}, e_{jt}^{-} \geq 0 \quad \forall t \in T, j \in J \end{gathered} \tag{4B}$$

Both models, (3) and (4B), are mixed 0-1linear programs (in the first case with a zero objective function) and a system of constrained linear equations. Theoretically, they can have multiple solutions, especially when the number of constraints (i.e., measurements) is low (less than the number of variables). Although we could use any optimization solver to obtain the solution to these 0-1 problems, integer programs are NP-hard and therefore some instances may require a computation time that grows exponentially.

Indeed, this is what happens in the large case study, with 10 feeders, 1351 meters, and 215 hourly measurements.

Then, although the proposed models require a binary value for the assignment variables, in the next section we will see how the dimensions of these models, in practice with up to thousands of customers per feeder and perhaps thousands of feeders in the distribution network, make their resolution impossible at reasonable time. In this case, the linear relaxation can be used to retrieve the integer solution to the models (3) or (4B), given a sufficient number of constraints (i.e., time measurements).

In that sense in [15] and [16] it is proved that the probability that a linear relaxation system of an integer one of the form given in (3) returns a unique integer solution, increases and approaches to 1 when the relationship between the number of constraints (measurements, $m = |T|$) and the number of variables (customers, $n = |I|$), or m/n ratio, is greater than a given threshold.

With the idea of building an iterative process to find the entire solution through linear relaxation, it is remarkable that the set of constraints (1) is unique to the problem and does not introduce any additional information over the time horizon.

However, the set of constraints (2) allows for adding information to the problem iteratively. Even so, it is noteworthy that at each time point, we add as many constraints as power lines are considered in the problem, with a peculiarity: they all share the left hand side but a different right hand side, one per each feeder and time measurement.

In order to reinforce the contribution of each set of constraints of type (2) to the convergence of the problem, we propose the identification of a new set of valid inequalities (cuts). These, although redundant in the integer case, provide explicit relationships among variables that reinforce the feasible region of the linear relaxation without eliminating integer solutions.

It is known in the literature, see [4], [6] and [7], that such cuts help the convergence to the integer solution. A simple analysis shows that cuts of the form: $x_{ij} = 0$ can be identified from each constraint (2), whenever there are values of *i* such that $c_{it} > T_{jt} - e_{jt}$.

Moreover, a procedure for identifying other type of cuts that are implied by single 0-1 knapsack constraints (2) as given in [14] can be used.

The induced inequalities (covers) are satisfied by any 0-1 feasible solution to constraint (2), but are typically violated by fractional solutions. They can be written as

$$\sum_{i\in C} x_{ij} \leq k_C, \text{ where } 1 \leq k_C \leq |C| - 1 \qquad (5)$$

This constraint is implied by constraint (2) provided that $\sum_{i\in C} \boldsymbol{c_{it}} > \boldsymbol{T_{jt}} - \boldsymbol{e_{jt}}$. The set C is called a minimal cover implied by (2) if $\sum_{i\in C-\{l\}} \boldsymbol{c_{it}} \leq \boldsymbol{T_{jt}} - \boldsymbol{e_{jt}}$, $\forall \boldsymbol{l} \in \boldsymbol{C}$.

The inequality $\sum_{i\in C} x_{ij} \leq |C| - 1$, is a facet defining inequality for the convex hull of the 0-1 points satisfying (2) if and only if *C* is minimal.

One of the difficulties that occur in the implementation of this procedure of cut identification and addition is that the right-hand-side value of each constraint (2), $T_{jt} - e_{jt}$, is not known in advance. The gap e_{jt}, is the error variable which is known only after solving the corresponding linear model including that measurement constraint. Then, at each iteration of the procedure, and after adding a new set of measurement constraints, we must obtain the solution (error variables) in order to identify new cuts. Then, these new cuts are considered and can be used in an automatic reformulation to further tighten the linear relaxation.

In this way we are able to reduce the linear feasible region, without leaving out any integer feasible solution as shown in Figure 1:

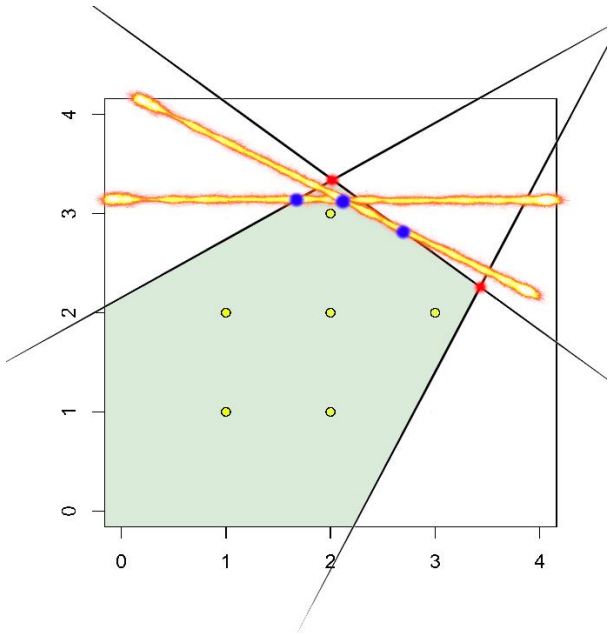

Fig. 1. Introduction of cuts

In Figure 1, the area in green delimitated by the black lines defines the linear feasible region. Therefore, the yellow points would be the integer feasible solutions, and the red points the linear optimal. Introducing the two straight flaming lines (cuts) reduces the feasible region for the linear model without eliminating integer solutions.

The following section summarizes the results obtained with these procedures by using a case study.

III. CASE STUDY

As mentioned above, in order to evaluate the efficiency of the identification procedure, we have simulated a distribution network assigning the available meters to a set of fictitious feeders. Moreover, the total consumptions of the feeders, T_{jt}, have been created for the noiseless and noisy variants of the problem.

As shown in [16], customer smart meter measurements exhibit time (over time) and spatial (between consumers) correlations. These correlations between customer load measurements tend to hinder the recovery of customer-to-line connectivity. When correlations are higher, a higher number of time measurements are needed to infer the line mapping accurately.

In the sample available, the time measurements to customer ratio (*m/n*) is small. In particular, the quotient is 215/1351 = 0.1591 < 0.5. In this situation, as indicated in [16], the probability of recovering the true network structure is very low, and the number of measurements are clearly insufficient to obtain an acceptable solution.

Based on these findings, we propose a procedure that helps to improve the probability of recovering the customer-to feeder mapping accurately with few time measurements.

The identification procedure proposed as solution is based on the linear relaxations of the mixed 0-1 models (3) and (4B), generating an iterative procedure in which at each iteration a set of constraints per unit of time is added to the model, saving the corresponding obtained solution. Thus, at each iteration the model increases the number of constraints, and does it up to a total number of constraints which in the last iteration is at most equal to the number of time measurements available. This iterative procedure doues not only add the knapsack constraint but also the cuts linked to it, that are identified by using the error terms obtained previously.

Moreover, analyzing these solutions, it is observed that there are variables that either take the value 1, or take repeatedly values near 1. In this case, setting the value of these variables to 1, (x_{ij}= 1), causes the identification of customer (meter) i in line j (that is, it is connected to line *j*). At the end of this iterative procedure, there may be variables fixed to 1 (that is, meters identified in lines) and unfixed variables that tell us that nothing can be concluded about their identification. Furthermore, each of these identifications may or may not be correct.

A. Small case study

Using the source data provided, a small case study is presented with 3 feeders, 14 meters and 6 hourly measurement constraints. The dimensions of the model (4B) would be the following:

- 42 x_{ij} (identifying) variables
- 18 e_{jt} (error) variables
- 14 type(1) constraints
- 18 type (2) constraints

Based on the values of c_{it}, we simulate new values for the parameters T_{jt} in order to model several situations in the electricity network. In these estimations, errors of 5 levels of magnitude are considered: no errors (noiseless case), and errors about 0.75%, 1.5%, 2.25% and 3% of the total line consumption respectively. As in this small case study there are only 6 time measurement constraints, the right hand side T_{jt} used for the cuts identification is this without considering noise.

The headings of Table 1 are interpreted as follows: #***ite*** denotes the number of iterations needed to achieve the best result; ***%correct(#)*** denotes the percentage of correct identifications, with the number of correctly identified customers in brackets; equivalently, ***%incorrect(#)*** denote the same for incorrect identifications; and #***cuts*** denote the number of cuts used.

TABLE 1: SMALL CASE

Case		***#ite***	***%correct(#)***	***%incorrect(#)***	***#cuts***
noiseless		4	100(14)	0(0)	0
		4	100(14)	0(0)	195
noisy	0.75%	6	85.7(12)	14.3(2)	0
		6	85.7(12)	14.3(2)	247
	1.5%	6	85.7(12)	14.3(2)	0
		6	85.7(12)	14.3(2)	247
	2.25%	6	71.4(10)	28.6(4)	0
		6	78.6(11)	14.3(2)	247
	3%	6	85.7(12)	14.3(2)	0
		6	85.7(12)	14.3(2)	247

The results shown in Table 1 summarize the appropriateness of the behavior of the proposed procedure. By introducing cuts and refining the iterative process, it is possible to rescue the exact solution in the noiseless case only in 4 iterations. In the noisy cases, between the 71% and the 85% of the feeders are identified. In this case, a better result with cuts, can be observed just in one of the cases, with more correct (and less incorrect) identifications.

B. Large case study

As stated in [16], the convergence of the iterative procedure is related to correlations (or, equivalently, sparsity) between the dataset variables, but above all it is necessary that the $m/n = |T|/|I|$ ratio reaches a certain threshold. In the largest case considered in this work with 10 feeders, 1351 meters and 215 hourly measurements, this ratio is 0.1591 which does not bode well.

In view of the robustness provided by the model cuts in the small case under study, we proceed to introduce a similar scheme in the large case. For this purpose, we have previously calculated the number of cuts induced by a single measurement constraint. In this case, a total of 253,571 cuts and more than 120 million of non-zero elements are generated per each knapsack constraint. This would imply, adding more than 108 million constraints and more than 50 billion non-null elements per iteration.

The introduction of cuts is a tool that brings great benefits; even though a too large number of cuts is not suitable. Therefore, we are currently experimenting with a strategy to determine the number of cuts (a moderate number of stronger cuts) and the appropiate iterations where they should be added, in order to increase the efficiency of the identification procedure.

In any case, we can verify the results obtained in [16] regarding the *m/n* ratio with the iterative process proposed in this work. For this purpose, we have constructed several samples leaving the number of measurements fixed and using a systematic sampling to select a smaller number of meters. In order to determine T_{jt}, the part of the simulated network corresponding to the selected counters has been recovered.

Noiseless Case

In this case, we have drawn 36 samples. In each dataset, customers are randomly assigned to different feeders in order to generate feeder measurements. Specifically, we have generated first a set of 12 samples named type I, consisting of 900 customers (*m/n* = 0.2389); then, a second set of 12 type II samples made up of 1000 customers (*m/n* = 0.215); and finally, a third set of 12 type III samples consisting of 1100 customers *(m/n* = 0.195).

Table 2 shows the results of the proposed algorithm (***PA***) (right column) compared to those obtained in a simple iterative process (***SI***) (left column) of adding new measurement constraints for each type of sample.

TABLE 2: LARGE CASE

Sample	**Proportion of correct identifications**					
	Type I (m/n=0,2389)		***Type II (m/n=0,215)***		***Type III (m/n=0,195)***	
	SI	***PA***	***SI***	***PA***	***SI***	***PA***
1	1	1	0.11	1	0.565	0.691
2	1	1	0.831	1	0.539	0.725
3	1	1	0.748	1	0.492	0.675
4	1	1	0.832	1	0.596	0.830
5	1	1	0.333	1	0.559	0.793
6	1	1	0.794	1	0.529	0.750
7	1	1	0.725	1	0.5	0.656
8	1	1	0.833	1	0.615	0.749
9	1	1	0.783	1	0.559	0.718
10	1	1	0.103	1	0.549	0.703
11	1	1	0.04	1	0.574	0.785
12	1	1	0.759	1	0.542	0.710

It is easy to notice that these results greatly improve those provided by Mangasarian et al. in [16].

Noisy Case

From the source data and using the simulated values of T_{jt}, we conduct again a systematic sampling, and select 60 samples. In each dataset, customers are randomly assigned to different lines in order to generate a reference case. Specifically, there are five sets of 12 samples each. In order to choose these sets,we have considered different combinations of the number of meters at each set (*n*) and the relationship between this number and the number of hourly measurements used (*m/n*).

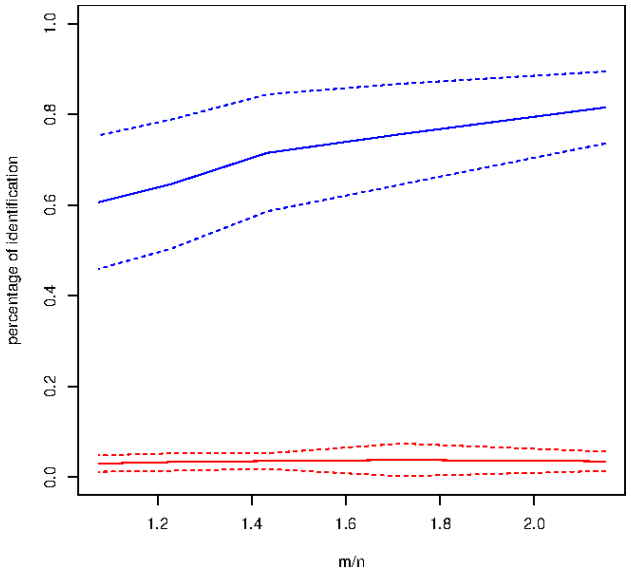

Fig. 2. Correct and incorrect identifications (%)

Figure 2 shows these average percentages of correct and incorrect identifications (with their corresponding 99% confidence intervals, represented by dashed lines). Thus, after running the analysis on 60% of the samples, when the ratio *m/n* varies between 1.08 and 2.15, the probability of recovering the true solution is between 60% and 80%. At the same time, the probability of obtaining an incorrect solution is practically nil.

IV. CONCLUSIONS AND FUTURE WORK

In this paper, and based on the mathematical formulation of the feeder mapping problem in distribution networks as an optimization model, we have developed an iterative procedure in which how and when we should include the cuts in an optimal way is still under experimentation. Even so, the computational experience carried out on a real data set, provide results that support the efficiency of the proposed scheme. Further computational work with the algorithm in its latest version is yet necessary. In addition, in order to support the results, its behavior should be tested with new real data collections, including the case of missing measurements or partial incorporation of known information, in order to further explore the performance of the proposed procedure.

ACKNOWLEDGMENT

This research has been partially supported by the project "Detección de línea mediante datos metrológicos" from ZIV Automation and the research projects Grupo de Investigación EOPT (IT-1252-19) from the Basque Government, GIU17/011 and GIU20/054 from University of the Basque Country (UPV/EHU), and PID2019-104933GB-I00 from the Spanish Ministry of Science and Innovation. This work is also part of the Flexigrid project that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 864579.

REFERENCES

[1] V. Arya, D. Seetharam, S. Kalyanaraman, K. Dontas, C. Pavlovski, S. Hoy, and J. Kalagnanam, "Phase identification in smart grids," presented at the 2nd IEEE International Conference on Smart Grid Communications, 2011.

[2] J. Bouford and C. Warren, "Many states of the distribution," in Power and Energy Magazine, IEEE, vol. 5(4), 2007, pp. 24–32.

[3] Meter phase identification, by K. Caird. (January 2010). US Patent Application 20100164473. Patent no. 12/345702.

[4] B.L. Dietrich, L.F. Escudero, A. Garín and G. Pérez,, "O(n) procedures for identifying maximal cliques and non-dominated extensions for consecutive minimal covers and alternates", TOP, vol. 1, 1993, pp. 139–160.

[5] M. Dilek, "Integrated design of electrical distribution systems: phase balancing and phase prediction case studies," Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA, 2001.

[6] L.F. Escudero, A. Garín and G. Pérez,, "O(n log n) procedures for tightening cover inequalities", Eur. J. Oper. Res, vol. 113, 1999, pp. 676–687.

[7] L.F. Escudero, A. Garín and G. Pérez,, "An O(n log n) procedures for identifying facets of the knapsack polytope", Op. Res. Letters, vol. 31, 2003, pp. 211–218.

[8] R. E. Gomory, "Outline of an algorithm for integer solutions to lineal programs," in Bulletin of the American Mathematical Society, vol. 64, 1958, pp. 275–278.

[9] M. Guignard, "Preprocessing and optimization in network flow problems with fixed charges," in Methods of Operations Research, vol. 45, 1982, pp. 235–256.

[10] M. Guignard and K. Spielberg, "Logical reduction methods in zero-one programming (minimal preferred variables)," in Operations Research, vol. 29, 1981, pp. 49–74.

[11] IBM ILOG CPLEX, High-performance software for Mathematical Programming and Aplications, [Online]. Available: https://www.ibm.com/es-es/products/ilog-cplex-optimization-studio.

[12] J. Fan and S. Borlase, "The evolution of distribution," in Power and Energy Magazine, IEEE, vol. 7(2), 2009, pp. 63–68.

[13] E. L. Johson, M. M. Kostreva and U. H. Suhl, "Integer programming problems arising from large scale planning models," in Operations Research, vol. 33, 1985, pp. 803–819.

[14] E. L. Johson and M.W. Padberg, "Inequalities,clique facets and bipartie graphs," in Annals of Discrete Optimization, vol. 16, 1983, pp. 169–188.

[15] O. Mangasarian and M. Ferris, "Uniqueness of integer solution of linear equations," in Optimization Letters, 4, 2010, pp. 559–565.

[16] O. Mangasarian and B. Recht, "Probability of unique integer solution to a system of linear equations," in European Journal of Operational Research, 214(1), 2011, pp. 27–30.

[17] G.L. Nemhauser and L.A. Wolsey, "Integer and Combinatorial Optimization," Wiley, New York, 1988.

[18] M. W. Padberg, "On the facial structure of set packing polyedra," in Mathematical Programming, vol. 5, 1973, pp. 199–215.

[19] P. K. Panda, V. Arya, D.A. Bowden, and L. Kohrmann, "Leveraging DERs to improve the inference of distribution network topology," presented at the IEEE International Conference on Smart Grid Communications, 2017.

[20] H. Pezeshki and P. Wolfs, "Correlation based method for phase identification in a three phase distribution network," presented at the 22nd Australasian Universities Power Engineering Conference (AUPEC), Bali, Indonesia, 26th-29th September, 2012.

[21] The R Project for Statistical Computing, [Online]. Available: https://www.r-project.org/.

[22] T.A. Short, "Advanced metering for phase identification, transformer identification and secondary modeling," in IEEE Transactions on Smart Grid Communications, vol. 4(2), 2013, pp. 651–658.

[23] J. Zhu, M.-Y. Chow, and F. Zhang, "Phase balancing using mixed-integer programming," in IEEE Transactions on Power Systems, vol. 13(4), 2013, pp. 1487–1492.