NTL Detection: Optimization of Inspection Routes Weighing Mobility Cost and Detection Likelihood

Pablo Massaferro
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
pmassaferro@fing.edu.uy

J. Matías Di Martino
Electrical & Computer Engineering Department
Duke University
North Carolina, USA
matias.di.martino@duke.edu

Alicia Fernandez
Electrical Engineering Institute
Universidad de la República
Montevideo, Uruguay
alicia@fing.edu.uy

***Abstract*—Countermeasures to Non-Techincal Losses (NTL) are primarily motivated by the high economic losses they represent. Most solutions focus on detecting suspicious behavior, aiming at identifying fraudulent activities in a data-driven and automated fashion. The direct economic impact associated with these solutions is commonly overlooked. In this work, we focus on the economic impact related to an NTL solution, where the list of customers to be inspected and the inspection routes are optimized. The clients to be examined are selected considering estimates of the fraud probability, the estimated magnitude of the fraud, and the costs associated with a particular inspection route. The mathematical formulation of the optimization problem leverages existing techniques developed in the context of a problem known as "the salesman problem." We validate our approach in real-world data obtained after inspecting over 168k (fraudulent and typical) customers in Uruguay, South America. The results show the relevance of taking routing costs into account and indicate that the economic harm of NTL can be reduced by 76%.**

I. Introduction

Due to the form of physical transport of energy in low voltage, companies are vulnerable to energy theft or other consumption measurement problems. This implies that a percentage of the total energy delivered to consumers is not monetized. Energy losses include technical losses (TL) resulting from dissipation of components on the grid, and non-technical losses (NTL), due to faulty equipment, billing errors, fraudulent meter manipulation or direct connections. NTL problem involves millions in losses to companies and has been a subject of academic research in the last decades. Even today, with the use of smart meters and the rise of deep learning techniques, it is an active research area [1].

To control NTL, companies carry out on-site inspections of clients' meters, if an irregularity is found, they apply a fine proportional to the estimate of the fraud carried out. Defining the set of customers to be inspected, is one main problem, as it is intractable to inspect all clients regularly. This work seeks to define the subset of clients and the inspection routes that maximize the economic return, in other words, which is the sequence of clients to be inspected that generates the highest income from fines, minimizing the inspection costs.

The problem of minimizing the rute's cost in a graph (given a set of nodes) is known as *Traveling Salesman Problem* (TSP). It is a classic NP-complete combinatorial optimization problem. In contrast with the canonical TSP, NTL involves a joint optimization of the route and selecting the nodes (potential fraudulent customers). On the other hand, there is a time limitation that prevents a single route from solving the problem (work shifts). Multiple routes are necessary to model realistically how NTL countermeasures are implemented in practice. The research questions are, given a set of electrical consumption data from residential and commercial customers; 1) What is the inspection planning method that generates the highest positive economic impact? 2) What is the optimal number of inspections to be performed? 3) Given a list of candidates to be inspected, what are the most suitable inspection routes?

II. Related Works

For several decades there has been a very important academic activity in the area of route optimization with multiple variants and diverse applications to the *Vehicle Routing Problem* (VRP) [2]. The problem of determining a route over a subset of points by optimizing a score is known as the *Orienteering Problem* (OP) [3]. This is a combination of selecting nodes and finding the shortest path between them. The objective is to maximize the total score of the visited nodes. This problem could also be seen as the combination of two classic combinatorial optimization problems, the *Traveling Salesman Problem* and the *Knapsack Problem* [4]. The TSP problem has also been addressed considering profits [5], which adds restrictions to the TSP optimization. An interesting and related approach considers route time restriction; these methods are commonly referred to as *Team Orienteering Problems* (T-OP) [6], [7]. As we shall see, these methods are particularly well suited for optimizing routes in the context of NTL detection. In [6] a heuristic based on simulated annealing is used where the compromise between exploration and exploitation of possible solutions is controlled by a "temperature" parameter that decreases with iterations.

If we take the simplified problem (TSP): and given a reduced set of points, suppose 20, if we want to find the route that passing only once through the 20 points minimizes the distance traveled, we would have to compute the permutation of 20 alternatives, meaning testing 2.4×10^{18} which is computationally very inefficient. There are several approximations to this problem that ensure local minimums. An efficient method with a low computational cost is the adaptation of self organized maps from a network to a ring configuration proposed by Brocki et al. [8]. This algorithm has a python implementation [9] which is used in the present work.

To the best of our knowledge, in the context of NTL, no prior work has study the joint optimization of detecting potential fraudulent customers and the actual inspection routes. The present work extends ideas introduced in Massaferro et al. [10], where a method was proposed to maximize the economic return taking into account a fixed inspection cost (independent of the inspection route). In addition, the present work considers a larger dataset (first introduced in [11]), which allows us to introduce novel discoveries and validate previous findings in a more extensive and diverse dataset.

III. Problem Formulation

A. Economic return estimation

Unlike the classic TSP where the profit collected at each visited node is known, we need to estimate the return value based on two factors. (i) The probability that a given customer is committing fraud, and (ii) if fraud is perpetrated, the expected magnitude of the power stolen. To this ends, the method described in [10] is used, training the eXtreme Gradient Boosting (XGBoost) algorithm, which is the one that reports the best results on the database to be used [11]. XGBoost is an efficient implementation of the Gradient Boosting algorithm [12]. This algorithm, used for both classification and regression, is an assembly of decision trees with a differentiable loss function, allowing gradient descent. Fraud probability $P(y_i = 1|x_i)$ is learned using a set of labeled training data where y_i represents the class label of th i_{th} customer and x_i the available data.

A regressor is trained using data from the positive class to estimate the economic return of detecting fraud (a_i). From now on, the estimated economic return for the customer i_k is $r_i = a_i P(y_i = 1|x_i)$.

B. Optimization Problem Formulation

We formulate this problem as a combination of *Team Orienteering Problem*(T-OP) and *Capacitated Profitable Tour Problem* (C-PTP) [2]. Defining each route as a graph $R = (V, A)$, where V represents the set of vertices i and A the set of arcs (i, j), we define the following variables:

- y_{ik} represents if the i^{th} customer is visited at the k^{th} path.
- x_{ijk} is equal to 1 if the node j is visited immediately after the node i. This variables form the adjacency matrix associated to the route graph.

The objective is to maximize the energy recovery r_i by minimizing the inspection costs c_i as seen in the objective function defined in (1).

$$max \sum_{i \in V} \sum_{k=1}^{M} r_i y_{ik} - \sum_{i,j \in A} \sum_{k=1}^{M} c_{i,j} x_{ijk} \tag{1}$$

$$s.t. \sum_{k=1}^{M} \sum_{j=1}^{n} x_{0jk} = \sum_{k=1}^{M} \sum_{i=1}^{n} x_{i0k} = M \tag{2}$$

$$\sum_{(i,j) \in V_k} (\alpha d_{ij} + \beta) x_{ijk} < T \;\; \forall k \in [1, 2, ..., M] \tag{3}$$

$$\sum_{k=1}^{M} y_{ik} \leq 1 \;\; \forall i \in [1, 2, ..., n] \tag{4}$$

$$\sum_{i=1}^{n} x_{ilk} = \sum_{i=1}^{n} x_{lik} = y_{lk} \;\; \forall l \in [1, 2, ..., n] \;\; \forall k \in [1, 2, ..., M] \tag{5}$$

$$c_{i,j} = c_{time}(\alpha d_{ij} + \beta) + c_{distance} d_{ij}. \tag{6}$$

Eq. (2) restricts all routes (M) to start and finish on the origin point. The second restriction (Eq. 3) limits the travel time of all routes to not exceed T (total time of the shift, in this case 7.5hs), α is a time/distance coefficient (inverse of the average speed), d_{ij} represents the distance between nodes i and j (Euclidean distance is used) and β is the average duration of an on-site inspection. The constraint of Eq. (4) ensures that customers are visited at most once while the constraint presented in Eq. (5) ensures connectivity on routes by making every client have an incoming connection and an outgoing connection if they are part of a route. Eq (6) shows the components of the cost of performing the j inspection after i inspection. The constant c_{time} is the cost linked to the hours of the personnel and $c_{distance}$ the cost of the trip linked to the cost of fuel and vehicles maintenance.

IV. Proposed approach

Since the optimization problem described is NP-complete, is common to obtain approximated solutions based on heuristics that approaches local optimal [4]. In order to maximize the economic return on the inspection routes, we study three methods that we referred as: *Naive, Nearest Neighbor top M* and *SOM/TSP*.

A. Naive Method

This method consists of a ranking of customers ordered by the estimated amount of energy recovery. The routes start at the point of origin adding clients following the order of the ranking until visiting the client m_k and returning to the origin. The amount of routes performed is the one that gives the highest gain according to equation 1. Unlike the methods presented below, this method only implies the calculation of n distances, where n is the number of samples in the dataset.

B. Nearest Neighbor top M Method

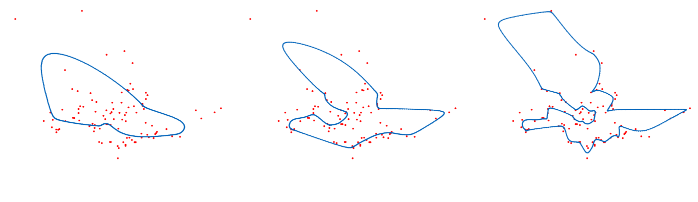
The *Naive* method does not optimize the inspection route because the cost is only defined by the customers with a fraud likelihood and volume, without considering their relative location. A simple method is proposed to minimize the distance traveled by maximizing the return. In this method, clients will not necessarily be visited in the order in which they appear in the loss recovery ranking. From the point of origin and onward, the distance to the M clients with the highest estimated recovery (not visited) is evaluated and the closest one is selected. This avoids computing the distance matrix for the entire data base and reduces the computation of $n(n-1)/2$ to $M\,n$ distances. The list of top M options ensures that clients with a high estimate of economic recovery are visited on the first routes and the use of *Nearest Neighbor* allows to reduce inspection costs by carrying out a greater number of inspections per route.

C. SOM/TSP Method

This method involves solving the TSP for a set of m clients and using *Self Organized Maps* (SOM) to choose the route. The m clients are chosen by ranking the estimated economic recovery. Determining the number of inspections, involves optimizing routes for different values of m on the same ordered list of customers.

1) Self Organized Maps: Also known as Kohonen maps [13], SOM is a type of unsupervised neural network that decreases the data dimensionality (generally to two dimensions) and allows to organize and visualize data. In its classical formulation, each neuron in the network is visualized as a node in a two-dimensional grid. Each neuron is associated with a vector of weights W of the same dimension as the input data. The objective of the training is to generate a new distribution in the mapped space that preserves the similarity relationship of the input data. The training is carried out as follows:

1) Weights are initialized randomly.
2) For each input data, the closest neuron is selected by measuring the distance of the data to all the weight vectors of the network. The selected node is called BMU (best matching unit).
3) BMU weights and neighboring nodes are updated according to the following equation: $W_v(k+1) = W_v(k) + \theta(u,v,k)\alpha(k)(X_i - W_v(K))$, where k represents the iteration step, u the BMU neuron and v are the nodes of the network. To update around BMU a neighborhood factor θ is used computing a convolution with a Gaussian kernel (kernel radius decreases with iterations). The weights W_v are updated in the direction of the data X_i using the learning coefficient $\alpha(k)$ that decreases monotonously with the iterations.
4) It iterates until the learning rate or kernel radius has fallen below a threshold.



(a) 1000 iter (b) 4000 iter (c) 7000 iter

Fig. 1: Routing evolution by iterations number with Self Organized Maps for 100 points in Montevideo

2) SOM to TSP adaptation: . It is proposed to adapt the SOM technique to the TSP by modifying the proximity structure of the network from a grid to a ring topology. In this way each node remains in a sequence. The input data and weight vectors are of dimension 2 (latitude and longitude). The location of a customer is associated with each BMU, so that each node in the ring gets closer to a customer as the iterations increase, as illustrates Fig. 1.

Exploration begins with a value of α large, which decreases with iterations in the exploitation phase. To improve the local fit, the neighborhood coefficient also decreases with iterations.

3) Routing with SOM/TSP: The routing method for maximum economic return with SOM/TSP is as follows:

1) The m clients with the highest estimated economic return are selected.
2) A route is set to minimize the distance between the m clients using SOM/TSP.
3) The route is divided into sub routes in order to meet the time constraints (equation 3 constraint). It is assumed that if r_o is the optimal path between the nodes $[1..m]$ then $r_o[1..k]$ is the optimal path of $[1..k]$ $\forall k < m$.
4) The total cost of all the routes that allow the m customers to be inspected and the actual return obtained is computed. Finally, we return to first step to cover the range of m that allows us to identify the number of inspections with the highest return.

Figure 1 shows an example of SOM/TSP application to solve the routing of 100 random clients in Montevideo.

V. Experiments and Results

A. Data Set

A dataset of 168k customers provided and in-site inspected by the company UTE are considered in this study. Inspections were performed during 2017 and 2020 by expert technicians. The

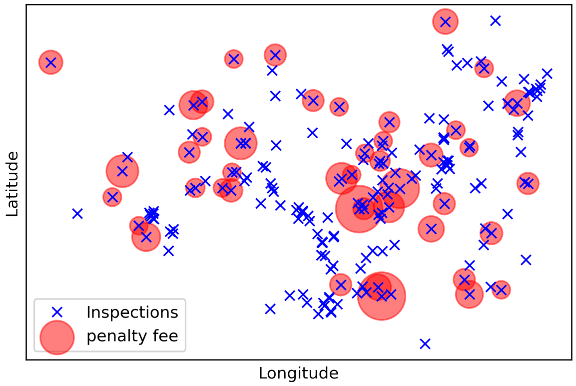


Fig. 2: We illustrate the geographic location of 200 random samples. Each cross represents a customer inspected. For those customers for which fraud was detected, the red circle illustrates the magnitude of the fine defined (estimated from an approximation of the magnitude of the fraud). To preserve the users privacy, we omit providing the absolute latitude and longitude values.

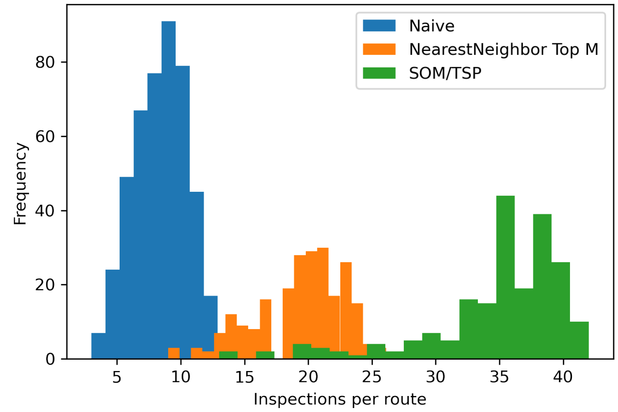


Fig. 3: Histogram of inspections per route after tunning the operating point that maximizes the return. The mean value of inspections per route is 8.5, 19.4 and 34.8 for the *Naive*, *Nearest Neighbor* and SOM/TSP methods respectively.

percentage of fraudulent clients is 15.5% (26k). This database is divided into a training set and a test set. The test set corresponds to the last 20k inspections carried out in 2020. Using the training set, we predict the estimated potential fraud using the method described in [10]. Once the prediction on the test set is made, a database is formed that contains the following fields: ID, latitude, longitude, estimated economic return and real economic return.

Considering the restriction of a single source routing, the city of Montevideo was selected for evaluation, since it concentrates 75% of all the data.

B. Experimental Results

For the experiments presented below, the following cost parameters were used in the Eqs. (3) and (6): $c_{distance} = 10\$/km$, $c_{time} = \$1000/h$, $\alpha = 5min/km$, $\beta = 10min$. For the experiments carried out with SOM/TPS, a learning coefficient $\alpha(k) = 0.8(1 - 0.3 \times 10^{-5})^k$ was used. The neighborhood coefficient of the BMU is calculated using a Gaussian filter whose variance decreases with the iterations in the following way $\sigma(k) = C(1 - 0.3 \times 10^{-4})^k$ where the constant C is proportional to the number of samples. The network is generated with a total of $8M$ neurons, being M the number of clients. The experiments were carried out for several M values between 50 and 15000. By achieving a more efficient route, the SOM/TSP algorithm allows more inspections within each route. Figure 3 shows the histogram of inspections per route for the three methods.

The objective is to maximize the economic return, so we evaluate the three strategies by calculating the maximum value

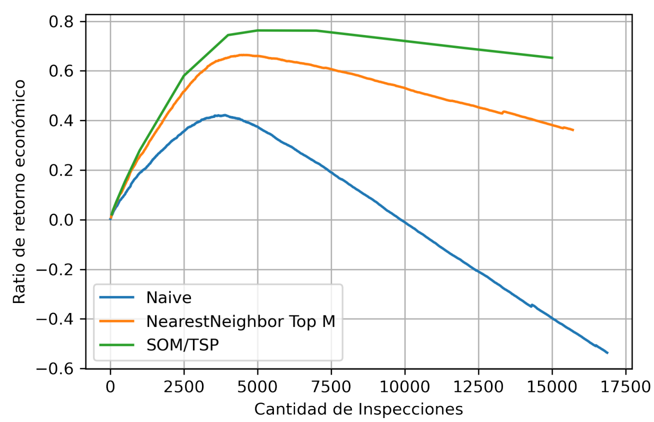


Fig. 4: Economic recovery ratio by method for different numbers of total inspections. The ratio is the quotient between the profit obtained with the method (income-costs) and the total possible income.

of the Eq. (1) for different numbers of inspections. Figure 4 shows the economic return ratio curves comparing the three methods. The results are expressed as a percentage of the total amount of fraud recovered. Table I shows the maximum economic return and for that operating point also shows the number of inspections and the number of routes to be done. The method that achieves the best result is SOM/TSP, generating a profit of 76.3% by performing 206 inspection routes with an average of 34.8 inspections per shift.

Methods	Naive	Nearest Neighbord Top M	**SOM/TSP**
Economic return	42.2%	66.4%	**76.3%**
Number of routes	460	232	206
Number of inspections	3895	4520	7000
#Inspections/#Routes	8.5	19.4	34.8
Precision	63.9%	57.9%	40.8%

TABLE I: Results of routing strategies at their point of maximum economic return over test data set.

VI. Conclusions and future work

We studied the problem of joint optimization of the detection and routing definition of inspections for customers with atypical consumption behavior. This work expands previous work and provides a useful and practical solution to help power companies minimize the impact of power theft. Results are reported on a new database of 168 thousand real customers, which were in-site inspected and labeled by expert technicians of the Uruguayan national power company UTE. One of the main contributions of this work is the mathematical formalization of the objective problem with references to the area of *Vehicle Routing* (VRP). Routing problems are NP-complete problems that cannot be solved with exact solutions when the number of samples is moderately large (e.g., larger than a few hundred). The problem is contextualized into the TSP, and recent advances in that field are leveraged to optimize NTL solutions. Three routing methods are evaluated and discussed, two of them based on simple heuristics and a third based on *Self-Organized Maps*. The results obtained show the relevance of addressing this problem using metrics that reflect the main objectives.

Works that perform classification or identification of fraud focus on standard classification metrics (Precision, Recall, F Measure) or metrics that compare models by modifying classification thresholds such as the area under the receiver operating characteristic or precision-recall curves. In this work, the economic return is used as target optimization metric. The selection of clients to be inspected takes into account the trade-off between profits and cost. A non-optimized inspection routing strategy could generate high economic costs. For example, customers for which the transportation costs associated with inspection are high should be inspected only if there is a substantial likelihood of fraud. This works formalizes this intuitive idea and frames the problem into a mathematical framework to quantitatively optimize this notion.

The experiments presented show that the SOM/TSP method allows an average of 34.8 inspections per route while the *Nearest Neighbor TOP M* method achieves an average of 19.4 inspections. The difference of almost ten percentage points in economic recovery between both methods can translate into a significant amount of money in the real problem. Future work could explore other solutions used in *Team-Orienteering Problems* and the *Travel Salesman Problem with Profits*, such as simulated annealing. Marginal gain heuristics should also be studied to discard nodes from routes and compare results with other state-of-the-art algorithms, e.g., Google's OR-Tools [14].

Acknowledgments

The authors thank UTE for providing the datasets and for sharing their expertise in particular, authors greatly acknowledge engineers Marcelo Álvarez, Gonzalo Caudullo, Ibero Fomichov, Agustín Heverling and Alexander Martins. We also want to thank Marcelo Fiori and Juan Bazerque fruitful discussions.

References

[1] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250–266, 2018.
[2] P. Toth and D. Vigo, *Vehicle routing: problems, methods, and applications*. Society for Industrial and Applied Mathematics, 2014.
[3] A. Gunawan, H. C. Lau, and P. Vansteenwegen, "Orienteering problem: A survey of recent variants, solution approaches and applications," *European Journal of Operational Research*, vol. 255, no. 2, pp. 315–332, 2016.
[4] P. Vansteenwegen, W. Souffriau, and D. Van Oudheusden, "The orienteering problem: A survey," *European Journal of Operational Research*, vol. 209, no. 1, pp. 1–10, 2011.
[5] D. Feillet, P. Dejax, and M. Gendreau, "Traveling salesman problems with profits," *Transportation science*, vol. 39, no. 2, pp. 188–205, 2005.
[6] S.-W. Lin, "Solving the team orienteering problem using effective multi-start simulated annealing," *Applied Soft Computing*, vol. 13, no. 2, pp. 1064–1073, 2013.
[7] M. Poggi, H. Viana, and E. Uchoa, "The team orienteering problem: Formulations and branch-cut and price," in *10th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS'10)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010.
[8] Ł. Brocki and D. Koržinek, "Kohonen self-organizing map for the traveling salesperson problem," in *Recent Advances in Mechatronics*. Springer, 2007, pp. 116–119.
[9] D. Vicen, "Solving the traveling salesman problem using self-organizing maps," https://github.com/DiegoVicen/som-tsp, 2018.
[10] P. Massaferro, J. M. Di Martino, and A. Fernández, "Fraud detection in electric power distribution: An approach that maximizes the economic return," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 703–710, 2019.
[11] ——, "NTL detection: Overview of classic and dnn-based approaches on a labeled dataset of 311k customers," in *2021 IEEE NA Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2021.
[12] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.
[13] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
[14] L. Perron and V. Furnon, "Or-tools," Google. [Online]. Available: https://developers.google.com/optimization/