

Is Big Data Sufficient for a Reliable Detection of Non-Technical Losses?

Patrick Glauner*, Angelo Migliosi*, Jorge Augusto Meira*, Petko Valtchev*[†], Radu State* and Franck Bettinger[‡]

*Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg

2721 Luxembourg, Luxembourg

Email: {first.last, jorge.meira}@uni.lu

[†]Department of Computer Science, University of Quebec in Montreal,

H3C 3P8 Montreal, Canada

Email: valtchev.petko@uqam.ca

[‡]CHOICE Technologies Holding Sàrl

2453 Luxembourg, Luxembourg

Email: franck.bettinger@choiceholding.com

Abstract—Non-technical losses (NTL) occur during the distribution of electricity in power grids and include, but are not limited to, electricity theft and faulty meters. In emerging countries, they may range up to 40% of the total electricity distributed. In order to detect NTLs, machine learning methods are used that learn irregular consumption patterns from customer data and inspection results. The Big Data paradigm followed in modern machine learning reflects the desire of deriving better conclusions from simply analyzing more data, without the necessity of looking at theory and models. However, the sample of inspected customers may be biased, i.e. it does not represent the population of all customers. As a consequence, machine learning models trained on these inspection results are biased as well and therefore lead to unreliable predictions of whether customers cause NTL or not. In machine learning, this issue is called covariate shift and has not been addressed in the literature on NTL detection yet. In this work, we present a novel framework for quantifying and visualizing covariate shift. We apply it to a commercial data set from Brazil that consists of 3.6M customers and 820K inspection results. We show that some features have a stronger covariate shift than others, making predictions less reliable. In particular, previous inspections were focused on certain neighborhoods or customer classes and that they were not sufficiently spread among the population of customers. This framework is about to be deployed in a commercial product for NTL detection.

Index Terms—Bias, big data, covariate shift, machine learning, non-technical losses.

I. INTRODUCTION

Losses in power grids can be grouped into technical losses and non-technical losses. Technical losses occur naturally, which are mainly caused by internal electrical resistance of infrastructure components. Non-technical losses (NTL) appear during power distribution and include, but are not limited to, the following causes [1], [2]:

- 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
- Technical and human errors in meter readings, data processing and billing



Fig. 1. Example of spatial bias: The large city is close to the sea, whereas the small city is located in the interior of the country. The weather in the small city undergoes stronger changes during the year. The subsequent change of electricity consumption during the year triggers many inspections. As a consequence, most inspections are carried out in the small city. Therefore, the sample of customers inspected does not represent the overall population of customers.

In practice, NTLs primarily consist of electricity theft and cause major problems to electricity providers, including financial losses and a decrease of stability and reliability. Furthermore, NTLs lead to an extra use of limited natural resources which in turn increases pollution. They can range up to 40% of the total electricity distributed in countries such as Brazil, India, Malaysia or Lebanon [3], [4].

The predominant research direction reported in the recent literature is the use of machine learning/data mining methods, which learn anomalous behavior from customer data and known irregular behavior that was reported through inspection results. However, carrying out inspections is expensive, as it requires physical presence of technicians. It is therefore important that the trained models make accurate predictions.

For about the last fifteen years, the Big Data paradigm followed in machine learning has been to gather more data rather than improving models. Hence, one may assume that having simply more customer and inspection data would help to detect NTL more accurately. However, in many cases, the data is biased as depicted in Fig. 1. Concretely, the customers

inspected are a sample of the overall population of customers. In this example, there is a spatial bias. Hence, the inspections do not represent the overall population of customers. As a consequence, when learning from the inspection results, a bias is learned, making predictions less reliable. In technical terms, this bias is called covariate shift or sampling bias [5]. Aside from spatial covariate shift, there may be other types of covariate shift in the data, such as the meter type, connection type, etc.

The main contributions of this paper are:

- We present a novel framework for quantifying and visualizing covariate shift at different hierarchical levels.
- We investigate the importance of covariate shift to NTL detection in Big Data approaches and show that it leads to unreliable NTL predictors.
- We report the covariate shift of different features in a real world data set consisting of 3.6M Brazilian customers and 820K inspection results.
- We visualize how different local areas are affected by covariate shift at different hierarchical levels.

To the best of our knowledge, we are not aware of any previously published research that addresses this topic for NTL detection. We are convinced that an accurate study of this topic is necessary in order to advance NTL detection by reducing the covariate shift in data sets in the future.

II. RELATED WORK

From an electrical engineering perspective, energy balance methods [6] can be applied to the detection of NTL. However, this requires topological information of the distribution network and does not reflect a change of network. In practice, the network topology undergoes rapid changes in emerging countries, i.e. the countries in which NTL is a particular issue.

The predominant methodology used in the literature is employing artificial intelligence (AI) to this anomaly detection problem. Historically, AI has been based on rule-based expert systems that incorporate expert knowledge. However, in many cases, describing a problem domain is challenging due to its complexity and temporal dynamics. Therefore, over the years, the fields of machine learning and data mining have become more popular. These methods learn models from data without being explicitly programmed.

A data set of ~22K customers is used in [7] for training a neural network. It uses the average consumption of the previous 12 months and other customer features such as location, type of customer, voltage and whether there are meter reading notes during that period. On the test set, an accuracy of 0.8717, a precision of 0.6503 and a recall of 0.2947 are reported. Consumption profiles of 5K Brazilian industrial customer profiles are analyzed in [8]. Each customer profile contains 10 features including the demand billed, maximum demand, installed power, etc. In this setting, a support vector machine slightly outperforms k-nearest neighbors and a neural network, for which test accuracies of 0.9628, 0.9620 and 0.9448, respectively, are reported.

We have extensively reviewed the state of the art in our previous work and identified the open challenges in NTL detection [2]. We have previously addressed the class imbalance and evaluation metric selection, when we showed that a large-scale machine learning approach outperformed rule-based Boolean and fuzzy logic expert systems [9]. Furthermore, we have shown that the neighborhood of customers yields significant information in order to decide whether a customer causes a NTL or not [10], [11].

III. COVARIATE SHIFT

In this paper, we address the challenge of covariate shift. It refers to the problem of training data (i.e. the set of inspection results) and production data (i.e. the set of customers to generate inspections for) having different distributions. This fact leads to unreliable NTL predictors when learning from this training data. In this section, we describe how to quantify it. Historically, covariate shift has been a long-standing issue in statistics. For example, The Literary Digest sent out 10M questionnaires in order to predict the outcome of the 1936 US Presidential election. They received 2.4M returns. Nonetheless, the predicted result proved to be wrong. The reason for this was that they used car registrations and phone directories to compile a list of recipients. In that time, the households that had a phone or a car represented a biased sample of the overall population. In contrast, George Gallup only interviewed 3K handpicked people, which were an unbiased sample of the population. As a consequence, Gallup could predict the outcome of the election very well [12].

A. Definition

Customers previously inspected are a sample of the overall population of customers. However, that sample may be biased, i.e. it does not represent the population of all customers, as visualized in Fig. 1. A reason for that is that previous inspections were largely focused on certain criteria and were not sufficiently spread among the population. The problem of training data and production data having different distributions has initially been addressed in the field of computational learning theory [13], which also calls it covariate shift, sampling bias or sample selection bias. It can be defined in mathematical terms as follows [5]:

- Assume that all examples are drawn from a distribution D with domain $X \times Y \times S$,
- where X is the feature space,
- Y is the label space,
- and S is $\{0, 1\}$.

Examples (x, y, s) are drawn independently from D . $s = 1$ denotes a selected example, whereas $s = 0$ denotes the opposite. The training is performed on a sample that comprises all examples that have $s = 1$. $P(s|x, y) = P(s|x)$ implies that s is independent of y given x . In this case, the selected sample is biased but the bias only depends on the feature vector x . This problem is called covariate shift [5].

B. Big Data

In the past fifteen years, the Big Data paradigm followed can be summarized: “It’s not who has the best algorithm that wins. It’s who has the most data.” [14] Concretely, this approach reflects the desire of deriving better conclusions from simply analyzing more data, without the necessity of looking at theory and models. However, taking covariate shift into account, just having more data is not sufficient. In this case, having less data that is more representative seems to be the answer.

C. Affected Classifiers

It has been shown that some machine learning algorithms are not affected by covariate shift, whereas others are [5].

1) *Local learner*: The prediction of the learner depends asymptotically only on $P(y|x)$. Hence, it is not affected by covariate shift. Examples are logistic regression and hard-margin support vector machine (SVM).

2) *Global learner*: The prediction of the learner depends asymptotically on both, $P(y|x)$ and $P(x)$. Hence, it is affected by covariate shift. Example: Decision tree learners such as ID3 or C4.5 [15] recursively split the input space by choosing the remaining most discriminative feature of a data set. To predict, the learned tree is traversed top-down. Other examples are naive Bayes and soft-margin SVM.

D. Quantification

The Kullback-Leibler divergence is a measure of the difference of two probability distributions. However, it is challenging (1) to adapt this measure to multi-dimensional data that is a combination of discrete and continuous features, which is common in machine learning, and (2) to define criteria when a distance is an indicator for a covariate shift.

Therefore, a preferred methodology for quantifying covariate shift is: First, we add a feature s and assign the values 1 or 0 to the training data ($s = 1$) or production data ($s = 0$), respectively. These data sets are furthermore merged into one data set. This latter is split into a training set X_1 (with no relation to the original training set) and a test set X_2 . The objective is to develop a supervised learning method capable of predicting the feature s using X_1 . The performance of the classifier on X_2 is then quantified using the Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (1)$$

which measures the accuracy of binary classifiers taking into account the imbalance of both classes, ranging from -1 to $+1$ [16]. The greater the MCC, the greater the covariate shift. A concrete threshold for covariate shift depends on the problem, however 0.2 has been proposed [17]. Though a low MCC does not automatically imply the lack of a covariate shift, a significant MCC value is an indicator of covariate shift.

We extend this methodology in Algorithm 1 by introducing the following novelties:

- 1) Tree classifier: Decision tree learning is affected by covariate shift. Decision trees scale to very large data

sets while they allow to learn non-linearities. Soft-margin SVMs are also global learners, however, for large data sets only a linear kernel is learnable in a feasible amount of time.

- 2) Model selection: We want to find a model which is able to distinguish between both distributions. Thus maximizing the MCC on the test set is equivalent to finding the best two-class classification between production data and original training data. For this, we optimize the five most important tree model parameters by randomly drawing from probability distributions: Max. number of leaves, max. number of levels, measure of the purity of a split, min. number of samples required to be at a leaf and min. number of samples required to split a node.
- 3) Cross-validation: We also split the data set into k folds in order to reduce the overfitting. This leads to a more reliable model for covariate shift quantification. The MCC per model, denoted by \overline{MCC} , is the average of the MCCs of the k test sets. The standard deviation of the k test MCCs serves as the reliability of \overline{MCC} . The lower the standard deviation, the more reliable \overline{MCC} .

Note: The inspection results are not taken into account as covariate shift only concerns the distributions of the inputs.

Algorithm 1 Quantifying covariate shift.

```

1: result  $\leftarrow$  0
2: reliability  $\leftarrow$  0
3: selected  $\leftarrow$  train_data.add_feature(s, 1)
4: not_selected  $\leftarrow$  prod_data.add_feature(s, 0)
5: data  $\leftarrow$  selected  $\cup$  not_selected
6: folds  $\leftarrow$  cv_folds(data, k)
7: for model in get_model_candidates() do
8:   mccs  $\leftarrow$  list()
9:   for fold in folds do
10:    Xtrain, Xtest, ytrain, ytest  $\leftarrow$  fold
11:    classifier  $\leftarrow$  DecisionTree(model)
12:    classifier.train(Xtrain, ytrain)
13:    ypred  $\leftarrow$  classifier.predict(Xtest)
14:    mccs.append(MCC(ytest, ypred))
15:   end for
16:   mcc_mean  $\leftarrow$  mean(mccs)
17:   if mcc_mean  $>$  result then
18:     result  $\leftarrow$  mcc_mean
19:     reliability  $\leftarrow$  std(mccs)
20:   end if
21: end for
22: return result, reliability

```

IV. EVALUATION

A. Data

The data used in this paper comes from an electricity provider in Brazil. First, it consists of 3.6M customers. A complete list of the customer master data used in the following experiments is depicted in Table I. The categorical features

TABLE I
ASSESSED FEATURES.

| Feature | Possible values |
|-----------------|--|
| Class | Power generation infrastructure, residential, commercial, industrial, public, public illumination, rural, public service, reseller |
| Contract status | Active, suspended |
| Location | Longitude and latitude |
| Meter type | 22 different meter types |
| Number of wires | 1, 2, 3 |
| Voltage | $\leq 2.3\text{kV}$, $> 2.3\text{kV}$ |

class, voltage, number of wires, contract status and meter type are converted to one-hot coding. Second, the data contains 820K inspection results, such as inspection date, presence of fraud or irregularity, type of NTL and inspection notes. 620K customers have been inspected at least once and the remaining $\sim 3\text{M}$ customers have never been inspected. Third, there are 195M meter readings from 2011 to 2016 such as consumption in kWh, date of meter reading and number of days between meter readings.

B. Implementation Notes

All computations were run on a server with 24 cores and 128 GB of RAM. The entire code was implemented in Python using `scikit-learn` [18] for machine learning. `scikit-learn` allows to distribute the training of the cross-validated classifiers among all cores. The maps were plotted using `cartopy` [19]. In total, all results and plots reported in this paper were computed in 12 hours using this infrastructure. Our implementation is available as open source: <http://github.com/pglauner/SpatialBiasNTL>.

C. Model Parameters

In the following experiments, we use $k = 10$ -fold cross-validation. In each experiment, we train 1K trees, which are 100 different tree models trained on each of the 10 folds. We optimize the five tree model parameters by randomly drawing from predefined uniform probability distributions depicted in Table II. We have chosen these ranges based on best practice recommendations and our own experience. Furthermore, the two classes ($s = 1$ and $s = 0$) are imbalanced, i.e. there are more examples of the non-inspected customers than the inspected ones. In order to take this into account during training, we associate weights with the classes such that the examples of the minority class have stronger impact.

D. Global Covariate Shifts

In the following experiments, we compute different global types of covariate shift by using all customers in each experiment. We therefore do not split the customers into different geographical areas.

We have previously presented the customer master data features available in Table I. We compute the global covariate

TABLE II
TREE MODEL PARAMETERS.

| Parameter | Range |
|---|-----------------|
| Max. number of leaves | [2, 20) |
| Max. number of levels | [1, 20) |
| Measure of the purity of a split | {entropy, gini} |
| Min. number of samples required to be at a leaf | [1, 20) |
| Min. number of samples required to split a node | [2, 20) |

TABLE III
GLOBAL COVARIATE SHIFT OF SINGLE FEATURES.

| Feature | \overline{MCC}_{max} | σ |
|-----------------|------------------------|----------------|
| Location | 0.22367 | 0.03453 |
| Class | 0.16255 | 0.01371 |
| Number of wires | 0.14111 | 0.00794 |
| Meter type | 0.13158 | 0.00382 |
| Voltage | 0.07092 | 0.02375 |
| Contract status | 0.03744 | 0.09183 |

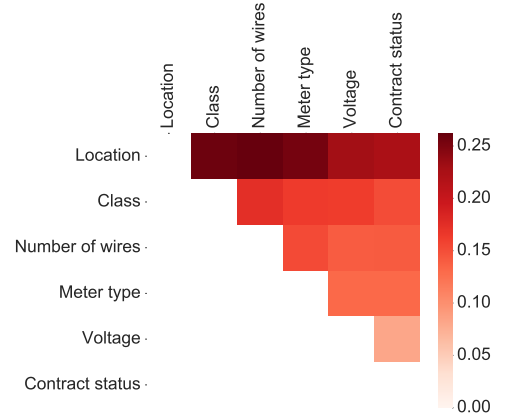


Fig. 2. Global Covariate Shift of Compound Features.

shift of each of these features. We report our results in Table III. \overline{MCC}_{max} denotes the maximum average of the MCCs of the $k = 10$ test sets among all 100 tree models trained on a feature. σ denotes the standard deviation of those $k = 10$ MCC test scores, which is a reliability measure of \overline{MCC}_{max} .

Overall, the strongest covariate shift is in the location with a MCC value of 0.22367. This means that previous inspections are mostly biased towards the location of customers. The location is also the only feature that is beyond the threshold of 0.2 mentioned in Section III-D. The features class, number of wires and meter type are below the threshold but are greater than 0.1. There is almost no covariate shift of previous inspections towards the voltage and contract status features. The standard deviation of the MCCs is the greatest for the contract status. The reason for this is a strong overfit in one of the folds. All other MCCs have a much lower standard

TABLE IV
GLOBAL COVARIATE SHIFT OF COMPOUND FEATURES.

| Feature | \overline{MCC}_{max} | σ |
|-----------------------------------|------------------------|----------------|
| All | 0.27325 | 0.03014 |
| Location + number of wires | 0.26206 | 0.03676 |
| Location + class | 0.25796 | 0.03540 |
| Location + meter type | 0.25479 | 0.03884 |
| Location + voltage | 0.22944 | 0.03544 |
| Location + contract status | 0.22335 | 0.03454 |
| Class + number of wires | 0.17501 | 0.00468 |
| Class + meter type | 0.16472 | 0.00309 |
| Class + voltage | 0.16322 | 0.01400 |
| Number of wires + meter type | 0.15283 | 0.00274 |
| Class + contract status | 0.15158 | 0.00992 |
| Number of wires + voltage | 0.14156 | 0.00800 |
| Number of wires + contract status | 0.14111 | 0.00794 |
| Meter type + voltage | 0.13165 | 0.00381 |
| Meter type + contract status | 0.13155 | 0.00382 |
| Voltage + contract status | 0.08213 | 0.08301 |

deviation, making them more reliable.

Next, we create compound features that are composed of multiple features. Due to the great number of possible combinations, we assess all 2-combinations as well as the 6-combination of all features. We visualize the MCCs for all 2-combinations in Fig. 2 and reported the MCCs in Table IV. For the 6-combination comprising all features, we computed $\overline{MCC}_{max} = 0.27325$, which is the maximum covariate shift of all compound features. Therefore, the spatial covariate shift contributes to this covariate shift the most, however, the other covariate shifts contribute a fraction as well.

E. Local Covariate Shifts

We now entirely focus on spatial covariate shift since it is the strongest one among the different types of covariate shift. In the following experiments we compute local covariate shifts by splitting the customers in different locations. The data set provides the following divisions in the following hierarchical order:

- 1) 9 regions
- 2) 261 municipalities
- 3) 1,380 localities
- 4) 19,026 neighborhoods

We plot the local spatial covariate shifts of these four levels of granularity in Fig. 3. All customers are located in one Brazilian state. We observe that spatial covariate shift is smoothened for regional level in Fig. 3a. It becomes increasingly more granular at municipal, local and neighborhood levels in Figs. 3a through 3d, respectively. We also notice that the spatial covariate shifts at lower levels tend to increase, which is depicted by increasing upper limits of the color bars.

F. Discussion

We have shown that covariate shift exists in our real world data set. The features of the customer data that are most

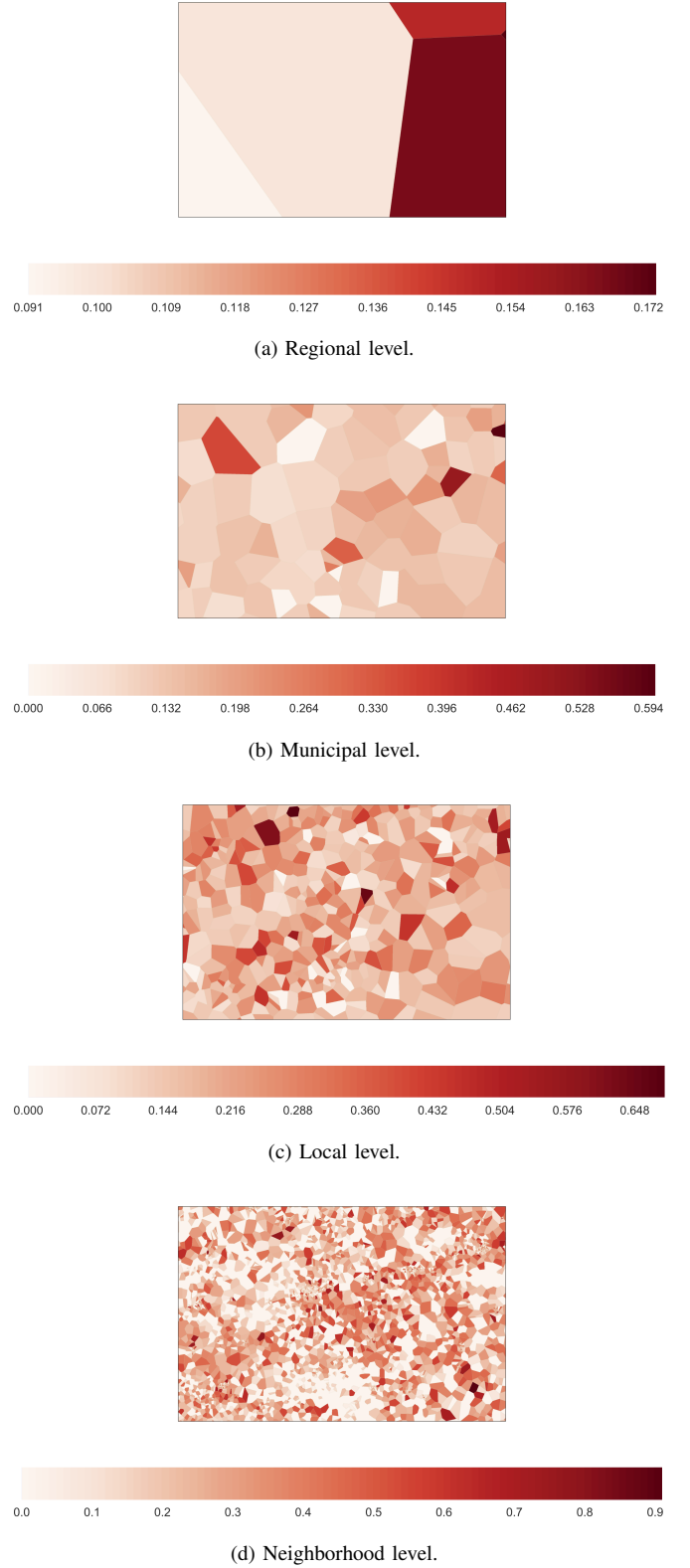


Fig. 3. Spatial covariate shift at different levels of granularity. For each division, we compute the median location of the respective customers and assign \overline{MCC}_{max} to it. We then use nearest interpolation to generate the local covariate shift maps. Note: The maps only visualize a fraction of all political divisions. This is because we have cropped the computed maps in this paper due to privacy purposes.

affected by covariate shift are the location, followed by class, number of wires and meter type. Classifiers that use other features such as the voltage or contract status instead tend to be more reliable. We have also shown that the spatial covariate shift exists on different levels of granularity and that municipalities and localities with very strong covariate shifts exist. Subsequently, these local covariate shifts have significant impact on the covariate shifts on higher levels or even globally on the entire data set. Therefore, using all inspection results of the data set for training a NTL predictor from a Big Data perspective leads to biased models that may not reliably detect NTL. As a consequence, reducing the spatial covariate shift in the data set must be a priority in order to learn reliable NTL predictors.

However, the finer the hierarchical granularity, the more divisions cannot be used for the computations for the following reasons. First, using $k = 10$ -fold cross-validation, training is only possible if a division has at least k customers. Second, the $k - 1$ folds used for training must have examples of both classes. Third, the MCC can only be computed for denominator $\neq 0$, which is the case for $(TP > 0 \wedge TN > 0) \vee (FP > 0 \wedge FN > 0)$. If the test MCC cannot be computed for a fold of a model, only the MCCs of the remaining folds are used in cross-validation. If no MCCs can be computed for a division, we skip it in the plotting. For instance, this effect has become most apparent at neighborhood level in the west of that state due to the low population density.

V. CONCLUSION AND FUTURE WORK

Covariate shift is the problem of training data and production data having different distributions. In this work, we have proposed a novel framework for quantifying and visualizing covariate shift in spatial data sets. In the context of non-technical loss (NTL) detection, we showed that there is a covariate shift between the inspected customers and the overall population of customers. We used a real world data set from Brazil that consists of 3.6M customers and 820K inspection results. We showed that some features have a stronger covariate shift than others. In particular, the spatial covariate shift is the strongest and appears in different hierarchical levels. Subsequently, machine learning models trained on this data will lead to unreliable NTL predictions. Our contribution will allow domain experts to model more reliable rules for NTL predictors.

Next, we will use and amend spatial point processes [20] for reducing the spatial covariate shift in our data set in order to train more reliable NTL predictors.

ACKNOWLEDGMENT

We would like to thank Lautaro Dolberg, Diogo Duarte and Yves Rangoni from CHOICE Technologies Holding Sàrl for contributing many good ideas to our discussions. This work has been partially funded by the Luxembourg National Research Fund.

REFERENCES

- [1] A. Chauhan and S. Rajvanshi, "Non-technical losses in power system: A review," in *Power, Energy and Control (ICPEC), 2013 International Conference on*. IEEE, 2013, pp. 558–561.
- [2] P. Glauner, A. Boechat, L. Dolberg *et al.*, "The challenge of non-technical loss detection using artificial intelligence: A survey," *arXiv preprint arXiv:1606.00626*, 2016.
- [3] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni, and R. C. Green, "High performance computing for detection of electricity theft," *International Journal of Electrical Power & Energy Systems*, vol. 47, pp. 21–30, 2013.
- [4] J. Nagi, K. S. Yap, S. K. Tiong *et al.*, "Improving svm-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Transactions on power delivery*, vol. 26, no. 2, pp. 1284–1285, 2011.
- [5] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 114.
- [6] C. d. Oliveira, N. Kagan, A. Meffe *et al.*, "A new method for the computation of technical losses in electrical power distribution systems," in *Electricity Distribution, 2001. Part 1: Contributions. CIREN. 16th International Conference and Exhibition on (IEE Conf. Publ No. 482)*, vol. 5. IET, 2001.
- [7] B. C. Costa, B. L. Alberto, A. M. Portela *et al.*, "Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process," *International Journal of Artificial Intelligence & Applications*, vol. 4, no. 6, p. 17, 2013.
- [8] C. C. O. Ramos, A. N. De Souza, D. S. Gastaldello *et al.*, "Identification and feature selection of non-technical losses for industrial consumers using the software weka," in *Industry Applications (INDUSCON), 2012 10th IEEE/IAS International Conference on*. IEEE, 2012, pp. 1–6.
- [9] P. Glauner, A. Boechat, L. Dolberg *et al.*, "Large-scale detection of non-technical losses in imbalanced data sets," in *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*. IEEE, 2016.
- [10] P. Glauner, J. Meira, L. Dolberg *et al.*, "Neighborhood features help detecting non-technical losses in big data sets," in *3rd IEEE/ACM International Conference on Big Data Computing Applications and Technologies (BDCAT 2016)*, 2016.
- [11] J. Meira, P. Glauner, R. State *et al.*, "Distilling provider-independent data for general detection of non-technical losses," in *Power and Energy Conference at Illinois (PECI), 2017*. IEEE, 2017.
- [12] T. Harford, "Big data: are we making a big mistake? ft magazine," <http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html>, 2014, [Online; accessed January 15, 2016].
- [13] C. Cortes and M. Mohri, "Domain adaptation and sample bias correction theory and algorithm for regression," *Theoretical Computer Science*, vol. 519, pp. 103–126, 2014.
- [14] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2001, pp. 26–33.
- [15] J. R. Quinlan, "C4. 5: Programming for machine learning," *Morgan Kaufmann*, p. 38, 1993.
- [16] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
- [17] F. J. Martin, "A simple machine learning method to detect covariate shift," <https://blog.bigml.com/2014/01/03/simple-machine-learning-to-detect-covariate-shift/>, 2014, [Online; accessed January 15, 2016].
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [19] Met Office, *Cartopy: a cartographic python library with a matplotlib interface*, Exeter, Devon, 2010 - 2015. [Online]. Available: <http://scitools.org.uk/cartopy>
- [20] A. Baddeley, I. Bárány, and R. Schneider, "Spatial point processes and their applications," *Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004*, pp. 1–75, 2007.