DATA MINING ON TECHNICAL AND CUSTOMER SERVICE DATA OF A BRAZILIAN DISCO TO INCREASE CUSTOMER SATISFACTION

Luciano Cavalcante SIEBERT
Lactec Institutes – Brazil
luciano.siebert@lactec.org.br

Eduardo Kazumi YAMAKAWA
Lactec Institutes – Brazil
eduardo@lactec.org.br

Eunelson José da SILVA Júnior
Lactec Institutes – Brazil
eunelson.junior@lactec.org.br

Lucio de MEDEIROS
Lactec Institutes – Brazil
lucio.medeiros@lactec.org.br

Angela CATAPAN
COPEL – Brazil
angela.catapan@copel.com

ABSTRACT

Customer satisfaction on power utilities is related to several factors and not easily predict nor understandable. This paper presents the development of a computational system to visualize, analyse and predict data related to residential customer satisfaction. Data mining techniques were used to develop regression models to predict satisfaction indices in moments different from a yearly field survey, therefore enabling managers to take informed, proactive actions.

INTRODUCTION

The utility industry is experiencing an intensive growth on the number of bytes that flow following and/or supporting power grids. In the smart grid, these data will come from several sources such as the billing process, advanced metering infrastructure (AMI), transmission and distribution operation centers, maintenance centers, customer's premises, distributed generation, electric vehicles, among many others. This unprecedented access to an immense volume of data, both historical and in real-time, will provide nearly instantaneous insights on how the electrical system and its consumers behave and relate, enabling improvements on the grid [1].

The aforementioned improvements should increase customer satisfaction, i.e. the consumer response to the perceived discrepancy between their expectations and the services the utility actually delivers. Within the power sector, understanding and assuring customer needs have not been a priority, since many markets residential customers are captive (the utility is assigned only due to the customer's geographical location).

Customer satisfaction is related to several factors, for instance, in [2] image of a service provider, the loyalty of consumers, consumer expectations, perceived value, perceived quality and the way complains are handled are considered some of the most important components of a customer satisfaction model.

Ref. [3] states that the assessment of customer satisfaction for power utilities starts with the establishment of appropriate drivers for its set of customers. These drivers rely on customers' perceptions on how the company is interacting with them. As an example, first-order drivers with direct impact on customer satisfaction could be service, price, and reliability; while second-order drivers could be customer responsiveness, company reputation, management reputation, and outage frequency and duration.

On Brazil, a yearly mandatory evaluation of customer satisfaction for all distribution companies (DISCOs) must be performed, through a field survey, where the companies must achieve a minimum level or face penalties. The survey, which is performed since 1999, has an error margin of ± 4 points with a confidence interval of 95%. The assessed quality areas are power delivery, information and communication, power bill, customer service, company's image, social responsibility and street lighting. For the main indicator of the survey, the satisfaction indicator with perceived quality, the areas social responsibility and street lighting are not taken into account. This evaluation nevertheless does not allow utilities to have a broader view on which factors affect the results, allowing utilities to act only reactively.

As regulation become stricter, game-changing technologies such as distributed renewable generation and energy storage come into practice meanwhile growing requirements in energy efficiency and demand response are presented understanding customer behavior, and satisfaction becomes a priority. For many companies, this will entail a shift in the mindset from a traditional inside-out operations-centric focus to outside-in thinking that starts with the customer's priorities [2].

This paper will present the development of a computational system that helps utilities to visualize in a dashboard data related to customer's satisfaction and the use of data mining techniques to help decision makers take effective actions. All the developments were carried within the Brazilian National Electricity Agency (ANEEL) R&D program and were financially supported by COPEL, the DISCO of the state of Paraná in Southern Brazil, which is responsible for a concession area of 194 thousand km² with ca. 3.5 million residential consumers. COPEL won the Brazilian's customer satisfaction award five times between 2011 and 2016, selected by the customer through the previously mentioned survey.

DATA MINING

Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive models from large-scale data [4]. It is either treated as a synonym for another

popularly used term, knowledge discovery from data (KDD), or it can be viewed as merely an essential step in the process of knowledge discovery. The steps for knowledge discovery are [5]:

1. Data cleaning;
2. Data integration;
3. Data selection;
4. Data transformation;
5. Data mining (application of intelligent methods to extract data patterns);
6. Pattern evaluation;
7. Knowledge presentation.

The data mining process comprises the core algorithms that enable fundamental insights and knowledge from data sets [4]. It is important to mention that the results of the data mining are strongly related to all other steps of the KDD. Therefore a solid analysis and development are necessary for all steps.

Data mining is an interdisciplinary field merging concepts from areas such as database systems, statistics, machine learning, artificial intelligence, and pattern recognition. This knowledge discovery and data mining process tend to be highly iterative and interactive.

According to [6], data mining offers solutions with low complexity and high-performance computing for challenging problems in many fields of the power system, including stability analysis, detection failure prediction, load prediction, and visualization of the power system. The four major areas of data mining for power systems considered in this study are:

- Data visualization: Helps the operators to intuitively monitor the state of the large-scale data provided by devices in the network;
- Clustering: Planning, identification of congested power lines, load forecasting, and event detection;
- Outliers detection: Detection of abnormalities and irregularities in the power system and its sensors;
- Classification: Classification of disturbances, planning of grid expansion.

The use of data science and predictive analysis will have a transformative impact on the power sector. With such vast amount of data coming from the grid, it becomes possible to break some paradigms of the power and energy community, treating the customers as “real clients”, taking into account their needs, behavioral aspects and how different aspects of power delivery and customer service affect them.

CONCEPTUAL MODEL

This paper presents the development of a computational system for power utilities to visualize, analyse and predict data related to residential customer satisfaction. It will be tested and validated in a case study with COPEL, the DISCO of the state of Paraná in Southern Brazil, which is responsible for a concession area of 194,000 km² with ca. 3.5 million residential consumers.

Fig. 1 presents the conceptual model of the system, which will be discussed on the remainder of this paper.

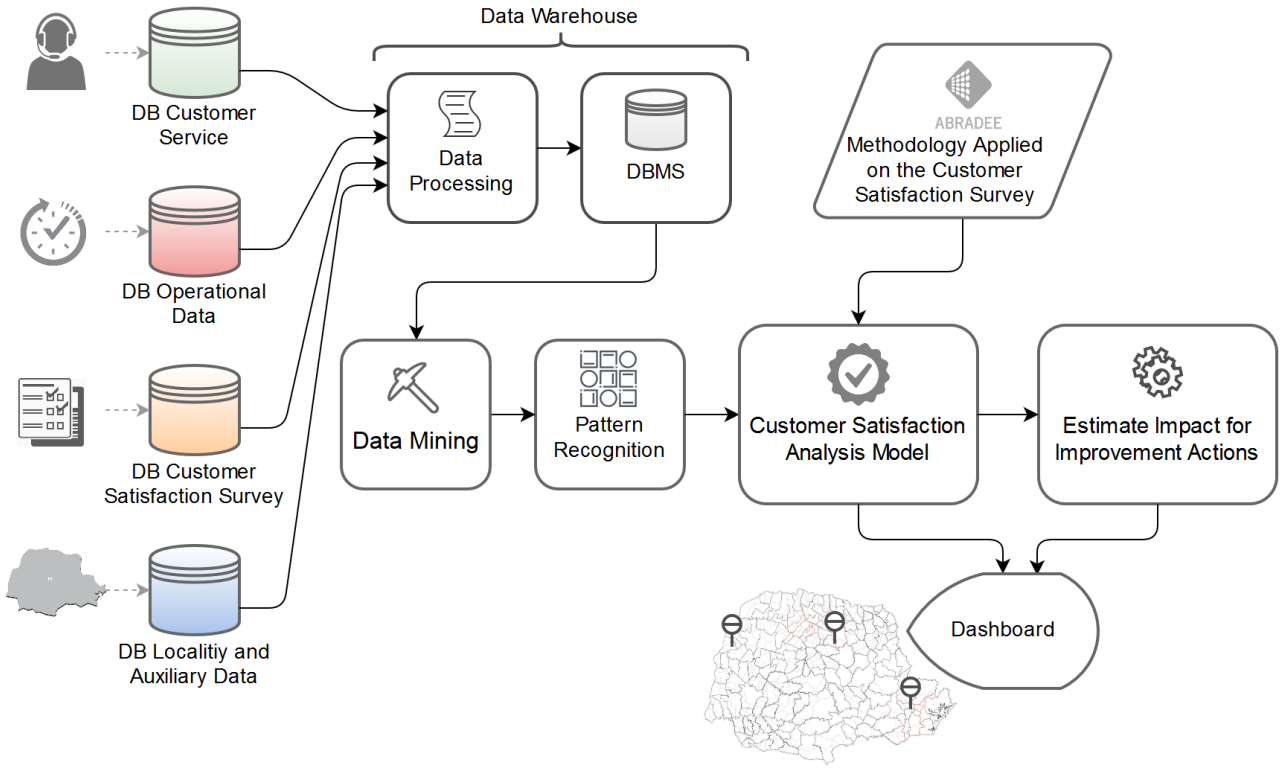

Fig. 1 –Conceptual model

Data Warehouse

To introduce a program for data analysis it is essential to understand in details the desired and available data as well as the business value of such data. It should always be evaluated regarding privacy issues, public available data, value, applicability and possible aggregations [1].

In this work, it was necessary to create a data warehouse to aggregate, in a unified scheme, the large volume of data collected from various information systems of the power utility. These data have varying formats, patterns, and periodicities. The collected data goes through cleaning routines, which corrects or ignores inconsistent data, and transformation routines, which, for example, changes the units of measure and date formats when needed. Also, to reduce data volume, in some cases, it is possible to store only the results of the data aggregation.

A very important aspect to be considered in the relationship between the data is their geographical location, so that one may assess the regions where the variation of the consumer satisfaction index occurred. The COPEL concession area is divided into 20 Maintenance Departments represented by different colors in Fig. 2. These departments may be further divided into cities (394) or operational areas (135).

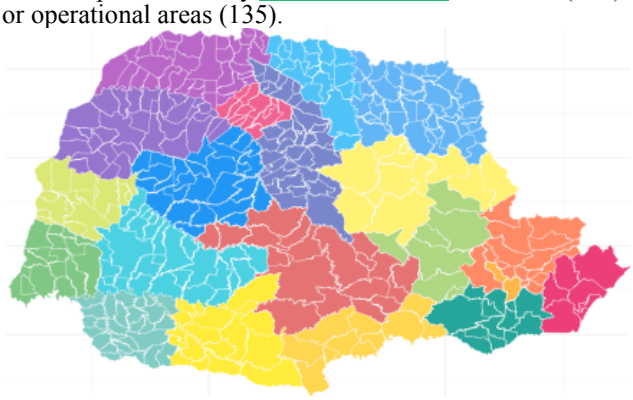

Fig. 2 – Map of the DISCO concession area

A conceptual data model was created to aggregate and organize all available data (Fig. 3). This model represents the business rules and specifies the information that will be kept in the system. The model presents the entities (boxes), the relationships between the entities (lines) and the main attributes (small circles).

The entities of the data model on Fig. 3 are related to the databases of the conceptual model of the system (Fig. 1) by its color. The data sources used were:

- Customer Service: daily total of services provided to the residential consumers, aggregated by request type, city and service channel (call center, agencies, online and mobile);
- Power Outage: power outages, including its location, duration, number of affected consumers, affected electricity demand, type, and cause;
- Reliability Indices: monthly values of reliability indicators such as SAIDI (System Average Interruption Duration Index) and SAIFI (System Average Interruption Frequency Index);

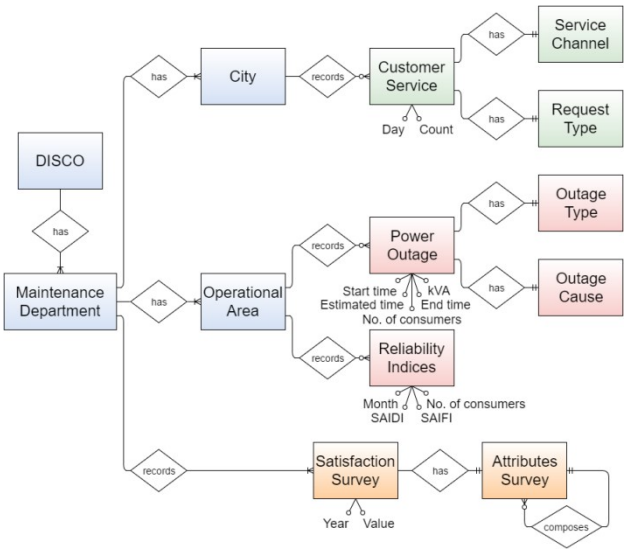

Fig. 3 – Conceptual Data Model

- Satisfaction Survey: annual field assessments of residential consumers satisfaction levels with the quality of the product and the services provided by the distributor, stratified by maintenance departments. The survey attributes are displayed ion a hierarchical fashion, considering seven quality areas, which can be aggregated to a global indicator of satisfaction with perceived quality.

Obtaining data from other sources, such as financial data (e.g. billings, defaults), was not feasible, since they are managed by an outsourced company. Additional data were necessary to enable all databases to be within the same time base and aggregate by locality.

This data warehouse will be updated daily through the execution of automatic data loading procedures. Table 1 summarizes the current volume of the data warehouse.

Table 1 - Summary of data warehouse size

Data source	Instances	Periodicity	Attributes
Customer Service	8,401,460	Monthly	113
Power Outage	2,099,777	By occurrence	7
Reliability Indices	8,670	Monthly	3
Satisfaction Survey	7,753	Yearly	66

After collecting, transforming, cleaning and consolidating all data, they were organized into a data matrix, since it facilitates the exposition of the subject under study and allows numerical calculations in an orderly and efficient way.

Data mining and pattern recognition

A regression model was developed to predict the satisfaction indicators in moments different from the yearly field survey, therefore enabling managers to take informed proactive actions. The model considers the already available operational and customer service data to predict customer satisfaction indices for the present moment in time. Regression analysis is a statistical

methodology for predicting value of one or more response variables (dependent) from a collection of predictors (independent) [7]. The independent variables are the customer service and operational data, and the dependent variables are the customer satisfactions indicators, defined by the yearly survey.
The indices are geographically distributed per maintenance department and can be aggregated for the DISCO. Besides that, the low-level indicators can be aggregated to a general indicator of satisfaction with perceived quality. Considering the KDD steps previously described, some actions were taken to transform the data into useful information for the regression model.
A correlation analysis was performed for the input data, which showed that 20.57% of the variables compared have a linear correlation level greater than 0.7, as illustrated in Fig. 4.

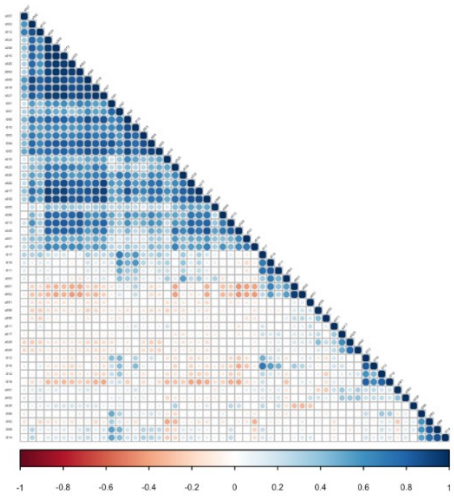

Fig. 4 – Correlation analysis of the independent variables

This multicollinearity leads to the need of further statistical treatment on data, to allow the building of adequate regression models since if they are built with highly correlated data increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. Therefore, a principal component analysis (PCA) was performed on the normalized input data; as well as a manual data aggregation and selection.
Since the satisfaction survey performed (dependent variable) is performed for each maintenance department region, where different people with a considerably different point of views, financial situation and behavioral aspects reside, it is not expected that they would have the same response to the same input. Indeed, many exogenous variables affect the result, and the same input data may result in considerable different outputs (satisfaction indices) for different departments. Since the amount of data provided in the data matrix is not abundant enough to allow the building of a different regression model for each maintenance department area, a cluster analysis was performed to understand which ones behave in a more similar way. Specifically, a k-means algorithm was used.
Finally, for the regression models several linear and non-linear regression techniques were tested with the real historical data, to assess which one would present the best metrics (R-squared, MSE, and MAPE). Namely, the regression models tested until the present moment were: linear regression, stepwise linear regression, polynomial regression, regression trees, machine ensembles (LSBoost and bag) and partial least squares (PLS) regression.

PRELIMINARY RESULTS

The computer system has several functionalities such as data import, data management, satisfaction prediction, executive graphics and monitoring of data evolution through a dashboard.
Data visualization is also a very important aspect of the system since until now they were not integrated into a unique platform. It allows executives to create a specific report, comparative graphics and to observe data evolution.
The system dashboard presents a map of the concession area, signaling the alert points where there are trends in the satisfaction levels and therefore needs further attention. Also, we present the stratified satisfaction predictions by area, the most influential indicators and graphs of the evolution of the indicators.
The interfaces are built with a Responsive Web Design approach, which can be viewed on any screen size, making the system accessible by any electronic device that has an Internet browser, as shown in Fig. 5. The dashboard will be presented on a TV screen located on the DISCO department responsible for the area, and the mobile access will allow a more flexible access to the data.

Fig. 5– Prototypes screens with responsive web design

Satisfaction analysis model and decision-making

Fig. 6 illustrates how a satisfaction indicator is presented on the dashboard. A colour scheme presents the results, which may be compared to fixed value. When the user selects a given department, the system presents the indices that compose the indicator analysed.

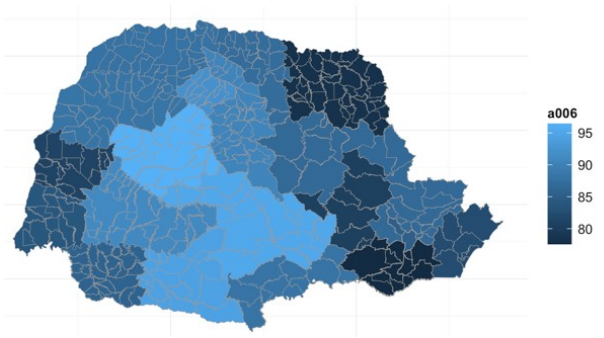

Fig. 6 – Dashboard with prediction for a given satisfaction indicator

The system also points out which operational, or customer service data influence the most each result, according to the regression models, allowing the utility to create efficient customer relationship actions that influence customer's satisfaction. For example, if there is a worsening on a given satisfaction indicator, the executives will have the information why it presumably will happen. Its main expected contribution is to allow DISCOs to understand the relation between technical and customer service data with customer satisfaction and therefore create a business model that is both profitable for the company and satisfactory for customers.

CONCLUSIONS

This paper presented the conceptual model of a computational system that allows power utilities to predict, assess and therefore improve customer satisfaction. Data mining techniques were used to transform the database, and a regression model was developed to predict customer satisfaction indices from customer service and operational data.

Preliminary results show a potential in effectively helping decision makers in power utilities to take proactive actions considering customer satisfaction as a drive. Future works should include the concept of big data using different data sources such micro and macroeconomic data, billing, social media, among others. Future publications will present the impact of the developed system for the power utility on a field test.

REFERENCES

[1] C. L. Stimmel, 2015, *Big Data Analytics Strategies for the Smart Grid*, CRC Press, Boca Raton, United States.

[2] J. Mutua, D. Ngui, H. Osiolo, 2012, "Consumers satisfaction in the energy sector in Kenya", Energy Policy, n. 48, 702-710.

[3] J. Elliott, C. Serna, 2005, "Managing customer satisfaction involves more than improving reliability",*The Electricity Journal*, v.18, n.7, 84-89.

[4] M. J. Zaki, W. Meira Jr., 2014,*Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, New York, United States.

[5] J. Han, M. Kamber, J. Pei, 2012, *Data Mining: Conecpts and Techniques*, Morgan Kaufman Publishers, Waltham, USA.

[6] M. Kazerooni, 2014, "Literature Review on the Applications of Data Mining in Power Systems", *Power and Energy Conference at Illinois*, PECI.

[7] R. A. Johnson, D. W. Wichern, 2007, *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey, United States.