redin Revista Facultad de Ingeniería, Universidad de Antioquia, No. 84, pp. 60-71, 2017

Identification of the characteristics incident to the detection of non-technical losses for two Colombian energy companies

Identificación de las características incidentes en la detección de pérdidas no-técnicas para dos empresas comercializadoras de energía colombianas

Carmen Cecilia Sánchez-Zuleta, Juan Pablo Fernández-Gutiérrez, Carlos César Piedrahita-Escobar

Departamento de Ciencias Básicas, Universidad de Medellín. Carrera 87 # 30 - 65. C. P. 050026. Medellín, Colombia.

ARTICLE INFO

Received December 15, 2016
Accepted August 01, 2017

KEYWORDS

Non-technical losses, MDS, cluster, Benford's Law, decision trees

Pérdidas no-técnicas, MDS, clúster, Ley de Benford, árboles de decisión

ABSTRACT: The study of non-technical losses affecting energy trading companies has guided the researchers' perspective on different techniques and tools that allow them to detect, and why not, to forecast such losses. In the search for a solution to the problem, the different researchers rely on variables that, in many cases, the same marketing companies, from their practical experience, have been considered as incidents in the identification of the problem. However, most of the studies carried out do not support their solutions with the fact that each trading company retains particular data in which both, technical and socio-economic characteristics recorded, are not necessarily shared in their databases. In this work, we follow up on some of the characteristics registered by two Colombian energy trading companies, which serve two different regions of the country in terms of topography and idiosyncrasy. In particular, attention is focused on two characteristics measured in both companies, which by their nature, will always be on the data of any energy trading company: Consumption in kWh, and the period, measured in months. For this purpose, Benford curves analysis, MultiDimensional Scaling (MDS), and hierarchical cluster will be implemented. Finally, it will be studied if the incidence of the variables visualized in the studies presented is reflected in the decision tree model.

RESUMEN: El estudio de las pérdidas no técnicas que afectan a las empresas comercializadoras de energía ha orientado la mirada de los investigadores hacia diferentes técnicas y herramientas que les permitan detectar y pronosticar dichas pérdidas. En la búsqueda de una solución al problema los diferentes investigadores se apoyan en variables que, en muchos casos, las mismas empresas comercializadoras, desde su experiencia práctica han determinado como incidentes en la identificación del problema. Sin embargo, la mayor parte de los estudios realizados no anteponen a sus soluciones el hecho de que cada empresa comercializadora registra en su conjunto de datos una serie de características, tanto técnicas como socioeconómicas, que no necesariamente comparten entre ellas. En este trabajo se hace seguimiento a algunas de las características registradas por dos empresas comercializadores de energía colombianas, las cuales atienden dos regiones diferentes del país en cuanto a topografía, e idiosincrasia. De manera particular, se centra la atención en dos características medidas en ambas empresas, y que, por su naturaleza, siempre estarán en los datos de cualquier empresa comercializadora de energía, El Consumo en kWh, y el Periodo, medido en meses. Con este propósito se implementarán análisis de curvas Benford, escalamiento multidimensional MDS, y clúster jerárquico, para finalizar estudiando finalmente si la incidencia de las variables visualizada en los estudios planteados se refleja en el modelo de árboles de decisión.

* Corresponding author: Carlos César Piedrahita Escobar
e-mail: cpiedrahita@udem.edu.co
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.n84a08

1. Introduction

The loss of power is a serious problem that affects both the distribution companies and the transmission and marketing of energy. However, the economy of the trading companies is mainly affected by this fact, since in their distributions the networks suffer by technical losses as well as by non-technical losses, the two kinds of losses that affect an electrical system.

In general, in the electric sector, technical losses are understood to be caused by physical effects, for example the Joule effect, that is to say, by the physical properties of the components of the power system, as happens with the power dissipated in the transmission lines [1], and [2]. On the other hand, non-technical losses are those that are generated by the intervention of the electrical connection or of the electric meter [3-6], or when failures occur in operating systems that are not detected in a timely manner, as well as errors in meter readings and billing.

As the nature of the loss problems is different, they are commonly treated independently. Technical losses, due to a response to a physical phenomenon, can be predicted with a low error rate and can be reduced by improving facilities [1]. However, it does not occur with non-technical losses. The human nature of the problem, the economic conditions that are not easy to observe and control, and their variability make the detection of this type of loss not so immediate, a problem to solve [7].

In [8] a definition of fraud was presented, from which five features are associated with this phenomenon. Three of them clearly show the complexity of the problem. The first of these characteristics speaks of having a not very frequent phenomenon, which leads to the treatment of unbalanced data sets such as those studied by Glauner [3]. The second describes the phenomenon of fraud as a situation that is hidden in such a way that it is not very perceptible, and the third indicates that fraud techniques evolve over time, that is, we have a dynamic problem [8].

In general, for the trading company, both technical and non-technical losses are energy quantities that are being distributed and not being billed, affecting the company's economy. Non-technical losses are an important variable in the impairment of the economy of a country, producing loss profits, stability in the power network and environmental damage [4].

The economic impact caused by energy losses, and in particular by non-technical losses, is reaching alarming figures in many countries around the globe. According to Glauner, estimates of losses can reach up to 40% of total electricity distributed in countries such as Brazil, India, Malaysia or Lebanon. In addition, this is a scourge that not only affects developing countries. Developed countries such as the United States or the United Kingdom present estimates of non-technical losses ranging from 1-6 million euros [4]. In India, losses in transmission and distribution systems were around 15% in the period 1966-67, and have increased gradually to 28.36% between 2011-2012 [9].

We could continue taking a more detailed look at non-technical loss rates in different parts of the world, such as those reported for the southern region of Jordan in the study by [10], where [6] and [11] seek to reduce these indices in the region using vector support machine models. However, this entire scenario leads to evidence of the need for energy distributors and marketers to make a timely and effective intervention of their losses. In particular, those related to cases of theft or systems failure because these are the ones most affecting their economy.

In Colombia, in addition to the economic losses, that represent non-technical losses for the trading company, these companies must assume the sanctions incurred by the Energy and Gas Regulatory Commission (CREG, acronym in Spanish) if they do not commit to this organization with a loss reduction plan [5]. It is expected that the company can reduce its energy losses to an efficient level, which will consist of technical losses and recognized non-technical losses (article 387 of the 2007 CREG). This takes us to the conclusion stated in the previous paragraph, of an imminent need to implement a loss detection plan, and in particular of the non-technical ones.

In general, it can be said that there are many and varied techniques that have been implemented around the world to detect, and even predict the non-technical losses of an energy company. However, as already mentioned, different factors that intervene in this class processes make this difficult. These factors constitute the characteristics of the problem, which are decisive when choosing and implementing a model to solve the problem.

Conventionally, energy companies store information from their users that goes beyond the monthly consumption kWh as for example, information related to technical aspects such as meter type or reading period. In addition, other aspects of the social type, such as the socioeconomic stratum, or of a personal nature such as address of the property, telephone, and more. All these elements are characteristics that relate to the user, and in a transversal way with their regular or irregular behavior.

Many papers found in the literature describe the implementation of possible solutions to the problem of non-technical losses. Some based on probabilistic models [12], spatial models [13], and others have been oriented from a more analytical or data mining perspective [4], [6] among others.

However, the purpose of this article is not to present a new solution to the problem. Rather the objective is to present a review of the cross-behavior of the characteristics included in the set of users for two energy trading companies. We identify in each one of the companies those characteristics that present a behavior, especially incident on the objective variable (hereinafter variable irregularity), as well as

its performance in the implementation of the models of decision trees and K-means.

2. Formulation of the problem

The search for a standard solution to the problem of non-technical losses in different energy trading companies has given us a new problem that would change the initial idea of a single solution. This problem appears with the simple exploration of the databases associated to each one of the companies under study. This first approach has shown that the characteristics that the company registers of a user: consumption, lighting value, period, among others; socio-economic: stratum, address, telephone and more, are not necessarily the same from one trading company to another. In addition, the suspicion arises that the proximity between the characteristics, and between the characteristics and the target variable, of a trading company, does not necessarily remain in a second company.

An important exercise is to find the subset of characteristics associated with a specific trading company, which maximizes the accuracy of the classification model used, and compare them with those variables that maximize the accuracy of the same model in a second company. In this situation, clearly, the non-coincidence of the influence of the same subset of characteristic would show that the problem of detecting and forecasting non-technical losses has no standard solution, and beyond that, there could be evidence of the influence of socioeconomic variables on the target variable.

According to [14], the problem of determining the set of characteristics that optimizes the model results is an open problem. However, in this paper, a case study is sought to show that the best analytical solution to the problem of losses for a trading company is related to the incidence variables and that these can vary from company to company.

Since the purpose of the research is to identify the characteristics of the electric sector users, who incur in irregularities of some kind in their consumption, this report intends to identify the variables that are more closely related or have a higher correlation with the variable of interest (irregularity). With this objective, it is sought to perform a multidimensional scaling to the data set that allows visualizing in some way the dependence between the variables that have been considered, based on the frequency that they present. In addition, the k-means models and decision trees are implemented for each of the data sets under study.

In light of the problem, the behavior of consumption in both companies will be studied and compared by statistical analysis according to the variables of social stratification of users, and given the characteristics of the database two of them will be considered in this case. Also, the consumption behavior, according to the variable of geographic type "zone", is always in relation to the objective variable.

A set of tools will be used to determine the variables incident on the response variable, so that a reduction of the dimension of each of the data sets can be achieved by implementing a multidimensional scaling, this forms a decrease in the estimated execution times of the models implemented. Additionally, some descriptive techniques are recommended to detect anomalous situations in a data set that enhances the selection of the mentioned characteristics. Finally, an implementation of decision trees and vector support machines is carried out, which allows us to determine the efficiency of the selection of the variables by means of a comparative analysis among the models obtained with the complete base and with the reduced set.

2.1. Nature of the data

In order to solve the problem, we have considered the databases of two Colombian companies, energy marketers, which serve two regions of the country that are culturally and topographically different. For this purpose, we will call them Company 1 and Company 2.

To get an idea of the sociocultural characteristics of the users of each of these companies, it will be said that Company 1 serves a region with a relatively homogeneous idiosyncrasy, and with an urban topography, which changes from municipality to municipality. However, it does not present abrupt jumps. Company 2, for its part, serves a more heterogeneous area, both cultural and geographical.

The data used are taken from the bases of the respective companies obtained from the complete set of each one of the companies, keeping the total number of users and keeping those variables of interest. Thus, for the two companies, there are three common characteristics: Consumption in kWh, which records the consumption in kWh per user in each period, the period or date in which the measurement is made, the account number, the socio-economic variable, and the objective variable, which in this case will be called Irregularity (Irreg). Each of the companies also has an extensive list of variables that complement their data, not all of them registered in the two companies, so for Company 2 we will see additional characteristics of its structure, such as zones, reading cycles, among others.

As non-technical losses are due to both fraud and infrastructure failures, the variable response has been called irregularity (Irreg) for the two case studies that have been carried out. It registers all users of the system that were visited by the company inspection group. It is a dichotomous variable that is marked with one or "True" if in the visited site a non-technical loss of any type was found, and with 0 or "False", if no abnormality was found.

Finally, it is worth highlighting that the problem being solved is oriented to the detection of fraud, according to the fraud feature proposed by [8]. This is a low-frequency event, it is then that the two sets are unbalanced, with a greater proportion of regular users than irregular ones.

3. Techniques and models

The following lines describe the statistical techniques that were implemented with the objective of determining the dependence or proximity between the variables. A description is also given of the mining models that are implemented in the solution of the problem.

3.1. Multidimensional scaling - PROXCAL analysis

Multidimensional scaling, hereinafter called MDS, is a multivariate technique that allows us to represent the measures of similarity (or dissimilarity) between a set of objects or subjects or variables as distances between two points in a space of low dimension [15]. It is a tool that, besides allowing us to establish similarities between objects, also allows us the evaluation of the relation between the variables that are measured in a set of data [16].

Basically, what the MDS does, is to take the item similarities and assign them a location in a low-dimensional space (usually it is intended to be in two or three dimensions), and through a series of iterations, the best representation is sought to guarantee an optimal solution to the problem [17].

When an MDS study is advanced, it is important to keep in mind the following essential requirements in the development of a multidimensional scaling analysis:

- A set of numbers, called proximities or similarities, that expresses all or most combinations of similarity pairs within a group of objects.

- Provide an algorithm to carry out the analysis.

The procedure, in very general terms, follows some basic ideas in most scaling techniques. The starting point is a dissimilarity matrix Δ among n objects p-dimensional in study, whose elements are given by δ_{ij}, which represents the dissimilarity of object i to object j. The number of dimensions k is set, and proceed as follows:

- Let n objects be in an initial configuration with k coordinates each one. In other words, each object has the coordinates $(x_1, x_2, ..., x_k)$ that are in a space of dimension k.
- Calculate distances d_{ij} between objects i and j, depending on the measure chosen.
- Perform a regression of d_{ij} on δ_{ij}. This regression can be linear, polynomial, or monotonic. For example, if you consider linear, you have the model (1)

$$d_{ij} = a + b\delta_{ij} + \varepsilon, \tag{1}$$

and using the least squares method, we obtain the estimates of the coefficients a and b, and from this can be obtained (2) what is generically known as a "disparity"

$$\hat{d}_{ij} = \hat{a} + \hat{b}\hat{\delta}_{ij} \tag{2}$$

If a monotonous regression is assumed, there is not exact relation between d_{ij} and δ_{ij}, but it is simply assumed that if δ_{ij} increases, then d_{ij} increases or remains constant.

A measure of similarity that is interesting for data of mix type, this our case, it has the characteristic of measuring the distance between both categorical and quantitative variables, is presented by Gower [17] and is given by (3):

$$S_{ij} = \frac{\sum_{r=1}^{p} w_{ijr} S_{ijr}}{\sum_{r=1}^{p} w_{ijr}} \tag{3}$$

where S_{ijr} is the similarity between objects i and j under the r-th variable only, and ω_{ijr} is unity if the i-th and j-th objects can be compared on the variable r, and zero in other cases.

- The goodness of the fit between the distances of the configuration and the disparities derives from the definition of the so-called stress index (STRESS). The most commonly used stress criteria are:

$$STRESS = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^{\,2}}} \tag{4}$$

and,

$$S - STRESS = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^{\,4}}} \tag{5}$$

The disparities depend on the type of regression used in the third step of the procedure.

The STRESS, levels indicates the goodness of fit of multi-dimensional scaling model, can be interpreted as in Table 1:

Table 1 Interpretation of STRESS

Size of STRESS	Interpretation
0.2	Poor
0.1	Regular
0.05	Good
0.025	Excelent
0.00	Perfect

- The coordinates $x_1, x_2, ..., x_p$ of each object are changed slightly so that the adjustment measurement is reduced.

The PROXCAL is a computational program designed to implement MDS and scaling in individual proximity differences. This is a consolidated version of the old SMACOF series.

3.2. Decision trees

A decision tree is a model of prediction or classification whose main objective is the inductive learning from observations and logical constructions. In general, the models of classification and prediction of patterns capture the relation existing among attributes variables $x_1, x_2, ..., x_p$ with objective variables $y_1, y_2, ..., y_q$. Consequently, we have that decision and regression trees are used to learn about classification and prediction patterns from the data, and to express the relationship among the variables x with a target variable y, with $y=f(x)$, expressed in the form of a tree. A decision tree is conceptually simple, user-friendly, computational speed, robustness in relation to missing data and outlier points, and mainly the interpretability of the rules generated [18], [19]. Depending on the nature of the target variable, a decision tree can be either classification or regression.

Graphically, a set of nodes, leaves and branches represent a tree. The head or root node is the attribute from which the classification process starts. The internal nodes correspond to each of the questions about the particular attribute of the problem. Each possible response to the questions asked is represented by a child node. The branches that come out of each of these nodes are labeled with the possible values of the attribute. The final nodes or leaf nodes correspond to a decision, which coincides with one of the levels of the target variable.

A tree starts by generating its root node, choosing a test attribute, and splitting the training set into two or more subsets. For each partition, a new node is generated and so on. When objects of more than one class of the target variable are present in a node, an internal node is generated. When it contains objects of a single class of the target variable, a leaf is formed which is assigned the label of the class. Decision trees provide readable rules, instructions of the form "If A, then B". These rules allow descending through the tree from its root node until some leaf is reached; when a new object is given, this can be classified in one leaf of the tree, which has the appropriate level of the target variable.

Methods of selection of the partition

In order to find a decision tree with the minimum description length, one must know how to divide a node to achieve the objective. For a dataset, there are as many ways to split the root node as individual attribute variables. To decide which selection of the division to perform, we must consider the variable, which offers greater homogeneity in the resulting groups. A homogeneous data set is a set of data whose data records have the same target value. There are several measures to determine the homogeneity of the data, among them are Entropy and the Gini index.

Entropy

Entropy is a measure of the homogeneity of the set that takes values in the interval $[0, \log_2(c)]$, with c the number of levels assumed by the target variable. It is originally introduced to measure the number of information bits needed to encode the data, and is defined as:

$$entropy(D) = \sum_{i=1}^{c} -P_i \log_2 P_i \quad (6)$$

Where: $-0\log_2 0=0$ by definition [18], and $\sum_{i=1}^{c} P_i = 1$.

- D denotes the given data set. When we are at the root node, D is the training data set

- C is the number of different levels of the target variable

- P_i is the probability that a data in the data set considered takes the i-th value of the target variable.

Gini coefficient

Another index of the measure of the homogeneity of a data set is the "Gini index", it is defined as follows:

$$gini(D) = 1 - \sum_{i=1}^{c} (P_i)^2, \quad (7)$$

with P_i and D as described for the case of entropy.

Regardless of the homogeneity index chosen, it is understood that the smaller the homogeneity coefficient, the more heterogeneous the group will be.

3.3. Hierarchical cluster

Cluster analysis is a multivariate technique whose main objective is the search for a structure that allows grouping the data set according to its characteristics; That is, the purpose of cluster analysis is the assignment of a set of objects into groups called clusters, so that two objects in a same group are more like each other than two objects that are in different groups.

Within cluster analysis, two main approaches can be determined which frame the two different methods implemented, namely:

- Hierarchical Clustering
- Fuzzy Clustering

Hierarchical cluster begins with the calculation of the distance matrix between objects. In this methodology, the groups are formed in the form of agglomerates or by division process. The hierarchical method has the characteristic that locates strongly each of the units of study within a cluster, that is, by a bivalent logic, i.e., a unit of study is or is not within a cluster.

One of the measures that allows us to determine the best cluster model for a data set within a hierarchical cluster is the Coefficient variable. This coefficient is widely used to compare different methods, it is defined as the correlation between the $\frac{n(n-1)}{2}$ elements of the upper part of the matrix of proximities or distances in front of the cophenetic matrix C, with elements C_{ij}, which are defined as those that determine the proximity between the elements i and j when these are joined in the same cluster. The cophenetic coefficient gives a measure of how to choose which method gives the best results, so that the method with a higher coefficient will be the one with the least distortion in the original relationships in the elements.

3.4. Benford´s Law

Benford's law is both a visual and numerical data exploration technique that is particularly interesting in fraud detection. It describes the frequency of occurrence (frequency distribution) of the first digit in many real-life data sets [19], and [8].

The use of this law has proven to be effective in identifying oddities in data. When comparing the expected distribution following Benford's law with the observed distribution in a data set, strong deviations from the expected frequencies may indicate suspicious and possible manipulation on the data. Therefore, this law has been used for case selection in fraud detection.

The mathematical formula describing this law expresses the probability $P(d)$ of the leading digit d to occur to be equal to

$$P(d)=\log_{10}\left(1+\frac{1}{d}\right) \text{ with } d = 1, 2, ..., 9 \quad (8)$$

Benford's law can be used as a screening tool for fraud detection, however, it is important to understand that Benford's Law needs data, which is not entirely random or highly conditioned, but is something in between. The data can be of a great variety and are often the typical result of diverse processes, with many influences, as with most data extracted from natural, social, and economic phenomena.

R and SPSS- Software to be used

For the development of this work, the R software libraries were used, and MDS was performed through the PROXCAL algorithm, that is implemented in the SPSS tool.

4. Principal results

The following results show the behavior of the comparative analysis between the variables of each of the companies, which are expected to allow groupings that lead to the reduction of the initial space dimension.

4.1. Study of Benford curves

As previously mentioned, the graph that derives from the implementation of Benford's law is a very adequate tool to identify rare behavior of a characteristic, in relation to a target variable. [8], and [18] argue that although this is a tool that is not always practical in the case of fraud, this law reflects strange behaviors of a variable.

The graphs presented in Figure 1 and Figure 2 corresponds to the Benford's law for the case of consumption in kWh for each of the companies under study and in relation to the variable irregularity. Note that while Figure 2 for Company 2 conforms satisfactorily to the behavior of the Benford's curve, the frequency of the consumption digits for Company 1 is far from expected.

Applying Benford's Law for Company 1, consumption reflects atypical frequencies for almost all digits and in virtually all cases. It emphasizes however the purple line, which is associated with the presence of fraud within the data. Note that digit four has a relatively frequency higher than that estimated by the law, while the total set and regular cases conform to the expected value. A similar case is found in digit seven, and the behavior of digits five and six on the other hand is lower than the expected frequency.

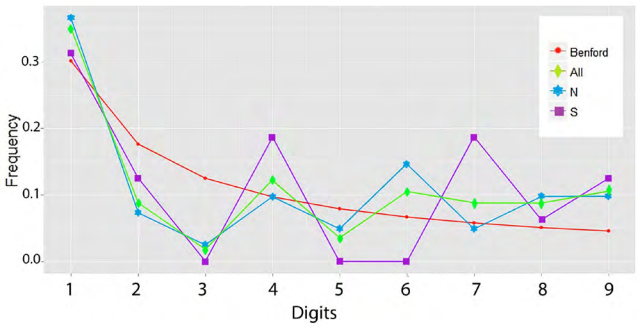

Figure 1 Benford's curve for consumption by Irregularity for Company 1

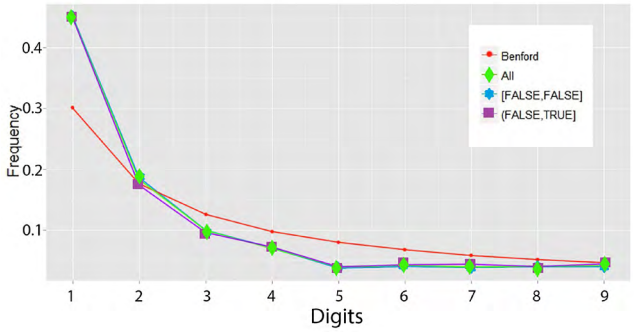

Figure 2 Benford's curve for consumption by Irregularity for Company 2

A comparative analysis of the behavior of energy consumption by the users of the two companies shows that in both cases the frequencies of company one are

above the 30% expected frequency. However, the behavior of company two, for both regular (non-fraudulent) and irregular (fraudulent) users does not reflect a particularly atypical trend. The curve of company 1, on the other hand, reflects apparently adulterated data, since the extraneous frequencies are part of each of the classes considered.

Benford's Law for the variable Period of each one of the companies under study is presented in Figure 3 and Figure 4. We can see again a strange behavior in the digits of Company 1, although it should be noted that the proportion of the digits of the variable Period for this company is preserved in all classes, with a strange behavior for the digit eight. The behavior of the digits of the variable Period (date of consumption) of Company 2, behaves the same in each of the classes. However, it can be seen in each of them a frequency much lower than the proportion of the digit.

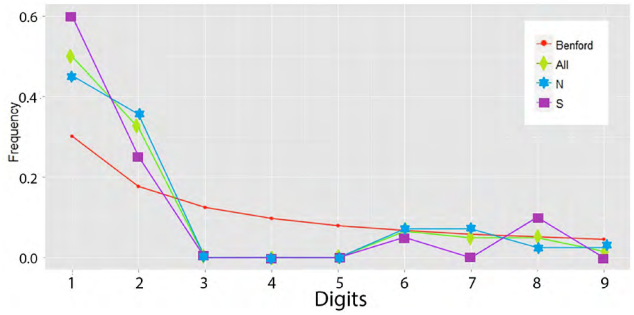

Figure 3 Benford's curve for period by Irregularity for Company 1

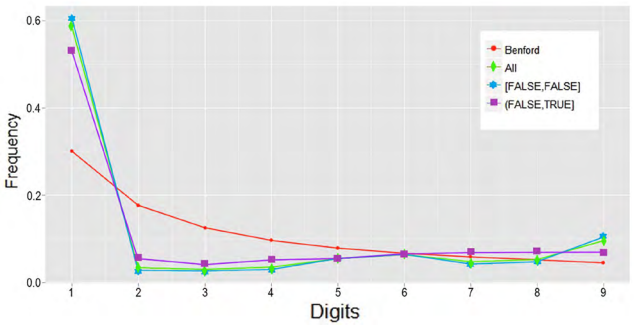

Figure 4 Benford's curve for ConsumptionDate by Irregularity for Company 2

From Figure 1, 2, 3, and 4 we have a first idea of the incidence that have both the variable consumption and the Period in the variable of irregularity, showing that it effectively depends on the company. This is the manifestation of the behavior of this variable.

It is also observed in the mentioned figures that the behavior of these two variables (consumption and period), in relation to the proportion per digit, in the same company, are significantly different. To observe the closeness between these two variables in each company, as well as the proximity they present with the other variables in the data, a multidimensional scaling was carried out through the correlation of the data. The PROXCAL algorithm was used, with which it is expected to observe the dissimilarity or similarity of the variables, and later also a cluster analysis will be applied to identify the most incident variable groups, in order to finally make a comparison of the results.

4.2. Study of MDS

For this study, there was a matrix that registers the value of each of the variables in all users and for each date, and since there are both quantitative and categorical variables, we used the PROXCAL algorithm developed in SPSS, which directly estimates the matrix of distances between the variables. Table 2 and Table 3 show the processing of the cases for each of the companies. It can be observed that the parameters were kept in the implementation of the algorithm for the data of both companies.

Table 2 Summary of processing of cases Company 1

info	comments	value
Cases		8
Sources		1
Objects		8
Proximities	Total Proximities	28[b]
	Missing proximities	0
	Active proximities[a]	28

[a]The active proximities include nonmissing proximities.
[b]Sum of all strict lower triangular proximities.

Table 3 Summary of processing of cases Company 2

info	comments	value
Cases		8
Sources		1
Objects		8
Proximities	Total Proximities	28[b]
	Missing proximities	0
	Active proximities[a]	28

[a]The active proximities include nonmissing proximities.
[b]Sum of all strict lower triangular proximities.

On the other hand, in Table 4 we can observe the proximity matrix between the characteristics in study for each of the companies. This matrix is constituted in the input for the algorithm PROXCAL that was implemented. From this matrix, we can have a first observation of the relationship between the variables. However, the strength of the technique lies in decreasing the dimensionality of the data to better visualize their relationship.

Table 4 Input data for algorithm Company 1, proximities

variables	Month	Monthly Consumption	Customertype	Regulated	Location	Stratum	Irregularity	IrregularESTud
Month	-							
Monthly Consumption	1189.723	-						
Customertype	201.550	1056.999	-					
Regulated	138.803	857.856	90.290	-				
Location	147.686	855.205	101.762	57.763	-			
Stratum	220.402	1107.398	107.865	94.962	100.092	-		
Irregularity	319.272	1087.456	236.583	215.988	215.906	303.096	-	
IrregularESTud	233.723	919.979	197.628	180.652	184.596	230.881	15.025	-

Note in the matrix of Company 1, the high value of the "Monthly Consumption" variable in relation to the other variables. A similar behavior is observed, although not so marked, with the simile variable "Period". This means that when graphing the low dimensions, it is expected that these variables are not very close to the remaining set of variables in this company, as shown in Table 4.

The behavior for Company 2 is not the same, in this case the characteristic that presents the greater distance with the rest of the set is "Period", or date of consumption. The energy consumption for this company is relatively close to the rest of the characteristic, as shown in Table 5.

Table 5 Input data for algorithm Company 2, proximities

variable	Consumptiondate	ConsumptionKWh	Cicle	Countnumber	Anomalitycode	Stratum	Readingzone	Readingcicle	FlagUrban	Irregu
Consumption date	-									
Consumption KWh	3735707341644.6	-								
Cicle	3735707368125.0	41194.640	-							
Countnumber	3734575477600.3	1132091618	1E+009	-						
Anomalitycode	3735707372990.6	45451.499	11117.608	1132122891	-					
Stratum	3735707376392.5	46889.036	8925.676	1132126299	10126.556	-				
Readingzone	3735707375911.5	46568.039	8482.325	1132125816	9978.582	786.155	-			
Readingcicle	3735707368112.9	41191.980	422.875	1132118022	11114.714	8943.88	8501.096	-		
FlagUrban	3735707376735.6	47134.134	9243.648	1132126642	10238.164	475.711	1003.643	9261.386	-	
Irregu	3735707376871.3	47245.896	9363.339	1132126777	10271.641	590.135	1091.983	9381.070	232.916	-

In Table 6 and Table 7, we can observe the STRESS of the scaling for each one of the companies, as well as the dispersion explained in each one of the cases. It is important to remember here that in the previous section it was established that this value validates the goodness of the model, and this model improves as its STRESS value approaches zero. Therefore, conducting an inspection on the values obtained for the measures of the goodness of fit, it can be noted that a Normalized Raw Stress value of 0.009 is available for Company 1 and 0.00009 for the second company, with S-STRESS values of 0.017 and 0.0000 respectively for each company, respectively. It is concluded that the proximity model adjusted by the scaling is excellent in both cases, although a better adjustment, by its values, is highlighted for company two. In this table, it is also possible to observe that the dispersion explained by the models is 99.05% for Company 1 and 99.99% for the second company.

Table 6 Measure of goodness and Stress, Company 1

Measure	Value
Normalized raw Stress	0.00950
Stress-I	0.09747[a]
Stress-II	0.13967[a]
S-Stress	0.01702[b]
Dispersion Accounted For (D.A.F)	0.99050
Tucker's coefficient of congruence	0.99524

PROXSCAL minimize the normalized raw Stress
[a]Factor for optimal scaling = 1,010.
[b]Factor for optimal scaling = 1,012.

Table 7 Measure of goodness and Stress, Company 2

Measure	Value
Normalized raw Stress	.00009
Stress-I	.00946[a]
Stress-II	.01060[a]
S-Stress	.00000[a]
Dispersion Accounted For (D.A.F)	.99991
Tucker's coefficient of congruence	.99996

PROXSCAL minimize the normalized raw Stress
[a]Factor for optimum scaling = 1,000.

The two-dimensional coordinates for each of the variables are presented in Table 4. Note in this table, for Company 1, that the only variables found in the fourth quadrant are the two irregularity variables. However, it is observed in Figures 5 and 6, where the two-dimensional representation of the variables is shown, that with the exception of the monthly consumption, the other variables are quite close to the irregular variable, which constitutes the variable of interest in this study.

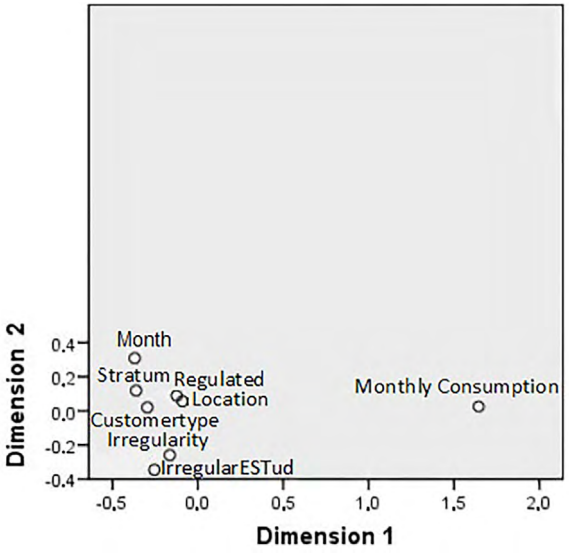

Figure 5 Common space object points Company 1

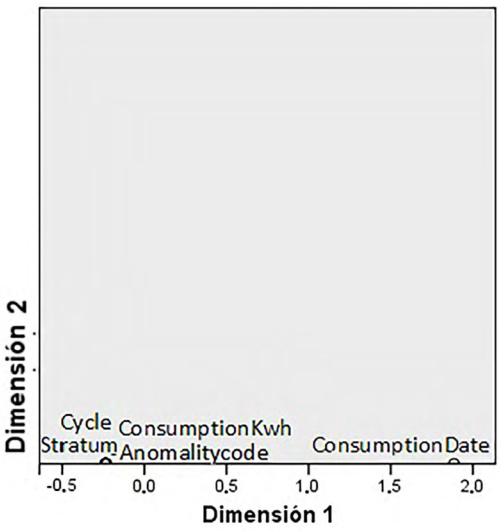

Figure 6 Common space object points Company 2

However, in the behavior of the variables of the second company, it can be observed that the most distant of the group, the most dissimilar, is the date of consumption (see Figures 5 and 6). It is also noted that for this company, consumption is closer to the variable of interest (Irregu), and in general, closer to the remaining group of the characteristics in the analysis. Once again, it has been observed that the behavior of these two variables (Consumption and Period) in the two companies being studied is not the same. Specifically, it can be observed in Figures 5 and 6 that its behavior is opposite in relation to the target variable.

A special behavior of the coordinates for Company 2, see Tables 8 and 9, where the only variables for this company, which show different coordinates from the coordinates of the target variable, are consumption and Period.

Table 8 First two coordinates Company

variable	Dimension	
	1	2
Month	-.370	.308
Monthly Consumption	1.644	.025
Customertype	-.301	.012
Regulated	-.125	.089
Location	-.090	.056
Stratum	-.338	.116
Irregularity	-.255	-.345
IrregularESTud	-.166	-.258

Note also that the difference in coordinates for dimension one is characterized by the Period (consumption date), while for dimension two, the difference is not as marked as the previous one and is caused by the Consumption variable. In general, it can be seen that for this company the difference in the first dimension marks the variable Period (date of consumption), and in the second dimension is determined by consumption.

Table 9 First two coordinates Company 2

variable	Dimension	
	1	2
Consumptiondate	-.370	.308
ConsumptionKWh	1.644	.025
Cicle	-.301	.012
Countnumber	-.125	.089
Anomalitycode	-.090	.056
Stratum	-.338	.116
Readingzone		
Readingcicle	-.255	-.345
FlagUrban	-.166	-.258

For the case of Company 1, it should be noted that based on the first dimension we have that, except for the monthly consumption, the other variables are quite related to the variable of interest. As mentioned before, the opposite happens in Company 2. In relation to the second dimension, we can observe a subtle distance between the variables, but we could speak of a group of variables that are quite related to the variable of interest, these variables are: Stratum, Type of Customer, Regulated, and Location.

4.3. Hierarchical cluster analysis

Continuing the purpose of the article, to identify the characteristics incident on the target variable and its behavior in relation to it, we propose now a hierarchical clustering analysis that allows visualizing its behavior.

Due to computational difficulties, dendrograms were constructed using the SPSS tool for Company 1, and R for company 2, in both cases the nearest neighbor method was used. The graphs are shown respectively in Figures 7 and 8.

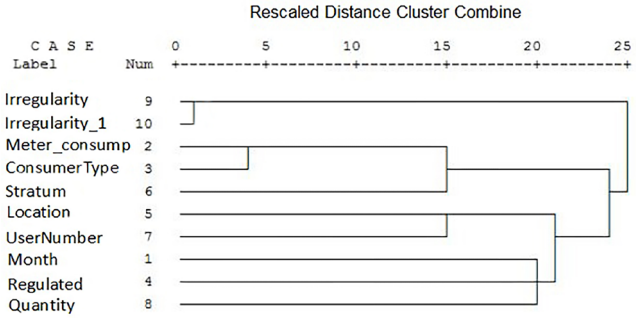

Figure 7 Dendrogram Company 1

For Company 1, it can be observed in the dendrogram of Figure 7 that, when constructing two clusters, all the characteristic variables remain in the same group, that is, separated from the two variables that constitute the outputs of interest. Additionally, it can be noted that the variables that have been monitored (Consumption and period (Month)) are in two distant groups.

It is striking in the dendrogram of Company 1 that the variable Monthly Consumption, which according to the results obtained in the MDS analysis is quite distant to the response variable (with respect to the first dimension), and is not precisely the most remote from the conglomerate analysis.

This phenomenon can be explained by the fact that although the Consumption variable is distant in terms of the first dimension, its position around the variables under study in relation to the second dimension is not as distant as in the previous case.

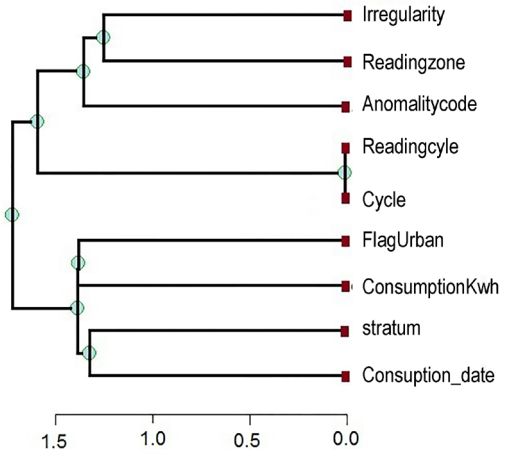

Figure 8 Dendrogram Company 2, clusters of correlation variables

Finally, it should be noted that the results obtained from this analysis for Company 1, are not showing a very satisfactory result, since from any grouping that is wanted to form, the two variables that contain the information of interest were isolated from the rest of variables, as can be observed in Figure 7.

The dendrogram of Company 2, presented in Figure 8, shows a main classification formed by two groups. From this perspective, it can be observed that the variables of interest (Consumption and Period (Consumption Date)), are in the same cluster. In fact, the distance observed between these variables in the previous analyze, is not clearly evident in the dendrogram.

An analysis of the dendrogram for company 2 let us see the closest feature to the target variable is the reading zone, followed by the reading code. It is important to highlight that for this company the zone where reading is donned, is a geographic variable. This allows estimating a georeferenced relationship of the problem, which was not visualized for the Company 1.

5. Implementation of the decision tree model

To finalize the follow-up of the variables Consumption and Period in the two companies, the decision tree constructed with the data of each one of the companies is presented in Figure 9.

For Company 1, it is found that both the Consumption and Period (Month) characteristics appear as incidents to estimate the irregular variable, despite the behavior that was found between them. However, this behavior allows us to think of the hypothesis that, given the panorama of the variables Consumption and Period in the previous analysis, it suggests that both variables should have incidence in the presence of irregularities. It is important to remember that the curves of Benford's law, that for this company; in the two variables were found infrequent behaviors, which led us to think that data was altered.

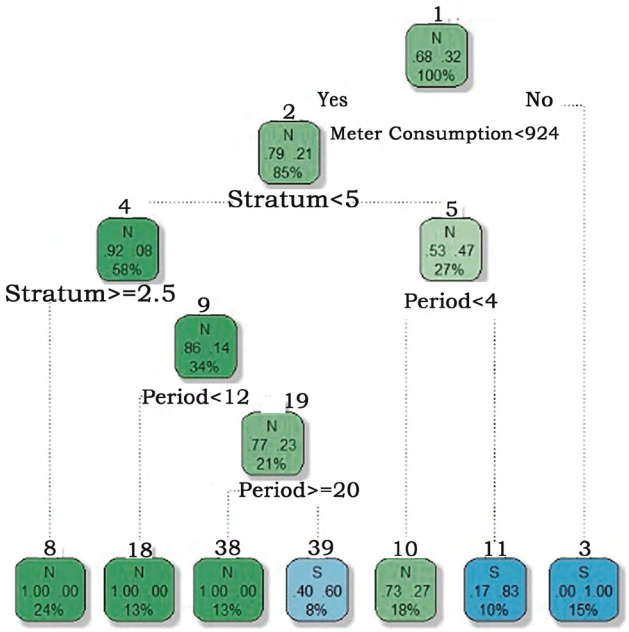

Figure 9 Decision tree Company 1 $ Irreg

On the other hand, surprisingly, the decision tree of Company 2 did not include the characteristic Consumption as an incident in the determination or identification of irregularities. However, the Period (Date of consumption) is incident in the determination of irregularities. It can be noticed in the Reading Zone characteristic, that the cluster related to the response was not determinant for this model either.

As validation of the models presented in Figures 9 and 10, the ROC curve (see Figures 11 and 12) is presented for the tree of each of the companies. These graphs allow us to evaluate the proportion of false positives and true positives that classifies the model. This also allows us to conclude that the model presented for Company 1, with an area under the curve of 0.91, is ranking better than that determined for Company 2, which has an area under the curve of 0.71.

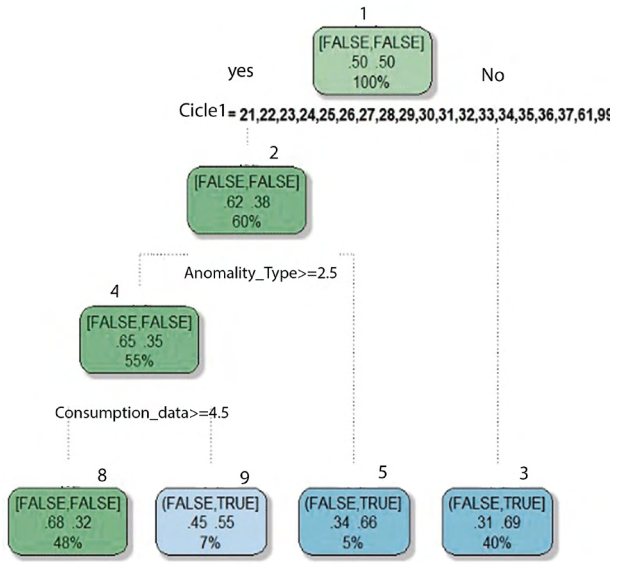

Figure 10 Decision tree Company 2 $ Irreg

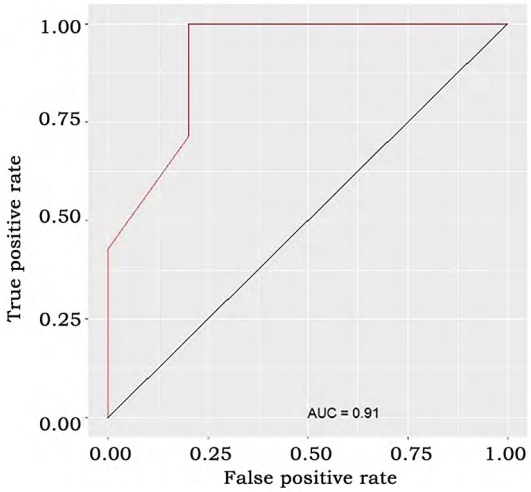

Figure 11 ROC curve for decision tree to Company 1

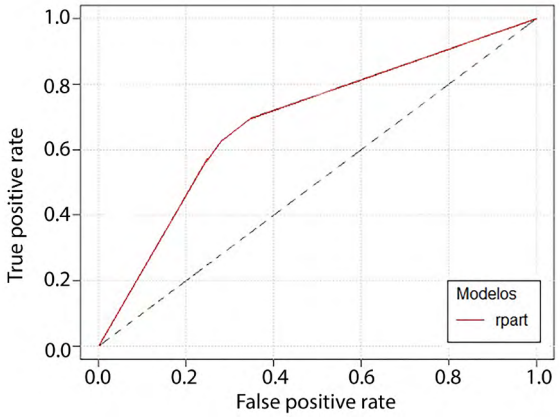

Figure 12 ROC curve for decision tree to Company 2

6. Conclusions

The results found in this analysis show a strange behavior of the variables consumption and periods in the Company 1, reflecting, for this company, that the consumption of the irregular users has a strange behavior in relation to both the Benford curve, as to the class of the regulars. This shows the alterations of this consumption. On the other hand, neither the consumption nor the period of Company 2 presents a special behavior. This behavior is particularly special in the decision tree model, since a better classification of the irregular users was found for the case of Company 1 than for Company 2.

It is important to emphasize in these conclusions that the behavior of the input characteristics in the study of an analytical model does not necessarily affect the objective variable in the same way. Additionally, the behavior of these variables within the input characteristics space is not necessarily the same for different energy trading companies. Consequently, it may be thought that the solution to the problem of non-technical losses has no unique solution.

7. Acknowledgement

The authors thank Colciencias for the support received for the innovation project "Analítica Avanzada para la gestión integral de eficiencia energética" code number 497770149036, MVM S.A.S –Universidad de Medellín, submitted research grant 701 of 2015.

8. References

1. J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, "Improving Knowledge-Based systems with statistical techniques, text mining, and neural networks for non-technical loss detection," *Knowledge-Based Systems*, vol. 71, pp. 376-388, 2014.
2. J. P. Navani, N. K. Sharma, and S. Sapra, "Technical and Non-Technical Losses in Power System and Its Economic Consequence in Indian Economy," *IJECSE*, vol. 1. no. 2, pp. 757-761, 2003.
3. P. Glauner *et al.*, *Large-Scale Detection of Non-Technical Losses in Imbalanced Data Sets*, 2016. [Online]. Available: https://arxiv.org/pdf/1602.08350.pdf. Accessed on: Mar. 13, 2016.
4. P. Glauner, J. Meira, P. Valtchev, R. State, and F. Bettinger, *The Challenge of Non-Technical Loss Detection using Artificial Intelligence: A Survey*, 2017. [Online]. Available: https://arxiv.org/pdf/1606.00626.pdf. Accessed on: Jun. 5, 2016.
5. Comisión de Regulación de Energía y Gas (CREG), *Propuesta para remunerar planes de reducción de pérdidas no tecnicas de energía electrica en sistemas de distribución local*, CREG, Bógota, Colombia, Jan. 2011.
6. J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohammad, "Detection of Abnormalities and Electricity Theft using Genetic Support Vector Machines," in *IEEE Region 10 Conference TENCON*, Hyderabad, India, 2008, pp. 1-6.
7. R. Jiang *et al.*, "Energy-Theft Detection Issues for Advanced Metering Infrastructure in Smart Grid," *Tsinghua Science and Technology*, vol. 19, no. 2, pp. 105-120, 2014.
8. B. Baesens, V. Vlasselaer, and W. Verbeke, *Fraud Analytics Using Descriptive, Predictive, and social Network techniques. A guide to data Science for Fraud Detection,* New Jersey, USA: Wiley, 2015.
9. J. P. Navani, N. K. Sharma, and S. Sapra, "Analysis of Technical and Non Technical Losses in Power System and its Economic Consequences in Power Sector," *International Journal of Advances in Electrical and Electronics Engineering IJAEEE*, vol 1, no 3, pp. 396-405, 2014. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.227.4747&rep=rep1&type=pdf. Accesed on: Aug. 28, 2017.
10. O. Refou, Q. Alsafasfeh, and M. Alsoud, "Evaluation of Electric Energy Losses in Southern Governorates of Jordan Distribution Electric System," *International Journal of Energy Engineering*, vol. 5, no. 2, pp. 25-33, 2015.
11. J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohammad, "Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 1162-1171, 2010.
12. E. A. Aranha and J. Coelho, "Probabilistic methodology for Technical and Non-Technical Losses estimation in distribution system," *Electric Power Systems Research*, vol. 97, pp. 93-99, 2013.
13. L. T. Faria, J. D. Melo, and A. Padilha, "Spatial-Temporal Estimation for Nontechnical Losses," *IEEE Transactions on Power Delivery*, vol. 31, no. 1, pp. 362-369, 2016.
14. C. C. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using Harmony Search and its application for non-technical losses detection," *Computers & Electrical Engineering*, vol. 37, no. 6, pp. 886-894, 2011.
15. I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. New York, USA: Springer, 2005.
16. I. Borg, P. J. Groenen, and P. Mair, *Applied Multidimensional Scaling*, 2nd ed. New York, USA: Springer, 2013.
17. T. F. Cox and M. A. Cox, *Multidimensional Scaling*, 2nd ed. New York, USA: Chapman & Hall, 2000.
18. N. Ye, *Data Mining. Theories, algorithms, and examples,* Boca Raton, USA: CRC Press, 2014.
19. G. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery,* New York, USA: Springer, 2011.