Electrical Power and Energy Systems 83 (2016) 402–410

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Determining a set of suspicious electricity customers using statistical ACL Tukey's control charts method

Josif V. Spirić [a,*], Slobodan S. Stanković [b], Miroslav B. Dočić [b]

[a] I. Strele 6/27, 16000 Leskovac, Serbia
[b] "Electric Power Industry of Serbia" – Section Leskovac, Stojana Ljubića 16, 16000 Leskovac, Serbia

ARTICLE INFO

Article history:
Received 18 February 2015
Received in revised form 8 April 2016
Accepted 11 April 2016
Available online 26 April 2016

Keywords:
Time series
Skewness
Total losses
The number of suspects
Indicator

ABSTRACT

Estimation a set of suspicious electricity customers is the first stage for detecting fraud of electricity. This paper deals with this stage. Based on historical data about monthly measurements of electricity, customers' time series are formed and then analysis of these series is performed using Tukey's control charts as one of the statistical method. The asymmetric control limit (ACL) Tukey's control charts are chosen due to a rather expressed asymmetry and dominantly presented right distribution of time series data. As usual, the choice of upper (*UCL*) and lower (*LCL*) control limits is not based on allowed number of observations outside of the control limits. The choice of these limits is based on the balance of total energy, registered energy and energy losses. The essence of this approach is the simultaneous observation all customers' time series of the controlled set with the same control limits and the same percentage of total and normalized energy losses in the observed distribution network. The criterion for finding the number of suspicious customers and their addresses is allowed error between the number of registered customers with one or more data (observations) outside the control limits for given losses and the number of suspicious customers' indicators based on certain balance of energy and the estimated total percentage of losses in distribution network.

Introduction

The subject of this paper is formation a list of suspicious customers who have signed a contract with the electricity supplier. Registered customers have to pay electricity each month for the previous month based on indications of measuring units. Analyzes and statements in this paper are based on monthly indications of consumed energy, which are recorded in supplier's billing database system.

For electricity supplier, the most relevant data about each customer is its time series of monthly invoiced energy. Regions with present fraud of electricity are characterized by increased losses of electricity. Increased losses were generally caused by electricity fraud of the certain number of customers. According to the theory of quality, a synonym for fraud activity is non-random factor in electricity consumption process. Also, a synonym for fraud may be an anomaly word. The anomaly is defined as a structural template (pattern) that is not customized with assumed normal behavior [1]. In other words, it means a deviation from usual rule, type or form [2].

There are many techniques for solving anomaly detection problem in terms of logic and mathematical tools application. Systematic division the most of these techniques is given in [1]. In this paper, ACL Tukey's Control Chart method is applied.

Statistical techniques are based on assumption that in stochastic model region an instance with normal data distribution is occurred with a high probability, and an anomaly in the same region is occurred with low probability. The basic division of these techniques is dependence of parameters. The first technique uses some knowledge about data distribution and is enabled the estimation of characteristic distribution parameters. The second technique is non-parametric where generally there is not assumed knowledge about available data distribution.

In further text of this introduction, authors will give a short overview of papers that consider data anomalies in order to detect electricity fraud in distribution systems of electricity utilities. Using of rough set theory and fuzzy set theory will be also mentioned.

In [3], electricity fraud and other non-technical losses detection in power distribution utilities are based on the use of Pearson's coefficient, Bayesian's networks and decision trees.

In [4], the MIDAS is name of the project which has developed two methodologies for fraud detection (dominant part of

* Corresponding author. Tel.: +381 64 836 7600.
E-mail address: josif.vspiric@gmail.com (J.V. Spirić).

http://dx.doi.org/10.1016/j.ijepes.2016.04.035
0142-0615/

non-technical losses). One is based on neural networks and other on statistical techniques.

In [5], *XMR* charts are used to indicate unnatural consumption profiles of registered customers. This checking method was tested on the time series sets of customers who were captured in fraud during the time series. It is shown that symptoms of non-random factors on time series of customers are discovered with a high percentage. This indirectly confirms the ability of method to successfully detect electricity fraud.

In [6], monthly development of a customer's selected variable is called the pattern, and is represented by a 12-dimensional vector for a period of one year. Assessment of density distribution requires a selection of a representative set of sample patterns, which is represented by a sample matrix. The following procedure leads to a pattern's degree of normality, which is defined as the measure of a pattern's frequency in the considered group of customers. High values of this coefficient will correspond to common (normal) behavior, whereas low values will reveal illogical situations.

In [7] is presented supervised anomalies detection process with classifiers that are result of an optimum-path forest (OPF) computation in the feature space induced by a graph. "These kinds of classifiers interpret the classification task as a combinatorial OPF computation from some key samples (prototypes) to the remaining nodes. Each prototype becomes a rout from its optimum-path tree and each node is classified according to its strongly connected prototype that defines a discrete optimal partition (influence region) of the feature space" [8].

In 1982, Polish mathematician Zdzislav Pawlak postulated the theory of rough sets as a tool for knowledge discovery in database (KDD) and it is based on indistinguishability relation [9,10].

Based on customer fraud results, a class of customers with anomalous series is formed and is characterized by corresponding patterns according to their consumption. After the discretization of conditional attributes, it is possible to find customers with consumption patterns that are identical to the class of customers with anomalous series patterns. Such customers belong to a boundary region and also, but not certain, may belong to a group of thieves. Because of that, they are considered as suspicious. Therefore, all customers in boundary region are the basis for the formation of a list according to which an on-site inspection should be performed [11].

To create a list of suspects, it is possible to use the fuzzy set theory. Firstly, at least two criteria should be defined that express the relationship of a specific customer's consumption characteristic to the appropriate average value of the selected consumption characteristic at the site the customer belongs to. The criterion can also be the ratio of the customer's consumption characteristic to that characteristic's average for an identified period for the same customer. Based on the selected criteria, their functions of belonging to fuzzy sets are formed, functions of belonging to fuzzy sets for suspicion assessment are determined and fuzzy rules according to the "if-then" system are set [12,13]. After the fuzzy reasoning procedure is finished, defuzzification is performed which turns a fuzzy conclusion into a real number that represents the suspicion evaluation. Values of suspicion evaluation (usually in %) make up the list of priorities for on-site testing.

Registered energy time series

The values of a customer's monthly electricity consumption are changed constantly over time and these changes are the result of a series of factors, many of which are random. Most random variables are distributed in nature according to the law of normal distribution.

According to the central border theorem, the observed random variable has normal distribution if it is affected by many factors in the capacity of an independent (or weakly dependent) variable, which is also distributed according to normal distribution [14]. This position may be extended under certain conditions and to a large number of factors, i.e. random variables distributed according to any law of distribution and with no restrictions regarding the dependence of a factor and the observed random variable. Based on the above, it can be considered that a series of customers' monthly energy, as a side effect of the process of electricity consumption, is distributed according to normal distribution.

A random process concept results from a random variable concept extension by associating each possible outcome s_i of a phenomenon or an experiment with an appropriate time function $X(t, s_i)$ instead of a number. Random process $X(t,s_i)$ is defined as a function that maps event space $S(s_1,\ldots,s_i,\ldots,s_n)$ into the family of time functions. A random process can be defined as $X(t)$ and the realization that corresponds to the i-th outcome s_i as $X_i(t)$ [15].

Energy which is measured in a time interval can be one of the characteristics of its usage process. That process is the result of many factors' effects and can be considered as random. Due to the adopted integer quantification of energy and the frequency of electricity meter readings, it can be considered as a process with discrete states and discrete time. The set of values $X_i(t)$ is called a time sequence or time series.

The observed data values of electricity consumption $W(t)$ and time of their occurrence (measurements) can be stored in a coordinate system $(W,0,t)$ that hosts the control limits of the process. The data arranged in such way form a time series $W_i(t)$. The process is out of control if some of points are located outside the region formed by control limits or specific point's series schedule in that region, depending on applied method.

Monitoring of electricity using process

The monthly invoiced energy values in time represent energy time series, but that series also represent the process of energy using. Statistical process control (SPC) is statistical method application for process monitoring and controlling.

The key SPC tools are control charts methods. Control charts detect changes in the process and suggests their users on non-random factors or process anomaly. Anomaly detection in series is carried out by defining upper *UCL* and lower control limit *LCL*, and observation position in relation to them [17,18].

There are many types of control charts. Chart type is adjusted to process. For proper use of control chart, in first of all, it should have enough number of process data or series data, the value of series mean change, member allocation of observed series, distribution symmetry, correlation between series members, etc. [16].

For larger changes in process average, *XMR* and Tukey's charts are dominant. They could be effectively used in cases of major mean changes of customer measured energy [19]. These two types of charts are linked by the fact that they belong to individual charts, which are characterized by single measurement at a given time.

Tukey's control chart

A Tukey's control charts method is chosen as suspicious customer's detection method. This method is suitable for process monitoring with individual values of process features. It has a characteristic for easy and simple control limits setting [19,21,22]. Tukey's control chart is also suitable for monitoring of major changes of the process mean. In comparison to other types of