Electrical Power and Energy Systems 71 (2015) 42–50

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

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

Fraud detection in registered electricity time series

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

[a] I. Strele 6/27, 16000 Leskovac, Serbia
[b] Company for Electric Power Supply "Jugoistok" Niš, Branch "Elektrodistribucija Leskovac", Stojana Ljubića 16, 16000 Leskovac, Serbia

ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form 8 February 2015
Accepted 20 February 2015
Available online 13 March 2015

Keywords:
Time series
The process
Fraud
Registered consumption drop
Control limits
Rules

ABSTRACT

This paper analyses time series without the seasonal component of consumers' power consumption at low voltage in order to detect fraud and illogical consumption by customers. Statistical process control is used, where the process represents the process of using electricity. *XMR* charts are used to indicate major changes (decrease) in registered customers' consumption. Verification of this method was tested on time series of a set of customers who were caught stealing during a time series. It shows that symptoms of non-random factors in time series of customers are revealed in a high percentage, which indirectly confirms the method's ability to successfully detect electricity fraud.

Introduction

The fraud of electricity can be committed by illegal, i.e. unregistered customers. Therefore, they are only consumers, not customers. Fraud may also be committed at the measuring points of legal, i.e. registered consumers who become customers by signing a contract. This paper deals with the fraud of electricity committed only by registered customers.

The fraud of electricity is a source of great non-technical losses and are a serious problem of functioning power distribution systems. This problem is especially visible in developing countries, countries in transition and generally in countries with low national income per capita. Furthermore, inspections of customer's measuring points are usually unprepared and unsophisticated which results low rate in discovering the fraud. The purpose of this paper is to promote statistical process control (SPC) strategy for detection of suspicious electricity customers.

Fraud of electricity by registered customers is a criminal activity that must be manifested in the time series data of monthly registered energy. Then, the time series can be termed as anomalous. The anomaly can be defined as behavior (structural template) that is not in line with the expected normal behavior. How to detect an anomaly on the basis of available data? To solve the problem of anomalies detection, there are the following techniques given in [1]:

- Detection technique based on classification.
- The nearest neighbor technique.
- Clustering technique.
- Statistical technique.
- Information theoretic technique.
- Spectral technique.
- Handling contextual anomalies.
- Handling collective anomalies.

The same source also lists several fraud detection applications including: intrusion detection, fraud detection, fault/damage detection, medical and public health anomaly detection, industrial damage detection, image processing, anomaly detection in text data, sensor networks and other domains. Here are a few papers that deal specifically with detection of electricity fraud.

In [2], detection of fraud and other non-technical losses in distribution companies is based on the use of Pearson's coefficient, Bayesian networks and decision trees. The key idea of these methods is identification of a customer pattern with a drastic drop in consumption and subsequent stabilization, but a gradual (significant) drop with subsequent stabilization of further consumption is also taken into consideration.

In [3], MIDAS is the name of the project which has developed two methodologies for fraud detection (the dominant part of non-technical losses). One is based on neural networks and the other on statistical techniques. The first methodology uses neural networks due to problem conditions and works with the Kohonen network structure. The second methodology is based on the detection of tolerant values outside the range (outliers).

* 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.2015.02.037
0142-0615/

In [4], 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 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 [5].

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 profiles (patterns) that are identical to the class of customers with anomalous series patterns. Such customers belong to a boundary region and may, although they do not have to, belong to a group of thieves. All customers in the boundary region are considered as suspicious and they are the basis for the formation of a list according to which an on-site inspection should be performed [6,7].

To create a list of suspects, it is possible to use "fuzzy" sets 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. In [8, 9], 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 "if–then" system are set. 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.

In this paper, on time series of fraud detected customers, the performance of *XMR* charts method is tested for fraud detection of those same customers. For detection of anomalies in time series of registered consumed energy, statistical method of *XMR* charts was chosen. In the article of Wheeler: "The Chart for Individual Values", it is stated this method was found by Jennet in the year of 1942. In fifties years of the last century (more precisely, from 1943 to 1953), *XMR* charts are frequently used in the General Electric Co. in Wembley. In 1980, Wheeler was started again to use this chart in chemical industry. The third and the fourth chapter, in more detail, will be described formation and interpretation of these charts.

For better overview of *XMR* charts method position, the basic performances of its use in SPC are given. Statistical process control is done by control charts which are based on statistic characteristics of the process. In the range of *X* chart, applied method is followed by basic (single) chart characteristic of the process and, in that meaning, it is an individual chart. This chart is followed by the process with sample size of $n = 1$. *X* chart is frequently used in non-manufacturing situations. One of this chart characteristics is long interval between observations. The chart range or *R* chart is followed by ranges between successive process characteristics. The ranges which are found on this way are the basis for process variability estimation. Simultaneously, *XMR* chart is also followed by process and range characteristic, so it is mutual or simultaneous chart. Process parameters can be known or "standard known" based on external specifications or long term experience. Processes with unspecified parameters are "standard unknown". There are two phases in the process control and monitoring. Phase I involves a retrospective study of the process and can be considered as preparatory process for Phase II which focuses on the ongoing monitoring [10]. By definition, *X* and *R* charts are most often used in Phase I of the process, which are characterized by larger changes of followed characteristics. Series anomaly on chart is determined by location of observations in regions within and outside control limits, and the existence of long sequences of consecutive observations. The characteristics of these charts are that they are relatively insensitive to small changes in the process, on the order of about 1.5σ [11].

In this paper, fraud detection is based on the analysis of monthly energy time series. Provided that the level of household electrification and the process of using a device which monitors that level are not changed, the main indicator of the series' anomaly in terms of potential fraud is a significant change in consumption value compared to the average consumption in the previous period. The analysis will be conducted on approximately uniform load diagrams of customers at low voltage in the household category.

The usual approach to anomaly detection is to define the region that represents normal behavior. Anomalies become instances where some observed data do not belong to the region of normal behavior. Much more rarely, in case when only anomalous instances are available, region of anomalous behavior is built, and then the region of normal behavior is defined [1].

In particular case, time series of fraudulent customers are available and these series are anomalous. The result of these series analysis should enable the identification of customer profiles with anomalous series. However, it is very important to choose an analytical method for this purpose. The suitable method should have a high success rate in anomaly time series estimation of customers who caught stealing. Taking into account the overall performances of *XMR* control charts as one of the statistical methods and the electricity consumption process with its performances, the *XMR* control charts method is chosen as a method for anomaly time series testing of registered electricity customers. In further text, the customer with the anomalous time series will be synonym for suspicious customer.

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 [12]. This position may be extended under certain conditions and to a large number of factors, i.e. random variables are 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