TIME SERIES CHANGE POINT IDENTIFICATION WITH APPLICATION OF THE DATA: VERIFICATION OF THE OCCURRENCE OF ABRUPT CHANGE POINTS IN THE TIME SERIES OF THE ESOCIAL DATA IN FRONT OF THE CAGED

Sérgio Luiz Rodrigues Torres

Dissertation

Supervisor:

In the FEP Advisor: Prof. *João Manuel Portela da Gama*

U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Master in modeling, Data Analysis and Decision Support System

Year 2019

BIOGRAPHICAL NOTE

Sergio Luiz Torres natural de São Paulo, was born in 1976. lived his childhood in Pernambuco and currently resides in the federal capital - Brasilia - Brazil.

It is a selfless person, responsible, fighter and hard worker, all that has been achieved through studies. It is demanding, but likes to live life with balance, there is time for everything, work, family and knowledge. The search is the quality of life.

In 2002, he entered higher education in economics at the Federal University of Pernambuco. After 5 years of graduation, finished as the second best student. However, it was not a pretense, but a consequence, because I liked to study the course in question.

The first job was the military police of the expert group; In the meantime, everyone who has this job has gone to law school, but curiosity about economic science arose from being in a developing country with various economic plans.

Later, entered the federal public service and works with employment data analyzing its structure and conjuncture. After 8 years, it was manifested the deepening of knowledge of data mining and its techniques. The main objective is related to realize that, currently, an economist has a direct relationship with machine learning and studies of data to diagnose the forecasts that the profession demands.

Thus, he enters the master's degree in modeling, data analysis and decision support systems of the Faculty of Economics of the University of Porto, which he joined in 2016, but by force majeure, cannot be present. However, pursuing a master's degree in 2017 and completing an average of 14 in total. There were numerous difficulties among them, the organizational culture, the English language, and little knowledge of the programming language and its derivatives.

In this context, it aims to apply the knowledge acquired to better analyze the entire economy and data structure and apply business analysis in public employment policies

ACKNOWLEDGMENTS

My biggest thank God, because I learned it from my mother, who without faith in the hour of anguish there is no overcoming. Very talk about resilience, i, in faith, so let this word. "(…) So the poor have hope, and injustice shut his mouth.

"How blessed is the man whom God corrects; therefore do not despise the discipline of the Almighty.

"Because it hurts, but it comes to treat; it hurt, but his hands also heal.

"Six misfortunes he deliver; in seven of them you nothing will suffer.

"In famine he will deliver you from death, and in war the deliver the blow of the sword.

"You will be protected from the scourge of the tongue, and does not need to be afraid when the destruction coming.

"You laugh from destruction and famine, and you do not need to fear the beasts of the earth.

It will make a covenant with the stones of the field, and the wild animals will be at peace with you.

"You shall know that your tent is safe; count their goods from thy dwelling and find nothing lack(…)". Job 5:18-24.

All my colleagues who accompanied me during the various stages of my academic path, in particular to the Raul, Bruno, Leopoldo and Nunes. With whom I have shared All uneasiness and joys that are part of the university spirit.

Also, I can't forget my father, who in the course of my journey, always helped me and this year has left us, "That God receives in his arms."

Finally, to all who make up the student body and faculty that somehow made part of this journey and helped and believed to be possible to conclude.

ABSTRACT

The objective of this operation is to investigate the basis of data from the System of Digital Tax Bookkeeping tax obligations to Social Security and Labor, nicknamed of and social, as a disruptive innovation for receipt of administrative records data from jobs, with the aim to assist with the optimization of the historical series of statistical data of the General Records of employment and unemployment, called CAGED. Because they understand that the temporal series are sequences of data point stipulated, measured in cycles of time stable during a period of time, it is, therefore, run the test of Page-Hinkley, once such a sequential analysis technique is commonly employed in the identification Monitoring changes in signal processing and enables effective identification of abrupt changes. The findings expose that the verification of the data base and social front of caged with the aforementioned test If the relevant show before the predetermined parameters. Are identified points of changes in temporal series, however, these were insignificant, proving the efficacy of and social.

Key words: Page-Hinkley; Time series; eSOCIAL; CAGED.

RESUMO

O objetivo desta exploração é investigar a base de dados do Sistema de Escrituração Fiscal Digital das Obrigações Fiscais Previdenciárias e Trabalhistas, alcunhado de eSOCIAL, como uma inovação disruptiva de recepção dos registros administrativos dos dados dos empregos, com o intuito de auxiliar com a otimização da série histórica dos dados estatísticos do Cadastro Geral de Emprego e Desemprego, denominado de CAGED. Por se compreender que as séries temporais são sequências de ponto de dados estipulados, medida em ciclos de tempo estáveis durante um período de tempo, é, portanto, executado o teste de Page-Hinkley, uma vez que tal técnica de análise sequencial é comumente empregada na identificação de alterações de monitoramento no processamento de sinais e possibilita a identificação eficaz de mudanças abruptas. Os resultados expõem que a verificação da base de dados do eSOCIAL frente ao do CAGED com o teste supracitado se mostram pertinentes diante dos parâmetros preestabelecidos. São identificados pontos de mudanças nas séries temporais, no entanto, estes foram inexpressivos, comprovando a eficácia do eSOCIAL.

Palavras-Chave: Page-Hinkley; série temporal; eSOCIAL; CAGED.

LIST OF FIGURES

LIST OF ACRONYMS

CAT	Communication of accident at work
CAGED	General records of Employment and Unemployment
CLT	Consolidation of Labor Laws
CNIS	Cadastro Nacional Social Information
CTPS	Portfolio of work and Social Security)
Dataprev	Technology company and information from Social Security
DIRF	Declaration of Income Tax)
eSOCIAL	Digital Tax Bookkeeping System of tax obligations to Social Security
FGTS	The Guarantee Fund of the time of service
GFIP	The Information tab to Social Security
LRE	Book of Record of employees
PH	Page-Hinkley
RAIS	Annual Relation of Social Information
SERPRO	Federal Service of Data Processing

"Behold this sambinha made a note only.

Other notes will enter, but the base is only one.

This is a consequence of what I have just said.

As I am the inevitable consequence of you.

How many people there are out there that talks so much and says nothing,

Or almost nothing.

I've used throughout the scale and in the end there's nothing left,

it came to nothing

And I returned to my note as I come back to you.

I count on my note as i like you.

And who is that all notes: Re, Mi, Fa, Sol, La, Si, dó.

You'll Always no stay in a note only."

(Letter: Newton Mendonça; Music: Antônio Carlos Jobim, 1961),

TABLE OF CONTENTS

1. INTRODUCTION

Firstly, it is necessary to justify the use of Music "Samba a Note Only" (Letter of Newton Mendonça and music of Tom Jobim, 1961) in the title of this work. The present scientific research has as its object of study to investigate the basis of data from the System of Digital Tax Bookkeeping tax obligations to Social Security and Labor, nicknamed of and social, as a disruptive innovation for receipt of administrative records data from jobs, with the aim to assist with the optimization of the historical series of statistical data of the General Records of employment and unemployment, called CAGED.

It is understood that there is an interrelationship between CAGED and eSOCIAL since, in the context of implementation, CAGED contains information on establishments and workers with a high level of breakdown of the establishment / worker profile. Its scope is all private employment sector formalized in the Brazilian territory. The supply of data is carried out via the web causing viability of time and space. These factors enable a greater understanding of market supply and demand distortions. In general, the system contributes positively in cost savings by searching for the way of reception and the possibility of existence of data in continuous periods of time.

To modernise and facilitate access to information, having in view the correlation of data, the federal government has deployed an information system that unifies the sending of information by the employer in relation to its employees, the eSOCIAL.

In this context, the preliminary analysis of this work is based on the importance of administrative records of the CAGED, because these are the metadata that will exist the new project.

This way, eSOCIAL will allow you to extract new information due to the correlation of data that was in different systems. These data can be described as events against CAGED administrative records. However, they may be affected structurally, including the loss of the historical series.

This study is justified by the need to give continuity to the disclosure of the employability of the labor market in Brazil. In this context, the general register of employed and unemployed (CAGED) is an important tool to analyze the Brazilian labor market, in its peculiarities. The present research is justified, too, for the possibility of producing scientific knowledge about the specific contribution of the identification of the point of change in temporal series with application of data with the verification of the occurrence of abrupt changes in time series of data eSOCIAL front to the CAGED. This study, which to be scientific, there is the concern with the use of methods of investigation.

It is, therefore, the relevance of research of the theme presented and defined in this study as a contribution to Statistical Science, in particular, by presenting appropriate to the area of "*data mining*".[1]

The application model for such a proposition was the Page Hinkley. The Page-Hinkley (PH) is a sequential analysis technique commonly used in the identification of monitoring changes in processing of signals[1,2,3]. The source of the data comes from companies that earn over 78 million in Brazil, eSOCIAL and CAGED in the last 5 years. These companies represent 29% of private jobs with contracts in Brazil.

This interrelation cited in the screen covers several points that, in this way, not investigated, in this exploration, only "one note" of the database of employability of private workers in Brazil. This research has as elements the knowledge regarding the importance of identifying points of change, as well as for the application of modeling and prediction of univariate and multivariate time series. Therefore, it is necessary to investigate with greater importance using the method or methods that best fit the study model.

Scientific research involving several "notes" of time series change point identification with data application, in a broad quadrant, of practices that seek a more assertive method, reveals that this is a complex totalizing that aggregates several aspects, considering that when there is a method that includes a good planning of the data to be worked, the results may go beyond the specificity questioned and, therefore, broaden the data analysis, seeking the completeness of the information, not only in the statistical scope.

There are some of the barriers that stand in the framing of a method for such identification that can be listed as specific objectives of exploration of this work. Thus, this study aims to: implement possible technical models (algorithms) to investigate change point detection. in the eSOCIAL against CAGED database, as well as identifying which relevant change point detection factors may break CAGED's historical series, as well as examining the eSOCIAL and CAGED databases.

To approach the problem it was necessary, firstly, to delimit the research object in its dimensional aspect, that is, the investigation is about how to identify the point of change in time series with data application. This delimitation took into account the current aspect of the theme and the social dynamism that involves information and data feeding of establishments and workers with a high level of disaggregation of the profile of the establishment and worker. Given the above, it is understood that it covers the entire private employment sector formalized in Brazil.

---

[1] Term in English for data mining

It is noteworthy, at this point, the not departing from the moment before the choice of the object of investigation, that is, the ideology of the researcher in question, and his practice and professional performance as a business analyst, in the conjuncture quality regarding the techniques of data mining, the theoretical model in which it is necessary for the formulation of the problem and the methodology adopted, evidencing its practice and the need to highlight such relevance of the delimited actor-theme as object of investigation.

In this epistemological moment of delimitation of the object, as well as in determining the way to investigate it, there is no need to talk about the separation between theory and practice, because the reflection of theory without any practice is idealism. Thus, the researcher, in this moment of delimitation, started from the set of theories and practices obtained after bibliographical revision concerning the definition of the time series and the identification of the point of change in the time series, but which do not reflect a dogma, ie, in the development. It is possible to establish new paradigms, to obtain new methods, in the light of theorists, to obtain the scientific answer formulated for the investigation and, therefore, to reach such importance, which is mentioned on screen, which involves the occurrence of abrupt points of change in the time series. eSOCIAL data compared to CAGED.

From the bibliographic review visited, it is noteworthy for the preparation of this study, it was possible to state the scientific-statistical problem with the following question: how to identify the occurrences of abrupt change points in the time series of eSOCIAL data compared to CAGED? This question can be considered a scientific problem, as it is supported by the bibliographical review of the researcher and is part of the trajectory of the process of knowledge production in relation to the theme delimited as a research object. In elaborating this question, the researcher sought to analyze the possibilities of explaining the problem through a method that best fits. Therefore, the following is a bibliographic review of this exploration.

2. BIBLIOGRAPHIC REVIEW

This section presents a review and research of the literature, introducing the time series process as well as the detection of change points. Comparative studies between the methods are also presented and discussed.

2.1 Definition of the temporal series

Time Series is a sequence of well-defined data points, measured at consistent time intervals, over a period of time. Data collected on an ad hoc basis or irregularly do not form a time series[4]. Time series analysis is the use of statistical methods to analyze time series data and extract significant statistics and characteristics on the data[5].

In this way, it is understood that the time-series analysis helps in understanding about what are the underlying forces that lead to a particular trend in the data points on the temporal series and, thus, helps to predict and monitor the data points, adapting them to the appropriate models.

Historically speaking, time series analysis has existed for centuries and its evidence can be seen in the field of astronomy where it was used to study the movements of planets and the sun in ancient ages[4]. Nowadays, time series analysis is employed in virtually every sphere around us - from everyday business matters, whether monthly sales of a product, to daily Ibovespa closing value, to complex research and scientific studies, such as in the parameterization of points of a specific evolution of some descriptor or seasonal changes[6].

The theme of the temporal series is defined as a set of historical series of data in sequence on a continuous or discrete interval of time. His analysis is required to observe and model the phenomenon studied to discuss what is the behavior of the series[6].

As to your range can present a set ordered in time that can be discreet, i.e., at intervals of time ($t = 1, 2, ..., n$), or continuous[7]. All factors have continuous intervals of time. In addition to being able to be discretized the series continues to be analyzed. The values of a series can also be obtained by aggregation, summed and media is obtained[7,4].

The temporal series can be investigated in parametric and non-parametric model, depending on the analysis. The first approach is through the temporal domain and the second approach is used in the domain of frequencies[8].

The historic series may submit components that, a priori, are not observed, however strictly important to influence the studies and models. These components are tendency (Tt) and seasonality (St) and random variation (TA) to zero mean and constant variance[7]. The decomposition allows

{Zt, t=1, 2,...,n} is decomposed into the sum or multiplication [7.9]. The model of the sum to be zt Zt=Tt+St+at and the multiplication Zt=Tt. St.at.

The trend can be understood as a gradual increase or decrease of the observations over time, the seasonality indicates possible performances occurred always in periods less than or equal to twelve months, and the random component that shows random irregulars oscillations[5,8].

As soon as there are two methods most often used to study the trend. The first option use the polynomial regression models, the second, autoregressive models [7.9].

Already the seasonality is necessary to study the phenomenon of study coupled with the statistic to fit the regression models that incorporate sine or cosine function the variable of time. Therefore, to discover the seasonality can use the spectral analysis[5].

The spectral analysis is the sum of the statistical analysis of the temporal series more methods of Fourier analysis. To remove the effect of seasonality of a series, the centered moving average is an important function.

Another important characteristic is to observe the stationarity. If is stationary or non-stationary. The simplest model and common in a temporal series called -if stationary.

I.e., free of trend and seasonality. According to Aminikhanghahi[10], this definition implies that,

- The function of the average value of μ t = (x t) is constant and does not depend on the time t .
- The automatic covariance function γ (s , t) = voc (x s , x t) = [(x s - μ s) (x t - μ t)] depends on the imprint of time s and t only through the time difference, or ∣ s - t ∣.

Technically, there are two ways of stationary models, the weak (wide) and the strong(strict), the second is very demanding and difficult to check[9]. The first is defined to a stochastic process as being weakly stationary or stationary second-order logic se[9]:

(i) $E[Z(t)] = \mu(t) = \mu$, constant for all;$t \in T$

(ii) $E[Z^2(t)] < \infty$, for the whole;$t \in T$

(iii) $\gamma(t_1, t_2) = Cov[Z(t_1), Z(t_2)]$ is a function of $|t_1 - t_2|$

To confirm the stationary random process is necessary to describe the functions mean, variance and autocovariância. In addition to autocorrelation function that is used to identify a suitable model for the temporal series[8].

2.1.1 Autocovariânce Function

$\{X_t, t \in T\}$ A real process is stationary discrete zero mean, its function of autocovariância (f.a.c.v.), as Morettin and Toloi[8] is defined by:

$$\gamma_\tau = E[X_t X_{t+\tau}]$$

2.1.2 autocorrelation function

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}, \tau \in Z$$

And has the same properties of autocovariância function, γ_τ except that now we have $\rho_0 = 1$.

In this way, the auto-regressive models are termed as air, MA and a weapon whose used for studies of stationary series [7:] The first describes how the random combination of previous values of Zt (Zt zt-1=b1+b2Zt-2 + ... +bp Zt-p) and, therefore, the whole series can be a function of white noise.

Therefore, it is the estimate of a prediction model as well as the analysis of residues for the assessment of the existence of biases and/or large errors of estimates. [7]

Thus, it is given [9]:

$$AR(p): Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \ldots + \emptyset_p Z_{t-p} + a_t$$

Already the second named as medium-sized pieces understands - if Zt is a function of the weighted algebraic sum at that move in time. The process is marked by a linear combination of white noise after in the present and past of time [7.]

This is portrayed as follows[9]:

$$MA(q): Z_t = a_t - \emptyset_1 a_{t-1} - \emptyset_2 a_{t-2} - \ldots - \emptyset_q a_{t-q}$$

The last one is the ARMA model externalized in the combination of the two previous models (AR and MA) in order to optimize the best solution found for a certain type of time series. This linear combination of close values and series[7] white noise. Has if [9]:

$$ARMA(p,q): Z_t = \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \ldots + \emptyset_p Z_{t-p} + a_t - \emptyset_1 a_{t-1} - \emptyset_2 a_{t-2} - \ldots - \emptyset_q a_{t-q}$$

However, much of the temporal series expresses non-stationary homogeneous characteristics. This is a number of different components has finite (Trend). Therefore, it can be analyzed through process of differences (DdZt) in the ARIMA model[7].

The model is appointed with ARIMA model (auto-regressive integrated moving average), it is necessary to withdraw the trends. This model solves the problem of seasonality when is simple periods[7].

However, when the temporal series presents seasonal factors from various periods is necessary to incorporate models of random seasonality called sarima. Incorporation of trigonometric functions[7].

Finally, there are numerous components of data of the temporal series that can be studied and analyzed such as: Making estimates, check what are the factors that influence the behavior of the serie, make predictions of future values and seek periodicities in the data.

2.2 TYPES OF TIME SERIES ANALYSIS

The various types of analysis of temporal series include[11]: descriptive analysis, spectral analysis, forecasting, analysis of intervention and explanatory analysis.

2.2.1 Descriptive Analysis

The descriptive analysis has the function to determine the trend or pattern in a temporal series using graphics or other tools. In this way, the descriptive analysis helps identify cyclical patterns, general trends and turning points [11.6.]

2.2.2 spectral analysis

The spectral analysis is also referred to as the domain of frequency and has as objective to separate periodic or cyclical components in a temporal series. For example, identifying cyclical changes in sales of a product [4,11,12, 13].

2.2.3 Forecasting

The forecast is used broadly in business forecasting techniques, in the budget, among other things, with foundations in historical trends[11]. The prediction of temporal series uses information about historical values and standards associated to predict future activities. In most cases, this is related to the analysis of trends, analysis of cyclical fluctuations and issues of seasonality. As in all methods of forecasting, success is not guaranteed[11].

2.2.4 Intervention Analysis

The analysis chart is used to determine if an event can bring a change in the time series, such as the level of performance of a fuincionário whether or not improved after an intervention in the form of training, to determine the effectiveness of a training program [4.11].

2.2.5 explanatory analysis

The explanatory analysis studies the cross-correlation or relationship between two time series and the dependence on one another., for example, the study of data from employee turnover and employee training data to determine if there is any dependence of training programs for officials in the rates of turnover of employees over time[8, 11.]

2.3 BENEFITS OF TIME SERIES ANALYSIS

The time series analysis aims to achieve several goals and the tools and models used vary according to each object of study[11].

The biggest advantage of using time-series analysis is that it can be used to understand the past and anticipate the future. Furthermore, the analysis of time series is based on past data e compared with the time, which is readily available in most areas of study[4]. For example, a provider of financial services may want to predict future movements in the price of gold to their customers. He can use data available historically to lead the analysis of temporal series and predict the rates of gold for a given future period[4.]

There are several other practical applications of time series analysis, including economic forecasting, analysis of census and projections of income. In addition, it is used by investment analysts and consultants to analyze the stock market and portfolio management[6].

Business managers use analysis of temporal series regularly for sales forecasting, budget analysis, inventory management and quality contro[l4].

2.4 TIME SERIES CHANGE POINT DETECTION

A temporal series represents a class of phenomena whose process, observational and consequent, numerical quantification, generates a sequence of data distributed in time, the shift point are sudden variations in the set of linear time series is not linear discorrido in time[8]. Throughout the study, the purpose is to show the relevance of the detection of point of change in a different approach, as well as the methods used. It is important to emphasize that detection of point of change is found in different areas of activity[8].

In addition, only depends on the model and the technique studied to explain the behavior of the activities recorded data over a given time period[8].

2.5 TIME SERIES BASIC

A temporal series can be taken in any variable that changes with time. When you invest, it is common to use a temporal series to track the price of a security over time. This can be tracked in the short term, as the price of one title at a time over a working day or long-term, as the price of a title in security on the last day of each month during the period over five years. [14.]

Time series analysis can be useful for seeing how an asset, security or economic variable changes over time. It can also be used to examine how changes associated with the chosen data point compare to changes in other variables over the same time period[15]. For example, suppose someone wants to analyze a time series of daily closing quotes for a given stock over a period of one year. This individual would get a list of all inventory closing prices for each day of the previous year and list them in chronological order[15]. This would be a time series of one year daily closing price for the stock. Going deeper, a person may be interested in whether the action's time series shows any seasonality to determine if it experiences ups and downs at regular times each year[16].

Analyzes in this area would require that the observed prices were correlated to a specific station. This may include traditional calendar seasons, such as summer and winter, or seasons of retail, as times of ballroom8. Alternatively, a person can record the changes in the price of the shares of an action, in what refers to an economic variable, such as the unemployment rate[8].

To correlate the data points with information related to the economic variable selected, a person can observe patterns in situations that exhibit dependency between the data points and the variable chosen[15].

2.6 SEASONALITY

The seasonality is a characteristic of a temporal series in which data experience regular and predictable changes that occur each year. Any fluctuation or predictable pattern that repeats or repeat over a period of one year is considered to be seasonal[17].

The seasonal effects are different from cyclical effects, once the seasonal cycles are observed within a calendar year, while the cyclical effects, such as increased sales due to low rates of unemployment, may cover periods of time shorter or longer than one year[18]

Soon[17]:

- Seasonality refers to foreseeable changes that occur during a period of one year in a company or economy based on the seasons of the year, including commercial seasons or calendar.
- The seasonality can be used to help analyze actions and economic trends.
- Businesses can use the seasonality to help determine certain business decisions, such as inventory and personnel.
- An example of a measure seasonal are retail sales, which typically record higher spending during the fourth quarter of the year.

Businesses that understand the seasonality of your business can predict and delaying the inventories, the teams and other decisions in order to coincide with the expected seasonality of associated activities, thereby reducing costs and increasing revenue[5,17]

It is important to consider the effects of seasonality to analyze actions of a fundamental point of view, because this can have a big impact on profits and in the portfolio of an investor[17.]

A company that experience higher sales during certain seasons may seem to obtain significant gains during peak periods and significant losses during the off-peak times[17]. If this is not taken into account, an investor, for example, you can opt to buy or sell securities based on the activity in question without accounting for the seasonal change that occurs later as part of the seasonal business cycle of the company[17]. The seasonality is also important to consider when tracking certain economic data. Economic growth may be affected by different seasonal factors, including the weather and the holidays[11,17]. Soon, economists may have a better idea of how an economy is moving when they adjusted their analysis based on these factors. If this seasonality was not taken into account, the economists do not have a clear vision of how the economy is actually moving. The seasonality also affects the sectors - called seasonal industries - that usually get the most for your money during small and predictable periods of the year[17].

2.6.1Adusting data for seasonality

Many data are affected by season of the year, and the adjustment to the seasonality means that more precise relative comparisons can be made between different periods of time. Adjust the data for the seasonality balances periodic oscillations in the statistics or the movements of supply and demand related to changes in the seasons. Using a tool known as a seasonally adjusted annual rate (SAAR), seasonal variations in data can be removed[8]. For example, the homes tend to sell more quickly and at higher prices in summer than in winter. As a result, if a person compare the selling prices of real estate in the summer the average prices of the previous year, it can have a false impression that prices are rising18. However, if he adjust the initial data based on the station, he can see if the values are actually increasing or only momentarily increasing with the warm climate[11].

2.7 Data Mining

Data mining is a way of exploring hidden patterns, valid and potentially useful in large data sets, that is, the data mining comes to discover unsuspected or unknown relationships between data, is a multidisciplinary skill that uses machine learning, statistical, artificial intelligence and database technology[16.]

The insights derived from data mining can be used para marketing, fraud detection and scientific discoveries, among others. Data mining is also called knowledge discovery, extraction of knowledge, data analysis / patterns, collection of information[19]

2.7.1 Data Types

Data mining can be performed in the following types of data[19]:

- Relational databases
- Data Warehouses
- Advanced database and information repositories
- Databases, object-oriented and relational objects
- Transactional Databases and spatial
- Heterogeneous databases and legacy
- The Multimedia database and streaming
- Text data banks
- Text mining and mining of the Web

2.7.2 The process of implementing data mining

The process of implementing data mining can be listed as primary element the understanding of business, then, the understanding of data preparation, data modeling, assessment and, finally, development, summarised as follows[19]:

Figure 1 - The process of implementing data mining

Source: Adapted from manhães[19]

2.7.2.1 understanding of business

The first part which refers to the understanding of business, at this stage, the business goals and data mining are established, so the individual understand the business objectives and customer needs, therefore, to define the objective. Thus, it is necessary to make an assessment of the current scenario of data mining, factors in features, supposition, restrictions and other significant factors for an evaluation.

2.7.2.2 understanding of data

In this phase, the sanity check of the data is performed to check whether it is appropriate for the goals of *data mining*[19].

- First, the data are collected from various sources of data available.
- These data sources can include multiple databases, filer plan or data cubes. There are problems such as the correspondence of objects and the integration of schemes that may arise during the process of data integration. It is a very complex process, because it is unlikely that the data from various sources correspond easily[19].
- Based on the results of the consultation, the quality of data should be verified. Lost data, if any, shall be recovered or acquired.

2.7.2.3 Preparing data

In this phase, the data are ready for production. The process of preparing data consumes approximately 90% of the time of the project. Data from different sources should be selected, cleaned, processed, formatted, anonymized and constructed, if necessary[19].

The data cleansing is a process to "clean up" the data, smoothing noisy data and completing missing values. For example, for a demographic profile of customers, the data are missing. The data are incomplete and must be filled in. In some cases, there may be conflicting data. For example, the age has a value of 300. The data can be inconsistent. For example, the name of the client is different in different tables[19].

The processing of data alter the data to make it useful in data mining, so that the processing can be applied.

2.7.2.4 Modeling

The processing of data would contribute to the success of the mining process, in part, to see if the smoothing, aggregation, the generalization, standardisation, the construction of attributes, the modeling, assessment and deployment.

Table 1- Modeling

Smoothing	Helps to remove the noise from the data.

Link Aggregation	The operations of summary or aggregation shall be applied to the data, i.e., the weekly sales data are aggregated to calculate the total monthly and yearly.
Generalization	At this stage, the low-level data are replaced by higher level concepts with the help of concept hierarchies. For example, the city is replaced by region.
Standardisation	Normalization is performed when the attribute data are scaled down or reduced. Example: The data should fall in the range of -2.0 to 2.0 after normalization.
Construction of attributes	These attributes are constructed and include the set of attributes available as useful for data mining. The result of this process is a set of data that can be used in the modeling.
Modeling	At this stage, mathematical models are used para determine patterns of data. • Based on business objectives, the appropriate modeling techniques should be selected for the data set ready. • Create a scenario to test the quality and validity of the model. • Run the model in the data set ready. • The results should be evaluated by all interested parties to ensure that the model can meet the objectives of data mining.

Evaluation	In this phase, the standards identified are evaluated in relation to the goals of the business. • The results generated by the model of data mining should be evaluated in relation to business objectives. • Gain an understanding of the business is an interactive process. It is understandable that the new business requirements can be raised because of data mining. • A decision to go or not is taken to move the model in the implementation phase.
Unfolding	In the deployment phase, it sends it to the discoveries of mining data for daily business operations. • The knowledge or information discovered during the process of data mining should be easy to understand for those interested não technicians. • A detailed deployment plan for submission, maintenance and monitoring of discoveries of Data mining can be created. • A final report of the project can be created with lessons learned and key experiences during the project. This helps to improve the business policy.

Source: Adapted from manhães et al[19]

2.7.3 Data Mining techniques

The data mining techniques can be listed, as shown in Figure 2:

Figure 2- Data mining techniques

Source: Adapted from Tan et al.,[16]

2.7.3.1 Classification

This analysis is used to retrieve information about important and relevant data and metadata. This method of mining data helps to sort data in different classes[16,19].

2.7.3.2 Clustering

The cluster analysis is a technique for data mining to identify data similar to each other. This process helps to understand the differences and similarities between the data[16].

2.7.3.3 Regression:

The regression analysis is the method of mining data to identify and analyze the relationship between variables. It is used to identify the probability of a specific variable, given the presence of other variables[16].

2.7.3.4 outdoor detection

This type of data mining techniques refers to the observation of data items in the data set that does not correspond to an expected pattern or expected behavior. This technique can be used in various fields, such as intrusion, fraud detection, or detection of failures etc. The external detection is also called *Outlier Analysis* or *Outlier Mining*[16].

2.7.3.5 Sequential Patterns

This data mining techniques help to discover or identify patterns or similar trends in transaction data for a given period[16].

2.7.3.6 Prediction

The Forecast uses a combination of other techniques of data mining, as trends, sequential patterns, clustering, classification, among others. Thus, it analyzes the past events or instances em a correct sequence to predict a future event[19].

2.7.3.7 Association Rules

This data mining techniques help to find the association between two or more items. He discovers, therefore, a pattern hidden in the set of data[16].

2.7.4 Challenges of implementation of Data

The challenges for the implementation of data, are[11]:

- Qualified specialists are needed to formulate the queries of data mining.
- Overfitting*:* Due to the database of training of small size, a model may not fit in the future.
- The data mining need large databases that are sometimes difficult to manage
- The business practices may need to be modified to determine the use of the information being exploited.
- If the data set is not diverse, the results of data mining may not be accurate.
- Integration Information needed from heterogeneous databases and systems of global information can be complex

2.7.5 Data Mining Tools

The following are two popular tools for data mining widely used

*2.7.5.*1 R Language

The R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical tests, classic, time series analysis, classification and graphical techniques. It offers ease of delivery and effective data storage[16].

2.7.5.2 Oracle Data Mining

Oracle Data Mining popularly known as mdgs is a module of Oracle Advanced Analytics Database. This data mining tool allows data analysts to generate detailed insights and make predictions. Helps to predict customer behavior, develop customer profiles and identify cross selling opportunities[16].

2.7.6 Data Mining Benefits

- The data mining techniques help companies to obtain profits, so generates[16].
- Data mining is an economical solution and efficient in comparison with other applications of statistical data[16].
- The data mining helps in the decision making process[16].
- Facilitates automated forecasting trends and behavior, as well as automated discovery of hidden patterns[16].
- Can be implemented in new systems, as well as in existing platforms
- Is the quick process that makes it easier for users to analyze vast quantities of data in less time[16].

2.7.7 Disadvantages of Data Mining

- There are chances of companies selling useful information of its customers to other companies for money.
- Many software for analysis of data mining are difficult to operate and require advanced training to work.
- Different data mining tools operate in different ways, due to the different algorithms used in your project. Therefore, the selection of the tool of data mining is a very difficult task.
- The data mining techniques are not accurate and, therefore, may cause serious consequences in certain conditions.

According to Tan et al[16], the process of mining data includes commercial understanding, comprehension of data, data preparation, modeling, development, deployment, making complex.

3. METHODOLOGY

Different time series are represented in several fields of activity. For example, in the field of meteorology, jobs, business, financias, medicine and among others. These data sets are sources for estimating, verify factors of behavior.

This time, it was necessary to address the problem to be studied and apply the model to solve and explain the phenomenon of the study. This section will make a narrative of what was studied among the alternatives to solve the problem of detection points of change in historical series in different data and expose the model proposed for such analysis.

First, it was necessary to verify the existence in the literature point detection algorithm for change in version online and offline. The online is described "run concurrently with the process they are monitoring, processing each data point as it becomes available, with the goal of detecting the change point as soon as possible after it occurs, ideally before the next data point arrives"[17]

Already the offline algorithm covers the entire set of data and seeks to identify all points of change. However, the online algorithm has a limited capacity to identify the point of change.

Considering that the computational capacity and their learning needs much effort to detect in time online. I.e. the workout data may not submit possible important change which may arise in implementing the process simultaneously.

In addition, temporal series can present a set of data from different characteristics. Therefore, it is important to address what is the dimension of these data. However, compare the operational cost of algorithms, this is done by parametric and non-parametric methods[10].

As discussed, "the computational cost of parametric methods is higher than nonparametric approaches and does not scale as well with the size of the dataset". Therefore, for dimension data rates the best application was the non-parametric[18].

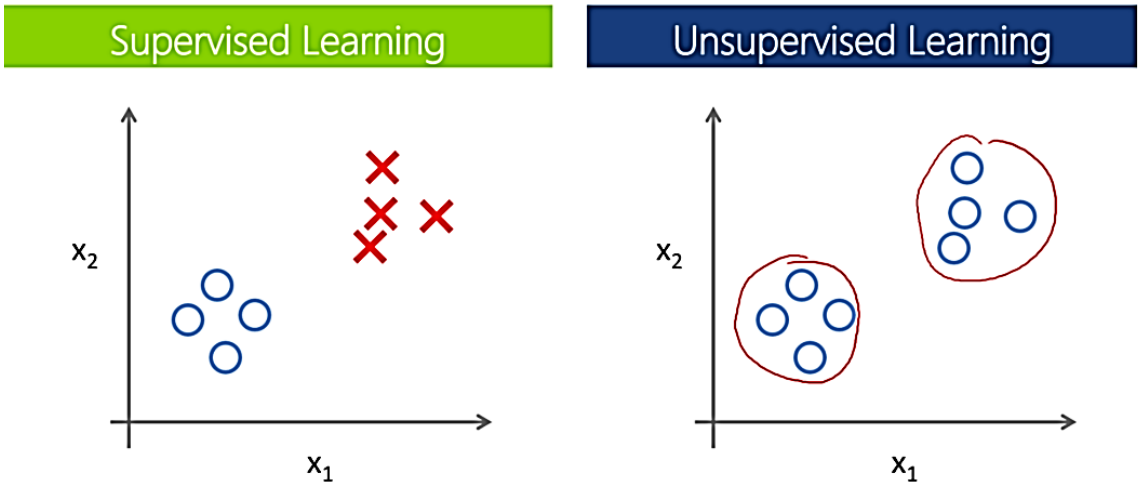
Another important component of the algorithm is as follows. I.e., algorithms of contract and interruptible algorithms. Contract the execution time is a parameter specified[10]. If interrupted, the quality of solution may not be good.

In turn, the interruptible algorithm is prepared for such an event. Their quality of solution and preserved. "In general, every interruptible algorithm is trivially a contract algorithm, but the converse is not true"[19.]

3.1 SUPERVISED AND NON-SUPERVISED METHODS: THE CLASSIFIERS FOR DETECTING OF POINTS OF CHANGE

The machine learning covers different types of methods for solving various problems exposed. The most common are the methods supervised and non-supervised[20].

Figure 3: Example of the differences of approach to learning



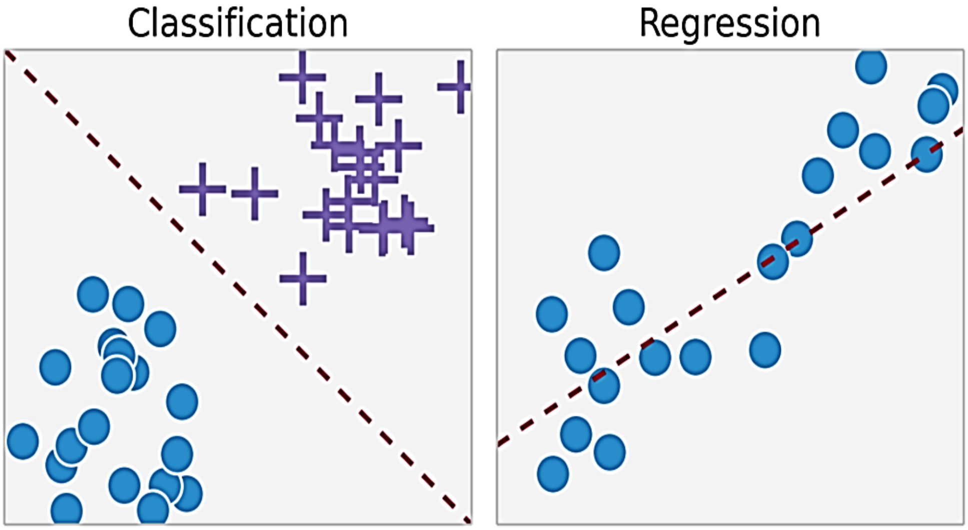
Source: Barros[20]

The supervised are defined by a system or programd from a set of pre-defined data or labelled.

There is a relation of data inputs and outputs. The program is able to make their own decisions after that is trained. The problems are presented as regression or classification.

The regression problem is the forecast of the results of the variables of inputs to a function continues. Since the problem of classification, the forecast has -if the results in a discrete output. The input variables are supervised in distinct categories.

Figure 4: Example of classification and regression

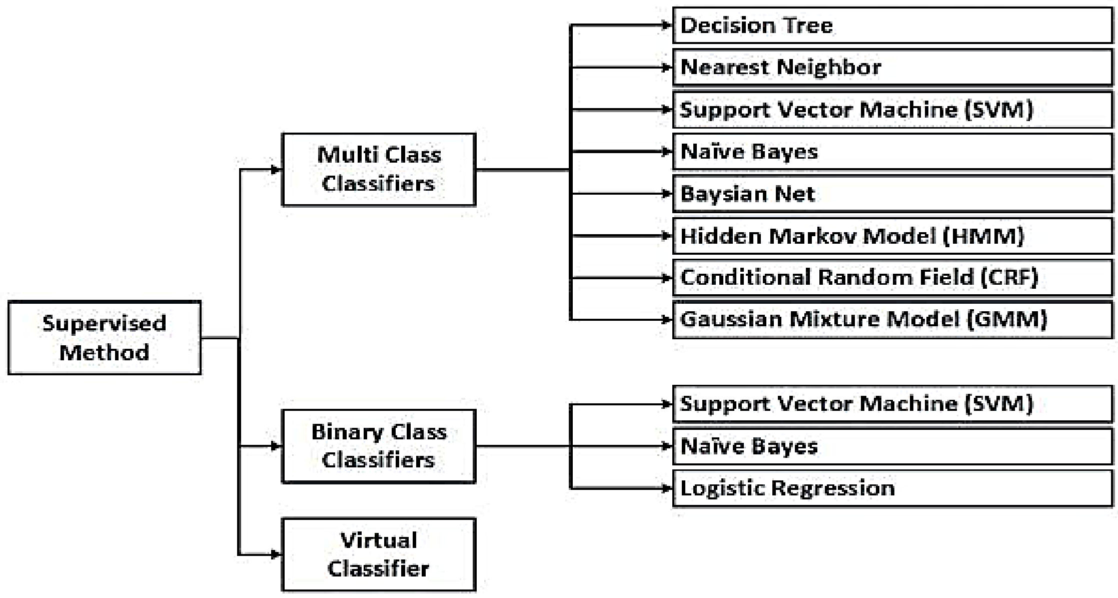


Source: Barros[20]

It is important to show that these algorithms can be employed in the form described: "although When the approach is employed for change point detection, machine learning algorithms can be trained the binary or multi-class classifiers. If the number of states is specified, the change point detection algorithm is trained to find each state boundary"[10].

Therefore, it is in the literature a wide variety of classifiers for machine learning with the aim of use for the detection point of change.

Figure 5: Example of supervised classifiers



Source: Aminikhanghahi et al[10]

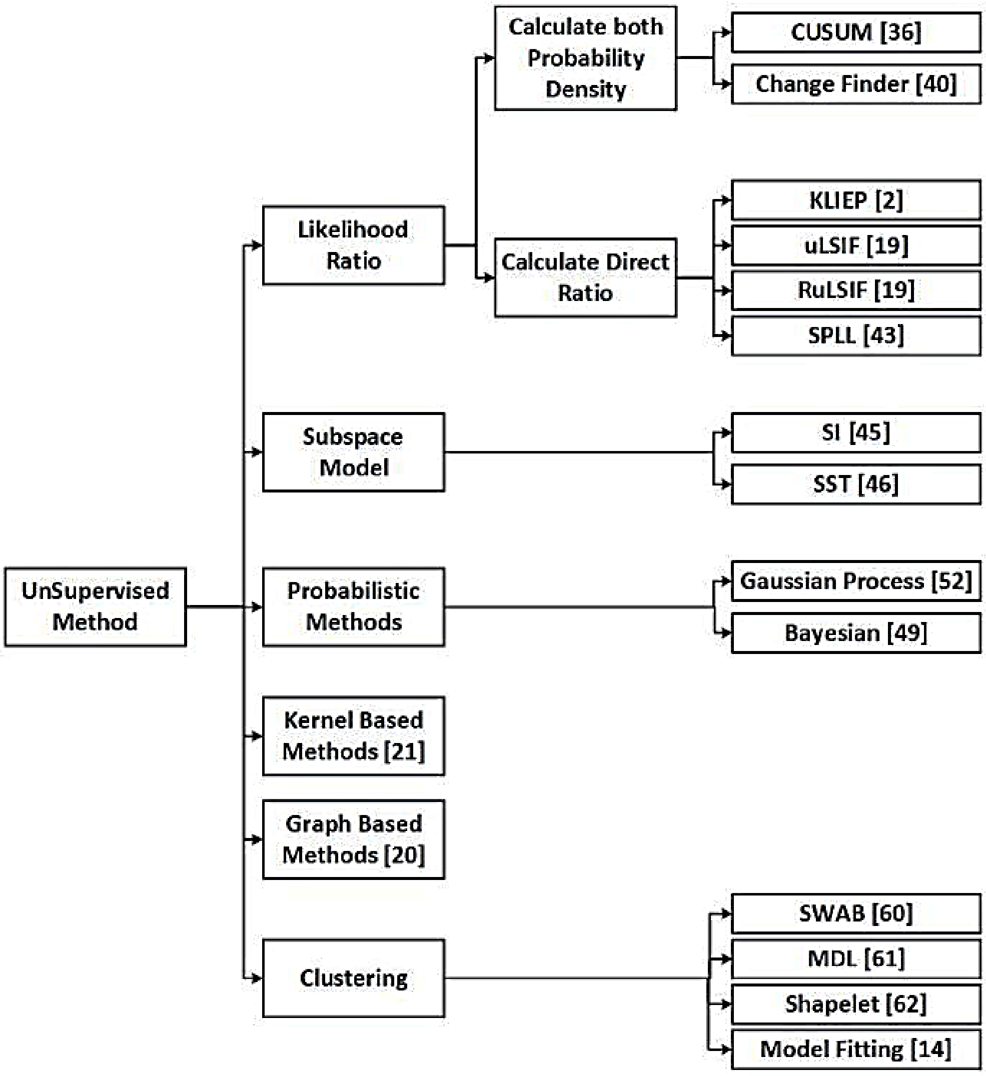
The non-supervised are defined by a system that solves the problem finding automatically from a set of data. I.e., without presenting a point of exit, at the conclusion. The algorithm seeks to find similarity between the data.

In this way, you can collate the data or reduce the size of the data set to obtain the attributes most feasible, in addition to detect trend. In forecasting there feedback to produce results.

In the context can be discorrido "change point detection, such algorithms can be used to segment time series data, thus finding change points based on statistical features of the data. Unsupervised segmentation is attractive because it may handle a variety of different situations without requiring prior training for each situation" [10]

Thus, the most common methods are represented by:

Figure 6: Example of unsupervised classifiers



Source: Aminikhanghahi et al[10]

3.2 THE SEARCH FOR ABRUPT TRANSITION IDENTIFICATION

Identify abrupt transitions depends on the methods used. However, in the literature there are appropriate techniques (algorithm) for a set of data[21].

As mentioned earlier, this section spoke about the model of point detection algorithm for change supervised and unsupervised. Now, there are other techniques which in principle will not be addressed. Therefore, the uncertainties of the temporal series are the challenges to ascertain the best optimization of results apply existing techniques. In the case of models supervised and non-supervised it is necessary to observe, in principle, dimensionality and robustness of the data [21.]

Comparison of the data supervised often outweigh the unsupervised methods in the detection of points of change. Considering that it is necessary feedback to provide output of data that give greater quality in the solution of the problem. However, its limitation is left to the discretion of the training data.

Moreover, the non-supervised operates with limited types of data of temporal series. That is, the temporal series non-stationary, generally, are not applied in the model. In this way, there are models that provide parametric versions for data sets of temporal series stationary not [10].

The corresponding parametric versions use a factor of forgetting to remove the effects of older observations. Example, the data of the univariate time series is the detection point of change can be used a technique called test Page-Hinkley (PHT) linked to the fading factors[1,2].

Page-Hinkley a sequential analysis technology typically used to monitor the detection of changes. The fading factors is a mechanism of gentle reminder that you can assign less weight to observations passadas[1,2].

Another problem is the parametric and non-parametric methods to investigate the performance of the model algorithm in shift point of the historical series. This done, described; in general, the non-parametric methods CPD are more robust than the parametric tests because the parametric methods strongly depend on the choice of parameters. In addition, the problem of CPD becomes more complex for parametric methods when the data have moderate to high dimensionality[10].

For more, it was necessary to check the stationarity of the temporal series, as well as the form of algorithm to be applied; online or offline. Nevertheless, it is understood that the performance of algorithms can be influenced.

Finally, to solve the problem of the temporal series applied to detection of point of change was required a combination of knowledge. For example, there are applied studies in the area such as: "Detecting Multiple Change Points Using Adaptive Regression Splines with Application to Large-Scale Neural Recordings[23], Unusual Time Series Detection [24.], On-line Detection of Continuous Changes in Stochastic Processes[25]" and among others.

Therefore, it has been an ongoing challenge to study the detection point of change in temporal series is stationary, is not stationary, univariate analysis or multivariate analysis. The aim, however, this topic is to explain, as was identified this change and which materials used in this research.

3.3 DATA COLLECTION

The main component is the data that originate from companies that billing over 78 million in the year 2016 in the case of Brazil, eSOCIAL. And CAGED in the last 5 years. For the title of knowledge there are 14.441thousand companies(array) eSOCIAL capable of representing 29% of jobs throughout with contracts in Brazil.

Making a correlation with the data eSOCIAL in for CAGED only in the month of January 2018. In numerical terms, the value is the total of 11,015,488 million, within a universe in total 37,868,331 million bonds regime of all companies in Brazil.

As the data that is divided by establishment (matrix more affiliates) who understand the value of 192,709 thousand establishments of the data of the CAGED in relation to data from eSOCIAL. Within a universe of 3,199,312 million establishments throughout Brazil.

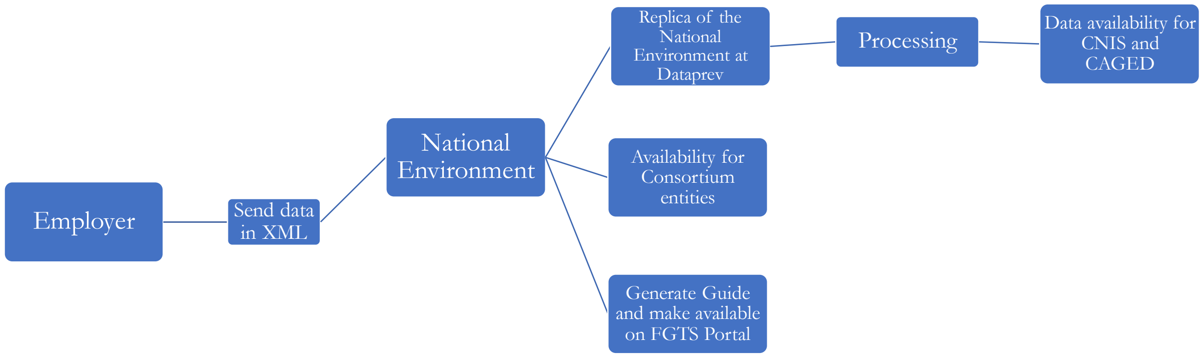
For this study arose the following instruments:

- Documentation;
- Results of data eSOCIAL, of the CAGED;
- Materials for internal and external communications;
- Records in files;
- The layout of the panels eSOCIAL, of the CAGED;
- Use of software programs such as the program R, Python, and software produced on own ministry for proper purposes;
- Strategic and Operating plans of the phenomenon that is studied.

Highlights - that for the purposes of tests in this study, makes it necessary to discuss how they are captured the data from admission, shutdown and the construction of the stock of jobs in both databases.

Thus, the composition for building the eSOCIAL database can be followed as shown below:

Figure 7: Flow eSOCIAL



Source: Prepared by the author (2019)

Identifies himself, that the employer sends events (periodicals and not periodicals) as admission and shutdown for the National Environment, in XML, the SERPRO. Later, the National Environment (SERPRO) receives, validates (according to the rules pre - existing) and stores the information.

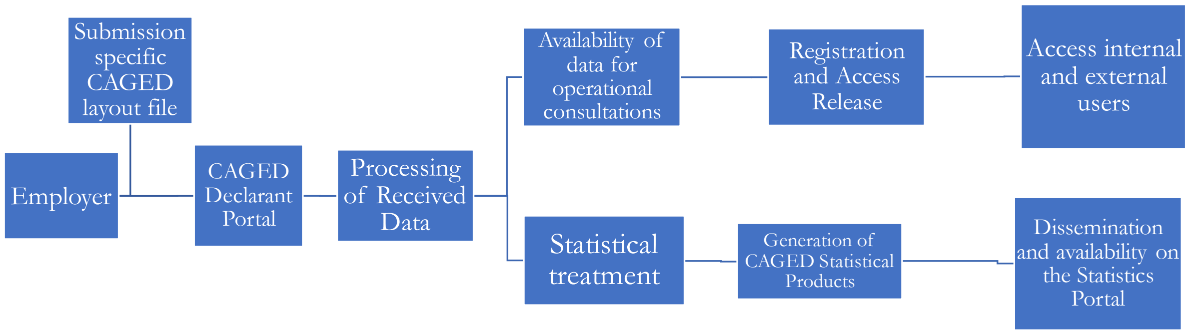
Then follows a Data Lake with eSOCIAL information for Consortium entities, with treatments, which are structuring the events. Therefore, once Dataprev receives and replicates the data, all data mining processes are applied. Importantly, in mining processing may present blocklist anomaly, data inconsistency. Hereafter, events are held back until the problem is resolved.

In addition, results in the generation of the FGTS Guide Portal (for events of shutdown, termination of the FGTS, or of remuneration, monthly gathering 8%. In the interim, Dataprev processes the events received for the National Environment and offers to the CAGED the data eSOCIAL.

In short, as the variables in this study were mentioned above, it is necessary to demonstrate how the eSOCIAL stock is constructed. Note that it is calculated from the sum of all admissions in the reference month, without shutdown. As the basis of the declaration is mandatory in eSOCIAL, companies only made their active employees available in the declaration.

For CAGED, the declaration follows the following model as shown below:

Figure 8: Flow of caged



Source: Prepared by the author (2019)

It is understood that the submission model is performed daily on the declarant's portal for unemployment insurance admissions and monthly, until the 7th of the subsequent month, for other movements, such as dismissals. Meanwhile, the CAGED declarant portal validates (this portal has about 400 input reviews) when it receives the CAGED file, issues a delivery receipt and makes the files available in the operating system in real time.

It then processes the received data by applying data mining techniques. Subsequently, it follows two objectives biases, the first is for operational consultations such as; to assist the citizen in accordance with the proof of their ties and source of law, to grant Unemployment Insurance and for inspection.

Already the second bias, refers to the present study, statistical treatment. In this case, is relevant to the reader and admission, shutdown and the composition of the stock of the employee. In this context, it generates statistical products for purposes of public policies available in the portal of statistics.

Note that the CAGED also builds an stock of employees with the declarations from administrative records that, in fact, is a little more complex than the stock of eSOCIAL, already mentioned. In the case of the CAGED, is made in annual process with the following objectives;

- Delete the reference basis for the stocks of establishments that have closed their activities during the previous year;
- Adjust the stocks with the incorporation of transfers and the correct answers; and,
- Update the geographical and sectoral levels based on the most recent information available.

The methodological procedure applies in the following way:

- New" establishments (those that made their first statement by CAGED) were integrated into the reference base with stock equal to the difference between admitted and dismissed or zero if total dismissals exceeded admissions. (there can be no negative stock);
- The establishments omitted from the RAIS of the previous year (RAIS are administrative records that compose the employees throughout, as well as the public, declared annually, and has the same level of breakdown similar to the caged.). that declared the caged in subsequent year were integrated to the reference base with the stock; and,
- The establishments considered "dead" (those who have not declared to the RAIS previous year nor the CAGED in the subsequent year) had their inventories disregarded for the purpose of Reference Base.

In view of the foregoing, the statements in both bases have different rules, but with the same purpose. In fact, the eSOCIAL has a broader scope, in theory, for the participation of other consortium entities.

However, the present study focuses on variables of admission, shutdowns and stock applied the technique of Page Hinkley and its possible breakage. However, can be explained, the reader, by the anomalies that are recounted below.

3.4 THE CHARACTERISTICS OF THE BASES: POSSIBLE ANOMALY AND THE REALITY IN WHICH THEY ARE INSERTED

The possible anomaly in both bases can be justified by the argument that for CAGED, organizations are required to make the statement, then called administrative records whenever there is a movement between admitted or dismissed so that the Ministry of Economy can know of all eventualities that permeate the relations of admission and dismissal of the job.

It is important to note that for eSOCIAL, this year, organizations are being required to report whether there is admission or dismissal. What happens is that companies are reporting simultaneously, both in CAGED and in eSOCIAL, so it is understood that the search and identification of time series in this paper reflects the non-appearance of abrupt points of change in the broad context.

Since eSOCIAL will replace CAGED. It can be seen, as well as justified, a priori, that in 2018 there may have been omissions on the part of the organizations in general. Considering that there is a bias of the lack of technological capacity of companies in this new reporting tool, once, in this scenario is that there may have been several events and the omission of reporting may not have been able to detect these events.

However, it is important to show that there may be inconsistency in the eSOCIAL base due to the "BlockList". Because, Blocklist is formed by a group of companies whose information / events were not processed in Dataprev due to inconsistencies in the reception of the event sent by SERPRO.

It is noteworthy that the omission of the company itself also permeates the understanding that eSOCIAL is currently considered a supervisory tool, ie it is added to the federal revenue, therefore, when the company reports the dismissal of a collaborator, especially in this type of movement. , the organization must pay to the tax authorities and the employee. However, there may be a lack of capital on the part of companies, for example, by postponing the statement.

It should be noted that theoretically the scope of eSOCIAL must be broader, since it included four government bodies into a single declaration, especially of federal revenue, so in theory, the admissions and off should be higher, even with omissions.

In the meantime, we highlight the legal obligations that eSOCIAL will replace, in the first place, it is necessary to state that the basic objective of this substitution is to compose a single database, with all the information extracted by the organizations.

The numbers of eSOCIAL are quite substantial: 8 million organizations and 80 thousand accounting offices impacted until 2019; 40 million workers covered and 15 obligations which will be replaced gradually. The information exposes the first reflections of the tool from the government that has become mandatory for organizations and they seek, in a general way, adapt to the new way of declaring their data.

It is understood that organizations, in general, periodically declare virtual eSOCIAL information. The replacements, which began to be made from January 2018, are: the FGTS Gathering Guide and Social Security data; CAGED under the CLT regime; RAIS, which is the Annual List of Social Information; The Employee Registration Book -LRE- emphasizes that this registration is now done by digital channels, simplifying the whole process, since such registration only needs to be sent through the eSOCIAL platform; The CAT, which is the Work Accident Report, in this part, it is understood that the CAT is delivered directly by eSOCIAL; The CD, which is Dismissal Communication; CTPS, which is the Work and Social Security Card; PPP, which is the Social Security Profile; the DIRF, which is the Withholding Income Tax Statement, the DCTF, which is the Federal Tax Debts and Credits Declaration concerning the sphere of labor; The data that makes up the work schedule; MANAD, which is the Normative Manual of Digital Archives, even if this is a supervisory archive, eSOCIAL will handle the obligations regarding MANAD .; The payroll; the FGTS Gathering Guide and the FGTS Provisional Gathering Guide; the Social Security Guide will be replaced by DARF, generated by DCTF Web, from the moment of SEFIP replacement, that is, according to the GFIP replacement schedule.

It should be emphasized, in this sphere, eSOCIAL is, indeed, a very impactful project in the sectors of HR organizations, there view, as shown in the screen, he alters several routines and procedures.

It is noteworthy, as explained, from January 2018, the implementation of eSOCIAL began; Therefore, it is seen that the project, funded by the State and several public entities, such as Caixa Econômica Federal, Federal Revenue, Ministry of Labor, among others, seeks to uncomplicate and unite the submission of data on tax, labor and social security obligations throughout the country.

The main reflexes of eSOCIAL can be highlighted below: it is identified that, this way, it will no longer have the possibility of an organization to hire a worker, in theory, without the latter having exposed all his data for such hiring. This data is required to be sent through the eSOCIAL platform at least one day before a new employee begins work at the company that hired him.

There is an exception, only at events that the organization submits the preliminary record, ie admission with the absence of some data. In this scenario, organizations will have to supplement the data until competence ceases. In this way, see, in fact, that have changes. An organization no longer has to hire an employee and it does not sign the wallet. Thus, it is identified that integration between sectors is generated, since eSOCIAL directly requires a broader integration between accounting, tax and labor, even legal sectors.

The data sent, therefore, can no longer be treated differently, that is, organizations may not make any movement of admission or shutdown that is not science of bodies responsible.

Note, in this case, that the relationship with the legal sector, with the entry of Social, will be further strengthened, as the new declaration tool.

It should be noted, this bias, that the information that may be requested by eSOCIAL, are:

- Admission and shutdown of the worker in real time;
- Temporary withdrawal;
- Amendment of the working day;
- Change in the situation of the employee, such as change of salary;
- Early warning;
- ASO (Certificate of Occupational Health);
- Environmental conditions of work that involve part of SESMT (Safety Health and medicine of work);
- CAT (Communication of accident at work) should be informed on the same day in case of death;
- Payroll movements;
- Monitoring the health of the worker.

Given the above, it can be stated to the reader that by trying to identify in a simultaneous regression scan, both based on the unconsolidated eSOCIAL and CAGED of 2018 and in the last five years, facing a priori and possibly obstacles at this time, By 2018, the admissions series, dismiss and, consequently, the stocks may turn out slightly differently, as explained, there are several replacements and many companies have had delays to adapt to these changes.

It is seen that on this basis to be treated is that there is distortion, in this way, seek to apply a technique for such identification and remove their results.

3.5 APPLICATION MODEL

The application model for such a proposition was the Page Hinkley[1,2]. The Page-Hinkley (PH) is a sequential analysis technique commonly used in the identification of monitoring changes in processing of signals[3]. The Page hinkley enables effective identification of abrupt changes and such test can be designed to detect a change in the mean of a Gaussian signal[3]. This test, as Gama et al.,[1] considers a cumulative variable m_T, defined as the accumulated difference between the observed values and your average to the present moment:

$$mT = \sum_{t=1}^{T} (\mathrm{x}t - \overline{x}\overline{T} - \delta)$$

Where e δ is the magnitude of the changes permitted $\bar{x}\overline{T} = 1/T \sum_{t=1}^{t} {}^{xt}$

According to Gama et al., [1] the minimum value of this variable is also computed: MT = min (mt, t = 1.. T). Finally, the test monitors the difference between HT and mT: PHT = mT - MT. According to Gama et al[1] when this difference is higher than the established threshold (λ) a change in distribution is signaled. According to the aforementioned authors, it is emphasized that the threshold λ is directly dependent on the allowable false alarm rate, that is, when it rises λ, it implies a smaller number of false alarms, but may lose or delay the detection of alterations[1,2].

In the meantime, it is understood that Page-hinkley can trace a Bayesian pre-sequential error, as well as identify any significant changes[1]. The PH and its progress in the statistical test can be verified with its detection threshold, therefore, according to Gomes et al., [1] and Hartl et al., [3] it is emphasized that the parameter λ must allow the algorithm to be resistant to false alarms, so the test can detect and react to changes as they occur, attenuating the detection delay time. Finally, by controlling this detection threshold parameter a balance can be established between false positive alarms[1].

3.6 TEST APPLICATION

For the present study, the code for the test was designed in Python. The exposed code was extracted from an online library. The delimiter for this selection, a priori, was to verify if with code it is possible to verify the noise in the time series, and it was identified as univariable, that is, it allows an analysis of one time series at a time.

The initial algorithms to "turn" the test did not accept short time series, in this way, the author used the same test in a specific library (scikt-multiflow) that has this test is already configured. Thus, the author used Anaconda, which is a free distribution of open source programming languages Python and R, to simplify the management and deployment of the codes.

Thus, the pre-established parameters, have been delta that is an exponent of the magnitude of changes permitted, lambda, which is the threshold of difference and the last alpha parameter that corresponds to the variable of adaptation to adjust lambda. For this method, we defined the input value for the detection of deviation in the Page Hinkley. It should be emphasized that the test of Page-Hinkley, in synthesis, scans of the temporal series point-to-point to find the break. However, this test can only be applied in a series at a time. Soon, the author had to repeat the test in all grades.

It stands out, in this context, and stresses to the reader that the test of Page-Hinkley does not compare temporal series, since it only makes the evaluation of a series of univariate form, as described in this study. The objective is, in the view of this, that the ideal is the Page Hinkley detect abrupt changes in series in nearby points, once this ideal is realized you can expound on such a coincidence. Therefore, it can be seen that the result of this analysis is to point.

3.7 TEST SIZING

For the generation of the measures that have been used for the creation of temporal series, if made necessary the creation of some parameters

- The period in which these measures were measured
- The variables used for the generation of these measures.
- Define the code to be applied

To test separately, the libraries were debts on importation, with loading vectors for the CAGED, as well as for the eSOCIAL.

For the vectors regarding the CAGED and eSOCIAL, we tested the admissions, shutdowns, and stock of the caged in the year 2018 in group by region, these being: south, southeast, north, northeast, central-west and the complex encompassing these sections, as well as for the CAGED was done the same loading vectors between the years of 2014 to 2018.

In order to refine the test parameters, apply it and consider a way to propose mean, median, minimum and maximum, as well as perform a regression on the data to identify some relevant result in this proposition, the author executed the following library codes, as well as made the refinement on an example basis:

Skmultiflow library code and example as per annex A

3.8 MATERIALS

For the development of this research were needed infrastructure materials of the researcher and materials of subsidy for research. On the materials of infrastructure, which involves various aspects, are not necessarily mapped, but were the responsibility of the researcher himself, who has all the necessary for the development of research, such as: material for food, leisure, security, housing etc. with respect to the materials of subsidies for research, lists: personal computer of the researcher "notebook", which has access to the worldwide network of computers and specific statistical programs, such as Python.

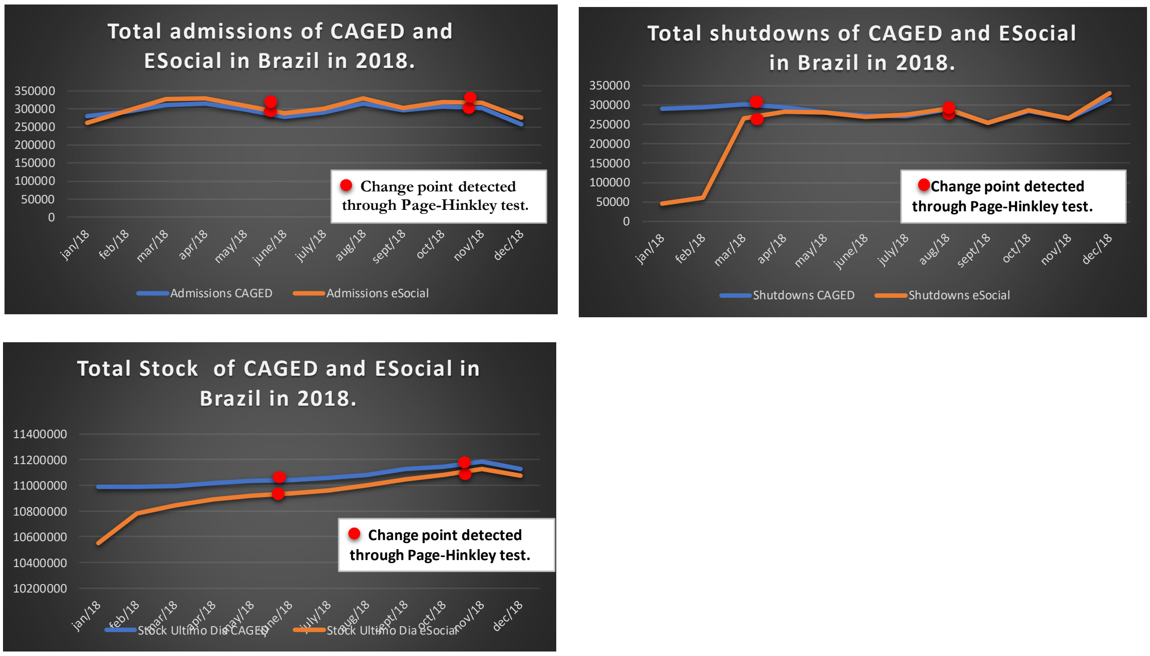
With respect to human resources, it was necessary to use auxiliary plant for the development of the research that has been done entirely by the researcher, with the use of own resources, withdrawn from the exercise of his activity as a business analyst. In the same direction, on the material resources, in addition to those already described above, the researcher used the “skmultiflow library”.

4.RESULTS

A correlation between eSOCIAL and CAGED was performed. Below, with the application of the test, the points of change are highlighted. For the analysis of eSOCIAL against CAGED in the sample base from 2014 to 2018, in Annex A of the project, as well as the codes that generated the following figures.

The following figures show CAGED and eSOCIAL total admissions, shutdown and stock in Brazil in 2018.

Figure 9 - Total number of admissions, shutdowns, and stock of the caged and and social in Brazil in 2018



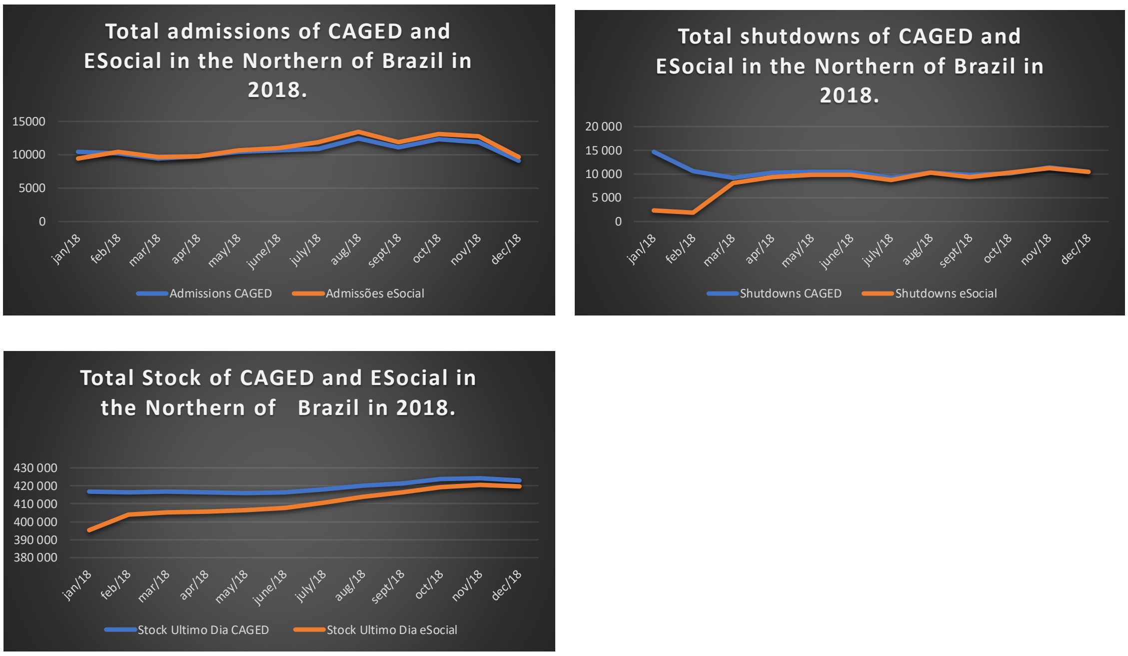
Source: Prepared by the author (2019)

It identifies, in these graphs that with the application of the test of Page-Hinkley, there were points of change in the exploitation of the three variables worked. It should be noted that for the admissions of CAGED and eSOCIAL in Brazil in 2018, the test identified two points of change in both databases relating to the months of May and October. Now for the sweep over the total shutdown of the CAGED and eSOCIAL in Brazil in 2018.

The application of the test Page-Hinkley identified two points of change in both bases in the months of February and August. For the total stock Page-Hinkley test identified two points of change in both bases in the months of May and October.

The following figures expose the total number of admissions, shutdown and stock of CAGED and eSOCIAL in Brazil, in the northern region, in the year 2018.

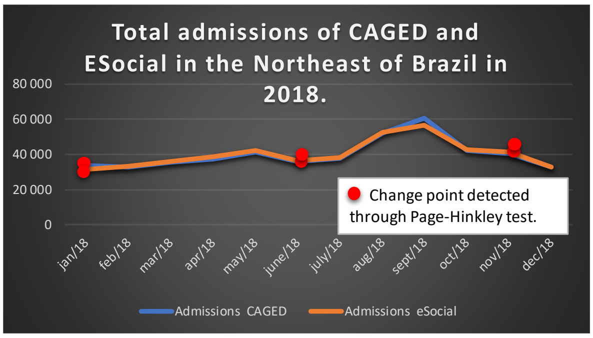
Figure 10 - Total number of admissions, shutdowns, and stock of the CAGED and eSOCIAL in Brazil in the northern region, in the year of 2018

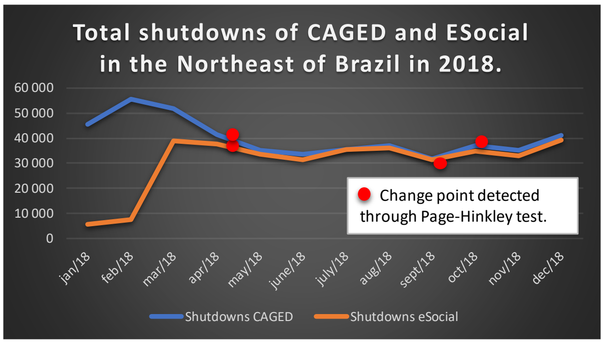


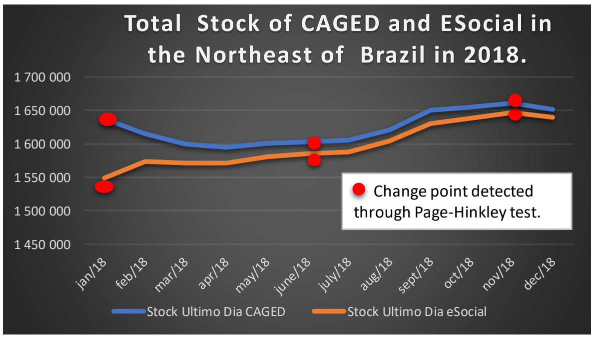
Source: Prepared by the author (2019)

You see, these graphs that with the application of the test of Page-Hinkley there was no identification of points of change in the exploitation of the three variables worked for the northern region of Brazil in both databases.

Figure 11 - Total number of admissions, shutdowns, and stock of the caged and and social in Brazil, in the northeast region, in the year of 2018







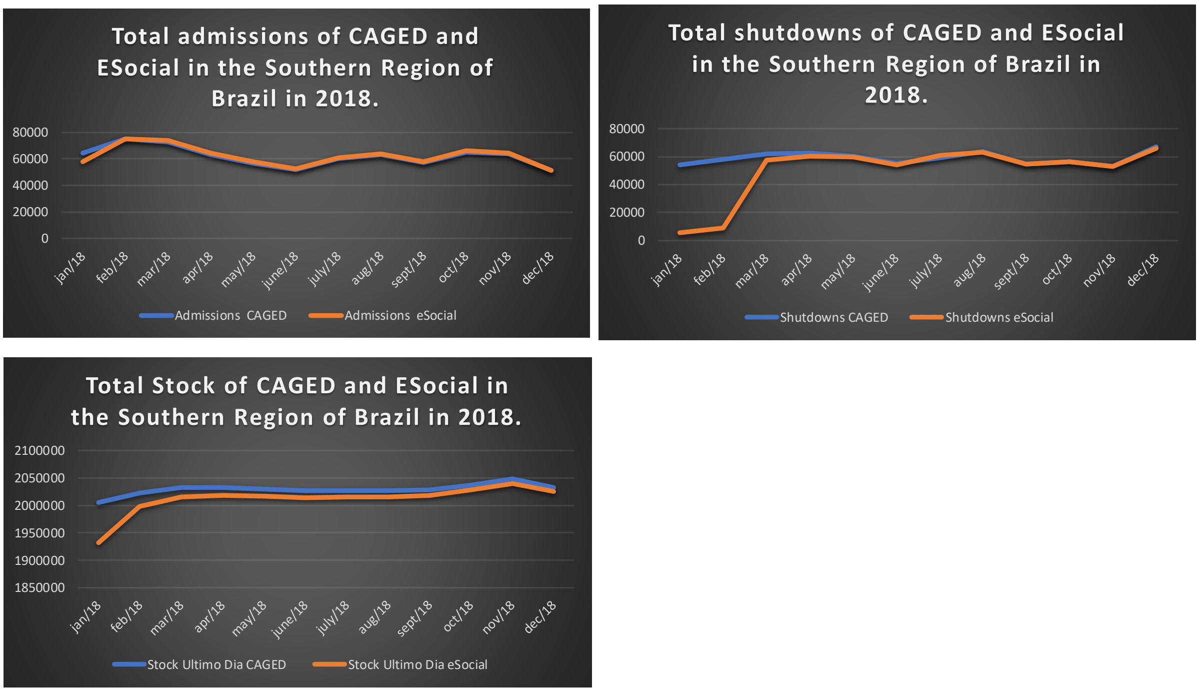
Source: Prepared by the author (2019)

These graphs that with the application of the test of Page-Hinkley there points of change in the exploitation of the three variables worked. It should be noted that for the admissions of CAGED and eSOCIAL in the northeast region of Brazil in 2018, the test identified three points of change in both databases relating to the months of January, February and November. Now for the sweep over the total shutdown of the CAGED and eSOCIAL in Brazil in 2018.

The application of the test Page-Hinkley identified three points of change, being the first point in both bases in the month of April, already for the month of September there was identification of shift point to shutdown of eSOCIAL, and for the month of October there was identification of the point of change for the shutdown of the CAGED. For the total stock Page-Hinkley test identified three points of change in both bases in the months of January, February and November.

The following charts exposes the total number of admissions, shutdown and stock of CAGED and eSOCIAL in Brazil, in the southern region, in the year 2018. See if, in these graphs that with the application of the test of Page-Hinkley there was no identification of points of change in the exploitation of the three variables worked in both bases for the southern region of Brazil.

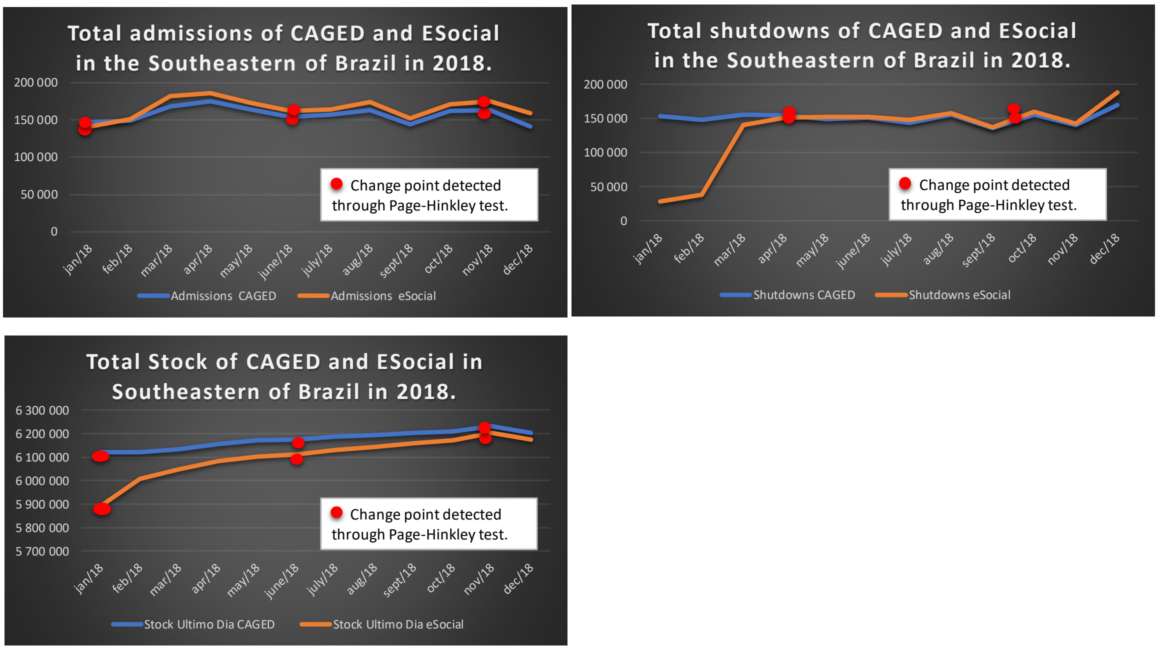
Figure 12 - Total number of admissions, shutdowns, and stock of the CAGED and eSOCIAL in Brazil, in the southern region, in the year of 2018



Source: Prepared by the author (2019)

The following figures expose the total number of admissions, shutdown and stock of CAGED and eSOCIAL in Brazil, in the southeastern region, in the year 2018.

Figure 13 - Total number of admissions, shutdowns, and stock of the CAGED and eSOCIAL in Brazil, in the southeastern region, in the year of 2018



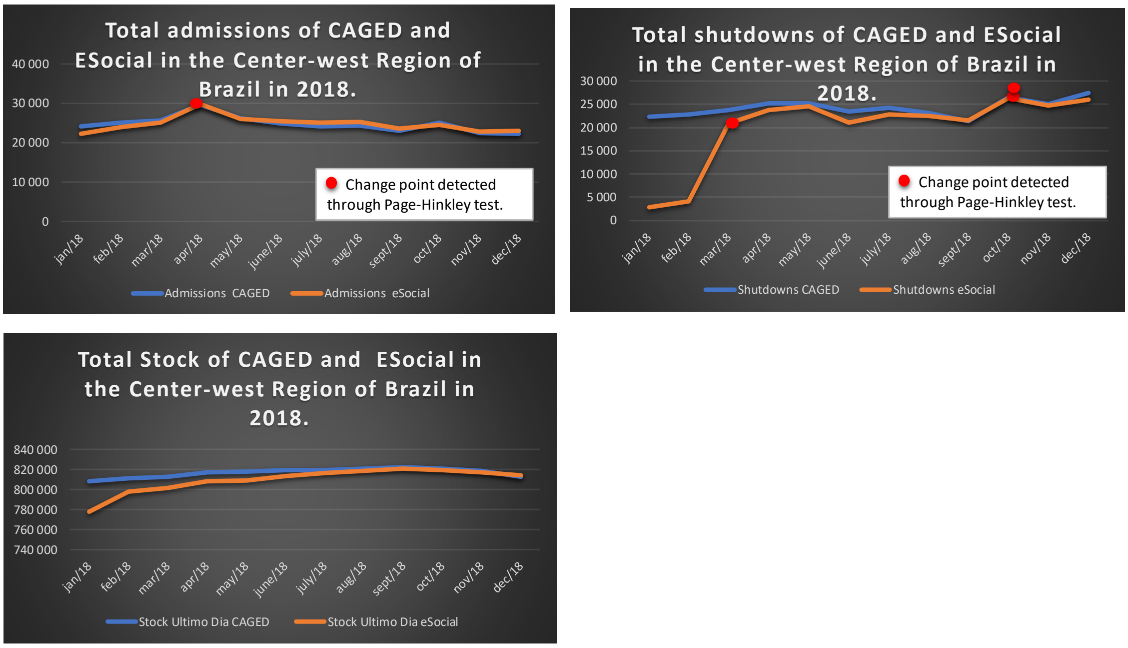
Source: Prepared by the author (2019)

These graphs that with the application of the test of Page-Hinkley there points of change in the exploitation of the three variables worked. It should be noted that for the admissions of CAGED and eSOCIAL for the southeast region of Brazil in 2018, the test identified three points of change in both databases relating to the months of January, February and November. Now for the sweep over the total shutdown of the CAGED and eSOCIAL in Brazil in 2018.

The application of the test Page-Hinkley identified two points of change in both bases for the months of April and October. For the total stock Page-Hinkley test identified three points of change in both bases in the months of January, February and November.

The following figures expose the total number of admissions, shutdown and stock of CAGED and eSOCIAL in Brazil, in the center-west region, in the year 2018.

Figure 14 - Total number of admissions, shutdowns, and stock of the CAGED and eSOCIAL in Brazil, in the center-west region, in the year of 2018



Source: Prepared by the author (2019)

You see, these graphs that with the application of the test of Page-Hinkley there points of change in the exploitation of the three variables worked. It should be noted that for the admissions of CAGED and eSOCIAL in the center-west region of Brazil in 2018, the test identified a shift point, this being for the month April on the basis of eSOCIAL already for the sweep over the total shutdown of the CAGED and eSOCIAL in the center-west region of Brazil in 2018.

The application of the test Page-Hinkley identified two points of change, being that in February the shift point is identified by the shutdown eSOCIAL, already in October, the shift point is identified in both databases. For the total stock Page-Hinkley test did not identify points of change.

It should be noted that, through the Page-Hinkley test were detected points of change, however, as above, were insignificant.

5. DATA ANALYSIS

There are two main objectives in time series analysis: the first is to identify the nature of the phenomenon represented by the sequence of observations, and the second to predict, that is, to predict future values of the time series variable. Both goals require that the observed time series data pattern be identified as well as described. Once the pattern is established, it can be interpreted and integrated with other data, that is, use it in the theory of the phenomenon investigated in this study. It is understood that regardless of the author's depth of understanding and the validity of the interpretation (theory) of the phenomenon, one can extrapolate the identified pattern to predict future events.

For the analysis of the data generated using the Page-Hinkley technique, it is important to highlight that for CAGED products at Brazil level, it was found that at admission there was no break in time, while in the shutdown only presented a break in time and stock showed two point breaks. In the meantime, it is emphasized to the reader that the basic justification for these occurrences permeates the understanding of the unique characteristics of such a technique: how Page-Hinkley that sweeps in a point-to-point time series, finding any point of change is due to any of the three factors that make up the possible rupture of the series, these being seasonality, trend and noise.

However, it notes that these breaks within the 60 months seen in CAGED from 2014 to 2018, both in admission, shutdown and stock, were an unimpressive result. Thus, it is understood that such detected changes did not interrupt a break that could occur.

Therefore, it is identified that these small interruptions give rise to why CAGED as a well-established database. In the correlation with eSOCIAL 2018 and CAGED 2018 in Brazil and the Regions, with the same variables, that is, checking the admissions, layoffs and stock, there were few breaks, so it is understood that eSOCIAL is well considering that the breaks occurred are quite identical.

From the reasoning that there are few breaks, therefore, there is a high similarity, so when applying this technique, it is clear that eSOCIAL can replace CAGED. In fact, there is no break in the historical series of CAGED and this is the central point of this study, because the product of this investigation has such similarity, it is concluded that the data are very robust, adjusted.

In a note, it is noteworthy that the three variables investigated in this exploration, admission, dismissal and stock are the main components of a declaration of an administrative record for the organization, with stock being the balance between admission and dismissal and the latter. The investigated variable was also identical when it broke.

Therefore, as they had 2 break points in a 12-month time series, see **figure 9**, and when they occurred they were similar in both CAGED and eSOCIAL, it can be said that these breaks are small anomalies, since Administrative records can be either declared or omitted by organizations, while some companies are not technologically prepared to make the declaration, as well omitted by the Blocklist barrier, others may have financial difficulties to make such a record and seasonality that employability offers.

6. FINAL CONSIDERATIONS AND FUTURE WORK

In general, it is considered that the methodology proved consistent and concordant with the results generated. It should be emphasized that the validations at the end of each step covered good results noting the efficiency of Page-Hinkley technique for such an approach.

The package of algorithms that the author used by means of a specific library pre-configured framework scikit-multiflow, proved to be efficient and valid for the detection of abrupt changes in the temporal series of this study. It should be noted, within this framework, that the automated method for checking segmented form univariate analysis enabled the conference in its detail. Therefore, to observe the temporal series front their spectral curves, generally, even the most tenuous, it can be inferred, even with the largest clippings, that the Page-Hinkley operative technique was operative to generate such results.

The predominance of congruent tracing in the time series of eSOCIAL data compared to CAGED was expected. To the author's perception that the new declaration tool is inclined to a more oversight, once it is profitable, simple and compulsory, i.e., in many respects is analogous to CAGED. However, and social is more comprehensive, considering that centralizes the ancillary obligations labor. Soon, to raise information about admissions, shutdowns and the stock and unite the series in the same basis of CAGED, it was interesting to note points of changes, however, these were not significant, considering that understands that there is seasonality, as well as have omissions, as explained in the analysis of this study.

The Mappings of the peculiarities of the tracings generated by the algorithm proved pertinent to predetermined parameters, i.e., to the exponent generated the magnitude of changes permitted, as well as the threshold of difference and the variable adaptation to such adjustment, it should be pointed out that when the author makes the refinements needed adjustments in this action is exactly to expose such changes and demonstrate the effectiveness of the technique that proved useful in aiding the understanding of these phenomena.

A suggestion for future work is the breakdowns of administrative records of data per unit of the Federation of Brazil (UF) and sectors of economic activity, as well as the cross-sectional study of organizational behavior forward the declaration obligations The administrative data seeking the causes and the reflexes of the taken decisions in procrastination in the Brazilian economic scenario.

BIBLIOGRAPHICAL REFERENCES

1) GAMA, J.R. SEBASTIÃO P.T,. RODRIGUES, P. P **Issues in evaluation of stream learning algorithms**," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, 2009, p. 329.

(2) GAMA, J; SEBASTIÃO, R; RODRIGUES, P.P On evaluating stream learning algorithms · **Mach Learn** (2013) 90:317–346 DOI 10.1007/s10994-012-5320-9 Published online: 24 October 2012

(3) MOUSS, H., MOUSS, D., MOUSS, N., & SEFOUHI, L. (2004). Test of Page-Hinkley, an approach for fault detection in an agro-alimentary production system. In **Proceedings of the Asian control conference** (Vol. 2, pp. 815–818).

(4) BEZERRA, M. I. S **Análise de séries temporais**. São Paulo 2006.Disponível em: http:pt.scribd.com/doc/55232564/Apostila-Series-Temporais.html. acessado em 05 de jun 2019

(5) BRAMER, Max. **Principles of data mining**. [s.l]: Springer, 2013.

(6) RIGO, S. J.; CAZELLA, S. C.; CAMBRUZZI, W. **Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades**. In: Anais do Workshop de Desafios da Computação Aplicada à Educação. 2012.

(7) LATORRE, M. do R.D, CARDOSO, M.R.A Análise de séries temporais em epidemiologia: uma introdução sobre os aspectos metodológicos. **Rev Bras. Epidemiol.** Vol. 4, no 3, pp 145-152, Nov. 2001

(8) MORETTIN, P. A TOLOI, C. M .C **Análise de séries temporais**. São Paulo: Edgard Blücher, 2006.

(9) NOMELINI, Q.S.S; FERREIRA, E.B; NOGUEIRA, D.A; GOLYNSKI, A,A; VILLA, T.E Uso de Modelagem Univariada e Multivariada com Séries Temporais como Ferramenta de Gestão do Agronegócio na Cultura de Soja do Brasil Using Univariate and Multivariate Modeling with Time Series as Agribusiness Management Tool in Soybean Culture of Brazil. **Rev. Espacios**. Vol.38. ano 2017

(10) AMINIKHANGHAHI, S; COOK, D.J A Survey of Methods for Time Series Change Point Detection., **Knowl. Inf. Syst**., vol. 51, no. 2, pp. 339–367, May 2017

(11) GARCIA, Alan Cássio. **Mineração de dados aplicada a sistemas de recomendação.** [Trabalho de Conclusão de Curso]. Santa Cruz do Sul: Universidade de Santa Cruz do Sul, 2012.

(12) DIGGLE, P **Time series : a biostatistical introduction**. Clarendon Press, 1990.

(13) AGUIRRE, A Uma introdução à análise espectral de séries temporais econômicas **Nov. Econ**. Vol. 5 no. 1. pp. 41-60, 1995

(14) MORAIS, E. F. **Análise de Regressão Versus Séries Temporais: Uma previsão para o Índice Ibovespa. Monografia** (Graduação) – Universidade Estadual da Paraíba, Campina Grande, 2010.

(15) EHLERS, R. S. **Análise de Séries Temporais**. 2007. Departamento de Estática, UFRP.Acesso em: 1 ago. 2019 Disponível em: http://www.each.usp.br/rvicente/ AnaliseDeSeriesTemporais.pdf. acessado em 02 jun 2019

(16) TAN, P.; STEINBACH, M.; KUMAR, V.; FERNANDES, **A. Introdução ao datamining: mineração de dados.** Rio de Janeiro: Editora Ciência Moderna, 2009.

(17) TURCHI, Sandra. **A Sazonalidade do consumo**. São Paulo, ago. 2011. Seção Artigos, negócios, varejo. Disponível em: . Acesso em 1 de agosto 2019.

(18) HIRANO, S; TSUMOTO, S. **Detection of risk factors as temporal data mining**. In: PAKDD WORKSHOPS, p. 143–156. Springer-Verlag, 2008.

(19) MANHÃES, L. M. B.; CRUZ, S. M. S.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. Instituto de Ciências Exatas – Universidade Federal Rural do Rio de Janeiro (UFRRJ), 2011.

(17) DOWNEY, A. B. **A novel changepoint detection algorithm** Dec. 2008.

(18) CHEN, H. ZHANG, N. **“Graph-based change-point detection**,” *Ann. Stat.*, vol. 43, no. 1, pp. 139–176, Feb. 2015.

(19) ZHANG, H . CAO R, S. ZILBERSTEIN, F. Wu, and X. Chen, “Optimal Composition of Real-Time Systems**,” Artif Intell**., 1996, pp. 181–213.

8(20) BARROS, B. “**Aprendizagem de Maquina: Supervisionada ou Não Supervisionada?,”** 2016**.** [Online]. Disponivel em: https://medium.com/opensanca/aprendizagem-de-maquina-supervisionada-ou-não-supervisionada-7d01f78cd80a. [acessado : 15-Jan-2019].

(21) GOSWAMI. B *et al.*, **“Abrupt transitions in time series with uncertainties.,”** *Nat. Commun.*, vol. 9, no. 1, p. 48, 2018.

HARTL, C., BASKIOTIS, N., GELEIA, S., & SEBAG, M. (2007). Change point detection and meta-bandits for online learning in dynamic environments. In **Conférence Francophone sur l’apprentissage automatique**, Cepadues (pp. 237–250)

(22) MIYAGUCHI K. YAMANISHI K, **“Online detection of continuous changes in stochastic processes**,” *Int. J. Data Sci. Anal.*, vol. 3, no. 3, pp. 213–229, May 2017.

(23) H. Toutounji and D.Durstewitz, “detecting multiple change points Using Adaptive

Regression splines with Application to Neural Recordings, “front.Neuroinform.,vol.12,p 67,Oct.2018.

(24) R. J. Hyndman, E. Wang, and N. Laptev, “Large-Scale Unusual Time Series Detection,” in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1616–1619.

(25) K. Miyaguchi and K. Yamanishi, “Online detection of continuous changes in stochastic processes,” Int. J. Data Sci. Anal., vol. 3, no. 3, pp. 213–229, May 2017.

ANNEX A

Skmultiflow library code

```
from skmultiflow.drift_detection.base_drift_detector import BaseDriftD
etector


class PageHinkley(BaseDriftDetector):
    """ Page-Hinkley method for concept drift detection.

    Notes
    -----
    This change detection method works by computing the observed
    values and their mean up to the current moment. Page-Hinkley
    won't output warning zone warnings, only change detections.
    The method works by means of the Page-Hinkley test [1]_. In genera
l
    lines it will detect a concept drift if the observed mean at
    some instant is greater then a threshold value lambda.

    References
    ----------
    .. [1] E. S. Page. 1954. Continuous Inspection Schemes.
       Biometrika 41, 1/2 (1954), 100-115.

    Parameters
    ----------
    min_instances: int (default=30)
        The minimum number of instances before detecting change.
    delta: float (default=0.005)
        The delta factor for the Page Hinkley test.
    threshold: int (default=50)
        The change detection threshold (lambda).
    alpha: float (default=1 - 0.0001)
        The forgetting factor, used to weight the observed value
        and the mean.

    Examples
    --------
    >>> # Imports
    >>> import numpy as np
    >>> from skmultiflow.drift_detection import PageHinkley
    >>> ph = PageHinkley()
    >>> # Simulating a data stream as a normal distribution of 1's and
 0's
    >>> data_stream = np.random.randint(2, size=2000)
    >>> # Changing the data concept from index 999 to 2000
    >>> for i in range(999, 2000):
    ...     data_stream[i] = np.random.randint(4, high=8)
    >>> # Adding stream elements to the PageHinkley drift detector and
 verifying if drift occurred
    >>> for i in range(2000):
    ...     ph.add_element(data_stream[i])
    ...     if ph.detected_change():
    ...         print('Change has been detected in data: ' + str(data_
stream[i]) + ' - of index: ' + str(i))
```

```
    """
    def __init__(self, min_instances=30, delta=0.005, threshold=50, al
pha=1 - 0.0001):
        super().__init__()
        self.min_instances = min_instances
        self.delta = delta
        self.threshold = threshold
        self.alpha = alpha
        self.x_mean = None
        self.sample_count = None
        self.sum = None
        self.reset()

    def reset(self):
        """ reset

        Resets the change detector parameters.

        """
        super().reset()
        self.sample_count = 1
        self.x_mean = 0.0
        self.sum = 0.0

    def add_element(self, x):
        """ Add a new element to the statistics

        Parameters
        ----------
        x: numeric value
            The observed value, from which we want to detect the
            concept change.

        Notes
        -----
        After calling this method, to verify if change was detected, o
ne
        should call the super method detected_change, which returns Tr
ue
        if concept drift was detected and False otherwise.

        """
        if self.in_concept_change:
            self.reset()

        self.x_mean = self.x_mean + (x - self.x_mean) / float(self.sam
ple_count)
        self.sum = max(0., self.alpha * self.sum + (x - self.x_mean -
self.delta))

        self.sample_count += 1

        self.estimation = self.x_mean
        self.in_concept_change = False
        self.in_warning_zone = False

        self.delay = 0
```

```
        if self.sample_count < self.min_instances:
            return None

        if self.sum > self.threshold:
            self.in_concept_change = True
```

Example:

```python
# Importa a biblioteca
import numpy as np
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
```

```python
# Simula um data stream com uma distribuição normal de 1's e 0's
data_stream = np.random.randint(2, size=2000)
```

```python
# Imprime a lista com os dados criados
for x in range (len(data_stream)):
    print (data_stream [x],end=" ")
```

```python
# Altera a concepção dos dados a partir do índice 999 até o índice 2000
for i in range(999, 2000):
    data_stream[i] = np.random.randint(4, high=8)
```

```python
# Imprime a lista com os dados modificados
for x in range (len(data_stream)):
    print (data_stream [x],end=" ")
```

```python
# Adicionando elementos de fluxo ao detector de desvio PageHinkley e verificando se ocorreu algum desvio
for i in range(2000):
    ph.add_element(data_stream[i])
    if ph.detected_change():
        print('Change has been detected in data: ' + str(data_stream[i]) + ' - of index: ' + str(i))
```

```
Change has been detected in data: 6 - of index: 1004
```

CAGED 2014 TO 2018 - Brazil - Admission, shutdown and stock.

In [35]:
```
# Importando as bibliotecas
import numpy as np

## Biblioteca disponível em https://scikit-multiflow.github.io/scikit-multiflow/_autosummary/skmultiflow.drift_detectio
n.PageHinkley.html#skmultiflow.drift_detection.PageHinkley
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
# Parâmetros utilizados no teste de Page-Hinkley (delta, threshould e alpha são valores defaut)
# (min_instances=6, delta=0.005, threshold=50, alpha=1 - 0.0001)
```

In [36]:
```
# Carregando vetores para o CAGED - Anos 2014 a 2018 (60 meses)
Admissoes_CAGED_2014to2018_BR = [396070,485214,439310,470203,444999,404482,422142,426702,425902,409474,409465,340382,35
7718,410644,435099,394669,356100,333403,319505,323038,318953,304527,293600,248743,256008,301729,329369,310902,291063,27
7585,280221,305265,280110,277837,278514,239429,265936,290657,306551,288634,294341,283089,281956,294437,278917,302723,28
9955,248681,280588,293050,311633,314701,298826,277705,289681,315301,296630,306793,302081,257147]
Desligamentos_CAGED_2014to2018_BR = [403671,412126,386232,424658,419241,382389,396955,401584,382867,420132,393234,43386
6,401510,403493,408410,423171,402262,385980,373305,370097,333254,349409,314456,393866,324765,350270,352763,331340,31517
5,311991,302845,310390,287060,290375,287318,350544,308678,312240,310749,275140,287123,284991,271316,301775,272384,27345
4,265663,312709,289792,295256,302635,294229,280187,273405,271964,289592,254208,285692,265074,315958]
Estoque_CAGED_2014to2018_BR = [11065862,11138950,11192028,11237573,11263331,11285424,11310611,11335729,11378764,1136810
6,11384337,11290853,11518139,11525290,11551979,11523477,11477315,11424738,11370938,11323879,11309578,11264696,11243840,
11098717,11180600,11132059,11108665,11088227,11064115,11029709,11007085,11001960,10995010,10982472,10973668,10862553,10
988838,10967255,10963057,10976551,10983769,10981867,10992507,10985169,10991702,11020971,11045263,10981235,10989277,1098
7071,10996069,11016541,11035180,11039480,11057197,11082906,11125328,11146429,11183436,11124625]
```

In [37]:
```
# Verificando se ocorreram mudanças na série temporal - Adminissões - Brasil
for i in range(60):
    ph.add_element(Admissoes_CAGED_2014to2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2014to2018_BR[i]) + ' - of index: ' + st
r(i))
```

In [38]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Desligamentos - Brasil
for i in range(60):
    ph.add_element(Desligamentos_CAGED_2014to2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2014to2018_BR[i]) + ' - of index: '
+ str(i))
```

```
Mudança foi detectada na série temporal:403671 - of index: 0
```

In [39]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Estoque - Brasil
for i in range(60):
    ph.add_element(Estoque_CAGED_2014to2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2014to2018_BR[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:11065862 - of index: 0
Mudança foi detectada na série temporal:11183436 - of index: 58
```

In [ ]:

In [ ]:

eSOCIAL 2018 - Brazil and the region - Admission and Shutdown.

```
# Importando as bibliotecas
import numpy as np

## Biblioteca disponível em https://scikit-multiflow.github.io/scikit-multiflow/_autosummary/skmultiflow.drift_detectio
n.PageHinkley.html#skmultiflow.drift_detection.PageHinkley
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
# Parâmetros utilizados no teste de Page-Hinkley (delta, threshould e alpha são valores defaut)
# (min_instances=6, delta=0.005, threshold=50, alpha=1 - 0.0001)
```

```
# Carregando vetores para o E-Scocial
Admissoes_Esocial_2018_BR = [262476,293923,326572,328526,309536,288067,300143,329023,301940,318597,316390,275755]
Desligamentos_Esocial_2018_BR = [45517,60560,266051,282659,280340,269331,275443,290043,254350,287760,265158,330429]
Admissoes_Esocial_2018_Sul = [57610,74866,73798,64055,57773,52322,60930,63783,57772,66363,64127,50973]
Desligamentos_Esocial_2018_Sul = [5733,8817,57609,60269,60024,54053,60631,62994,54508,56311,53126,65781]
Admissoes_Esocial_2018_Sudeste = [141686,151117,181988,186306,172842,162619,164181,173789,152022,171539,175464,159204]
Desligamentos_Esocial_2018_Sudeste = [29095,38199,140362,151384,152303,152975,147832,158210,137548,159957,143100,18857
2]
Admissoes_Esocial_2018_Norte = [9450,10410,9616,9803,10623,11048,11855,13451,11870,13089,12783,9658]
Desligamentos_Esocial_2018_Norte = [2248,1789,8105,9401,9821,9860,8722,10352,9372,10255,11272,10526]
Admissoes_Esocial_2018_Nordeste = [31378,33473,35969,38450,42178,36564,37997,52581,56553,42860,41200,32909]
Desligamentos_Esocial_2018_Nordeste = [5538,7663,39028,37808,33516,31374,35388,35947,31323,34810,32810,39356]
Admissoes_Esocial_2018_Centro_Oeste = [22323,24022,25161,29854,26090,25477,25054,25365,23651,24617,22777,22966]
Desligamentos_Esocial_2018_Centro_Oeste = [2895,4082,20910,23741,24621,21014,22812,22480,21539,26368,24808,26015]
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Brasil
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_BR[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:309536 - of index: 4
Mudança foi detectada na série temporal:318597 - of index: 9
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Brasil
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_BR[i]) + ' - of index: ' + st
r(i))
```

```
Mudança foi detectada na série temporal:266051 - of index: 2
Mudança foi detectada na série temporal:290043 - of index: 7
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Sul
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_Sul[i]) + ' - of index: ' + str
(i))
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Sul
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_Sul[i]) + ' - of index: ' + s
tr(i))
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Sudeste
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_Sudeste[i]) + ' - of index: ' + s
tr(i))
```

```
Mudança foi detectada na série temporal:141686 - of index: 0
Mudança foi detectada na série temporal:162619 - of index: 5
Mudança foi detectada na série temporal:175464 - of index: 10
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Sudeste
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_Sudeste[i]) + ' - of index: '
+ str(i))
```

```
Mudança foi detectada na série temporal:151384 - of index: 3
Mudança foi detectada na série temporal:159957 - of index: 9
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Norte
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_Norte[i]) + ' - of index: ' + str
(i))
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Norte
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_Norte[i]) + ' - of index: ' +
str(i))
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Nordeste
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_Nordeste[i]) + ' - of index: ' +
str(i))
```

```
Mudança foi detectada na série temporal:31378 - of index: 0
Mudança foi detectada na série temporal:36564 - of index: 5
Mudança foi detectada na série temporal:41200 - of index: 10
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Nordeste
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_Nordeste[i]) + ' - of index:
' + str(i))
```

```
Mudança foi detectada na série temporal:37808 - of index: 3
Mudança foi detectada na série temporal:31323 - of index: 8
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Centro-Oeste
for i in range(12):
    ph.add_element(Admissoes_Esocial_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_Esocial_2018_Centro_Oeste[i]) + ' - of index:
' + str(i))
```

```
Mudança foi detectada na série temporal:29854 - of index: 3
```

```
# Verificando se ocorreram mudanças na série temporal do E-Social - Centro-Oeste
for i in range(12):
    ph.add_element(Desligamentos_Esocial_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_Esocial_2018_Centro_Oeste[i]) + ' - of ind
ex: ' + str(i))
```

```
Mudança foi detectada na série temporal:20910 - of index: 2
Mudança foi detectada na série temporal:26368 - of index: 9
```

CAGED 2018 - Brazil and the region - Admission and Shutdown.

```
[ ]: # Importando as bibliotecas
import numpy as np

# Biblioteca disponível em https://scikit-multiflow.github.io/scikit-multiflow/_autosummary/skmultiflow.drift_detectio
n.PageHinkley.html#skmultiflow.drift_detection.PageHinkley
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
# Parâmetros utilizados no teste de Page-Hinkley (delta, threshould e alpha são valores defaut)
# (min_instances=6, delta=0.005, threshold=50, alpha=1 - 0.0001)
```

```
[ ]: # Carregando vetores para os dados do CAGED de 2018
Admissoes_CAGED_2018_BR = [280588,293050,311633,314701,298826,277705,289681,315301,296630,306793,302081,25714]
Desligamentos_CAGED_2018_BR = [289792,295256,302635,294229,280187,273405,271964,289592,254208,285692,265074,315958]
Admissoes_CAGED_2018_Sul = [64370,75275,72555,63239,56490,51833,59949,63418,57263,65236,63891,51435]
Desligamentos_CAGED_2018_Sul = [53970,57936,62236,62470,60463,54958,59387,63697,54718,56639,52929,67010]
Admissoes_CAGED_2018_Sudeste = [147608,149438,168208,174886,164477,154193,156905,162780,144600,161793,163309,141730]
Desligamentos_CAGED_2018_Sudeste = [153372,148423,155515,154787,149010,150974,143714,155390,136593,155474,140307,16964
0]
Admissoes_CAGED_2018_Norte = [10446,10258,9499,9730,10397,10705,10881,12510,11102,12346,11939,9151]
Desligamentos_CAGED_2018_Norte = [14756,10644,9285,10317,10420,10539,9199,10331,9767,10089,11430,10531]
Admissoes_CAGED_2018_Nordeste = [33900,32912,35730,37222,41279,35974,37689,52227,60592,42366,40524,32630]
Desligamentos_CAGED_2018_Nordeste = [45382,55407,51759,41429,35109,33531,35370,37006,31798,36942,35267,41260]
Admissoes_CAGED_2018_Centro_Oeste = [24264,25167,25641,29624,26183,25000,24257,24366,23073,25052,22418,22201]
Desligamentos_CAGED_2018_Centro_Oeste = [22312,22846,23840,25226,25185,23403,24294,23168,21332,26548,25141,27517]
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Brasil
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_BR[i]) + ' - of index: ' + str(i))
```

```
Mudança foi detectada na série temporal:298826 - of index: 4
Mudança foi detectada na série temporal:306793 - of index: 9
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Brasil
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_BR[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:302635 - of index: 2
Mudança foi detectada na série temporal:289592 - of index: 7
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Sul
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_Sul[i]) + ' - of index: ' + str(i))
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Sul
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_Sul[i]) + ' - of index: ' + str
(i))
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Sudeste
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_Sudeste[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:147608 - of index: 0
Mudança foi detectada na série temporal:154193 - of index: 5
Mudança foi detectada na série temporal:163309 - of index: 10
```

```
[ ]: # Verificando se ocorreram mudanças na série temporal do CAGED - Sudeste
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_Sudeste[i]) + ' - of index: ' +
str(i))
```

```
Mudança foi detectada na série temporal:154787 - of index: 3
Mudança foi detectada na série temporal:155474 - of index: 9
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Norte
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_Norte[i]) + ' - of index: ' + str
(i))
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Norte
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_Norte[i]) + ' - of index: ' + s
tr(i))
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Nordeste
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_Nordeste[i]) + ' - of index: ' + st
r(i))
```

```
Mudança foi detectada na série temporal:33900 - of index: 0
Mudança foi detectada na série temporal:35974 - of index: 5
Mudança foi detectada na série temporal:40524 - of index: 10
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Nordeste
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_Nordeste[i]) + ' - of index: '
+ str(i))
```

```
Mudança foi detectada na série temporal:41429 - of index: 3
Mudança foi detectada na série temporal:36942 - of index: 9
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Centro-Oeste
for i in range(12):
    ph.add_element(Admissoes_CAGED_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Admissoes_CAGED_2018_Centro_Oeste[i]) + ' - of index: '
+ str(i))
```

```
: # Verificando se ocorreram mudanças na série temporal do CAGED - Centro-Oeste
for i in range(12):
    ph.add_element(Desligamentos_CAGED_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Desligamentos_CAGED_2018_Centro_Oeste[i]) + ' - of inde
x: ' + str(i))
```

```
Mudança foi detectada na série temporal:26548 - of index: 9
```

```
:
```

eSOCIAL 2018 - Brazil and the region - Stock

```
# Importando as bibliotecas
import numpy as np

# Biblioteca disponível em https://scikit-multiflow.github.io/scikit-multiflow/_autosummary/skmultiflow.drift_detectio
n.PageHinkley.html#skmultiflow.drift_detection.PageHinkley
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
# Parâmetros utilizados no teste de Page-Hinkley (delta, threshould e alpha são valores defaut)
# (min_instances=6, delta=0.005, threshold=50, alpha=1 - 0.0001)
```

```
# Carregando vetores para os dados do Estoque do Esocial de 2018
Estoque_Esocial_2018_BR = [10549072,10782435,10842956,10888823,10918019,10936755,10961455,11000435,11048025,11078862,11
130094,11075420]
Estoque_Esocial_2018_Sul = [1932257,1998306,2014495,2018281,2016030,2014299,2014598,2015387,2018651,2028703,2039704,202
4896]
Estoque_Esocial_2018_Sudeste = [5893146,6006064,6047690,6082612,6103151,6112795,6129144,6144723,6159197,6170779,620314
3,6173775]
Estoque_Esocial_2018_Norte = [395124,403745,405256,405658,406460,407648,410781,413880,416378,419212,420723,419855]
Estoque_Esocial_2018_Nordeste = [1548460,1574270,1571211,1571853,1580515,1585705,1588314,1604948,1630178,1638228,164661
8,1640171]
Estoque_Esocial_2018_Centro_Oeste = [777501,797441,801692,807805,809274,813737,815979,818864,820976,819225,817194,81414
5]
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Brasil
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_BR[i]) + ' - of index: ' + str(i))
```

```
Mudança foi detectada na série temporal:10918019 - of index: 4
Mudança foi detectada na série temporal:11078862 - of index: 9
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Sul
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_Sul[i]) + ' - of index: ' + str(i))
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Sudeste
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_Sudeste[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:5893146 - of index: 0
Mudança foi detectada na série temporal:6112795 - of index: 5
Mudança foi detectada na série temporal:6203143 - of index: 10
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Norte
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_Norte[i]) + ' - of index: ' + str
(i))
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Nordeste
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_Nordeste[i]) + ' - of index: ' + st
r(i))
```

```
Mudança foi detectada na série temporal:1548460 - of index: 0
Mudança foi detectada na série temporal:1585705 - of index: 5
Mudança foi detectada na série temporal:1646618 - of index: 10
```

```
# Verificando se ocorreram mudanças na série temporal do Esocial - Centro-Oeste
for i in range(12):
    ph.add_element(Estoque_Esocial_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_Esocial_2018_Centro_Oeste[i]) + ' - of index: '
+ str(i))
```

CAGED 2018 - Brazil and the region - Stock

In [2]:
```
# Importando as bibliotecas
import numpy as np

# Biblioteca disponível em https://scikit-multiflow.github.io/scikit-multiflow/_autosummary/skmultiflow.drift_detectio
n.PageHinkley.html#skmultiflow.drift_detection.PageHinkley
from skmultiflow.drift_detection import PageHinkley
ph = PageHinkley()
# Parâmetros utilizados no teste de Page-Hinkley (delta, threshould e alpha são valores defaut)
# (min_instances=6, delta=0.005, threshold=50, alpha=1 - 0.0001)
```

In [3]:
```
# Carregando vetores para os dados do Estoque do CAGED de 2018
Estoque_CAGED_2018_BR = [10989277,10987071,10996069,11016541,11035180,11039480,11057197,11082906,11125328,11146429,1118
3436,11124625]
Estoque_CAGED_2018_Sul = [2004684,2022023,2032342,2033111,2029138,2026013,2026575,2026296,2028841,2037438,2048400,20328
25]
Estoque_CAGED_2018_Sudeste = [6121312,6122327,6135020,6155119,6170586,6173805,6186996,6194386,6202393,6208712,6231714,6
203804]
Estoque_CAGED_2018_Norte = [416942,416556,416770,416183,416160,416326,418008,420187,421522,423779,424288,422908]
Estoque_CAGED_2018_Nordeste = [1637841,1615346,1599317,1595110,1601280,1603723,1606042,1621263,1650057,1655481,1660738,
1652108]
Estoque_CAGED_2018_Centro_Oeste = [808498,810819,812620,817018,818016,819613,819576,820774,822515,821019,818296,812980]
```

In [4]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Brasil
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_BR[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_BR[i]) + ' - of index: ' + str(i))
```

```
Mudança foi detectada na série temporal:11035180 - of index: 4
Mudança foi detectada na série temporal:11146429 - of index: 9
```

In [5]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Sul
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_Sul[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_Sul[i]) + ' - of index: ' + str(i))
```

In [6]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Sudeste
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_Sudeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_Sudeste[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:6121312 - of index: 0
Mudança foi detectada na série temporal:6173805 - of index: 5
Mudança foi detectada na série temporal:6231714 - of index: 10
```

In [7]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Norte
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_Norte[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_Norte[i]) + ' - of index: ' + str(i))
```

In [8]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Nordeste
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_Nordeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_Nordeste[i]) + ' - of index: ' + str
(i))
```

```
Mudança foi detectada na série temporal:1637841 - of index: 0
Mudança foi detectada na série temporal:1603723 - of index: 5
Mudança foi detectada na série temporal:1660738 - of index: 10
```

In [9]:
```
# Verificando se ocorreram mudanças na série temporal do CAGED - Centro-Oeste
for i in range(12):
    ph.add_element(Estoque_CAGED_2018_Centro_Oeste[i])
    if ph.detected_change():
        print('Mudança foi detectada na série temporal:' + str(Estoque_CAGED_2018_Centro_Oeste[i]) + ' - of index: ' +
str(i))
```

In [ ]: