Middle East Journal of Science

https://dergipark.org.tr/mejs

MEJS

e-ISSN:2618-6136

Research Article

TIME SERIES OUTLIER ANALYSIS FOR MODEL, DATA AND HUMAN-INDUCED RISKS IN COVID-19 SYMPTOMS DETECTION

Ahmet Kaya [1] *Rojan Gümüş* [2] *Ömer Aydın** [3]

[1] Ege University, Tire Kutsan Vocational School, Tire, İzmir, Turkey
[2] Dicle University, Ataturk Vocational School of Health Services, Diyarbakır, Turkey
[3]Manisa Celal Bayar University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Manisa, Turkey
* Corresponding author; omer.aydin@cbu.edu.tr

Abstract: *Information systems are important references aiming to support the decisions of decision-makers. Information reliability depends on the accuracy and efficacy of data and models. Therefore, some risks may emerge in information systems concerning models, data, and humans. It is important to identify and extract outliers in decision support systems developed for the health information systems such as the detection system of Covid-19 symptoms. In this study, the risks that are important in decision making in Covid-19 symptom detection were determined by the statistical time series (ARMA) approach. Potential solutions are proposed in this way. Moreover, outliers are detected by software developed by using the Box-Jenkins model, and the reliability and accuracy of data are increased by using estimated data instead of outliers. In the implementation of this study, time-series-based data obtained from laboratory examinations of Covid-19 test devices can be used. With the method revealed here, outliers originating from healthcare workers or test apparatus can be detected and more accurate results can be obtained by replacing these outliers with estimated values.*

Keywords: *Covid-19, Health information systems, time series, outlier analysis, ARMA*

Received: July 12, 2021 Accepted: October 26, 2021

1. Introduction

Health institutions are very important centers where financial values and risks must be managed effectively because, for all health institutions, it is essential to employ qualified personnel, purchase expensive equipment and machinery and use full-fledged buildings. In addition, it is particularly important to manage information systems accurately and use them effectively. A transformation is realized on the way from data to information and such transformation is implemented by information systems [1]. Thus, useful and meaningful outputs can be achieved by transforming so-called raw data into information [2]. Transformation into the required strategic information after a line of transactions such as preparation, processing, and communication of raw information is realized thanks to the information systems [3]. It is obvious that information systems also have a strategic role for organizations. Information systems are highly important in the development of products, services, and qualifications needed by organizations to gain an advantage over their rivals [4].

Strategic information systems are described as instruments using development, practice, transformation, and communication of organizational strategies [5]. Thus, information systems are a set of associated elements generated to enable control and coordination in the organization and collect, process, store and disseminate data to be used in decision-making [6].

Information system, on the other hand, is a series of organized transactions that support decision-making and provide control. An information system has a 7-phase hierarchical structure that contributes to the decision-making process. This structure is as described below [7]:

(1) Data collection (through observation or measurement),

(2) Data organization,

(3) Data processing,

(4) Outputs,

(5) Transformation from outputs into information,

(6) Presentation to the decision-maker,

(7) Decision-making (contributed by the expertise of decision-maker).

As physicians manage and supervise diagnosis, treatment, and control processes of a disease, they are often required to make accurate and effective decisions. The Decision-making process aims to encourage versatile thinking and transform uncertain environments into relatively stable environments. On the other hand, it is necessary to take into consideration the conditions of patients who will be positively or negatively affected by the decisions made, as well as the potential gains, losses, and risks. Decisions taken by physicians have a broad area of impact. These decisions are virtually managerial and each decision results in an array of changes and costs. Changes are generally directed toward increasing the competitive edge and efficiency and can be used to create new areas of health investment [8].

Health institutions are significant tools in the transition from conventional therapy methods to modern therapy [9]. These institutions have developed a major reliance on information and communication technologies from the past to the present [10]. This reliance is undoubtedly out of necessity. The most significant tools of information technologies are information systems. Data entered into information systems must be protected to maintain the reliability and efficacy of these information systems. Such protection is called security. In fact, data entered into health information systems are data regarding patients and, thus, they must be given the utmost care and consideration [10]. While electronic data enable health care organizations to identify areas of strength and weakness within their own operational environments, sometimes unintended negative consequences can occur even among the highest-functioning healthcare systems [11]. Data concerning a patient contain clues used in the diagnosis and treatment of the disease and are of vital importance to the patient. If such clues cannot be protected or become biased due to human or instrument-induced errors, then a physician who makes decisions based upon such data may be mistaken. As a result, diagnosis and treatment decisions will be erroneous as well. In such cases, serious vital risks occur for the patient. The physician and health institution may also inevitably encounter problematic processes.

Such errors not only put the patient's life in danger, but they also cause irreparable loss of prestige for health institutions that need to be managed in a serious manner. Furthermore, the hospital is faced with numerous question marks, and the health institution and physician encounter a series of material and immaterial sanctions. In addition, this kind of news is presented to large populations via mass communication tools, media devices, and social media platforms. It is apparent that the diagnostic decisions of physicians who are significant elements of health institutions have a broad area of influence. Prior to making decisions bearing very significant material and vital risks for both health institutions and physicians, and most importantly for the patients, data utilized in disease diagnosis and treatment must be subjected to a series of control processes before their entry into the information systems [12].

Information that is acquired from information systems and contributes to the decision-making process may contain models, data, and human-induced outliers. However accurately the information system and the physician works, such errors may result in biases in information generation and cause physicians who are in the decision-making position to make wrong decisions. Data that provide a basis for models, thus for information, maybe misentered by individuals, or if data is obtained via measurement and weighing instruments, misuse of such instruments may cause a bias on data. Also, data is by its nature acquired in different ways than expected, which is explained as an expression of an unforeseen outlier [13].

In this study, a method is proposed for verifying the accuracy of data used in health information systems. Statistical time series (ARMA) approach and Box-Jenkins model were used in the method. The method was applied to data known as Box-Jenkins A series and outliers were detected and eliminated.

This study uses time series analysis in health information systems. Despite the fact that the method was used in different kinds of fields, no studies, projects, or reports using the method on health data are available in the literature. In this regard, it is considered that the present study will serve as a source for this type of study.

2. Literature Review

Time series is a highly effective forecasting and analysis tool in statistics and various other disciplines in recent years, in relation to the analysis of datasets obtained in equal and unequal time intervals. Datasets obtained at equal time intervals are defined as discrete time series, and datasets obtained at unequal time intervals are defined as continuous time series. Detection of outliers in time series is known as outlier analysis. Outliers in time series were first studied by Fox in 1972 [15]. Fox developed a method called the likelihood ratio test to detect outliers (effects) in autoregressive (AR) models and defined the outliers detected with this method as first and second-type outliers. Fox also conducted studies on the power functions of outliers. Henceforth, many researchers developed methods encompassing all ARIMA models for detection of multiple outliers based on and building upon, Fox's studies. As well as Hilmer (1984), Tsay (1986), Pena (1987), Abraham and Yatawara (1988), Chang, Tiao, and Chen (1988), and Bruce and Martin (1989) contributed to the literature with similar studies [13,16-20].

Besides, Abraham and Yatawara studied the method of Lagrange multipliers and score-based outlier tests in 1988. Pena (1987), Abraham and Chuang (1989), Bruce and Martin (1989) studied and published articles about tests depending on the elimination of outlier observations during outlier detection and effective observations in time series [21].

Initial studies on the effect of outliers in time series were conducted by Box and Tiao in 1975 [22]. Another study on the detection of outliers is the method called Robust Procedure developed by Denby and Martin in 197923. This method was further studied by Martin and Yohai in 1985 [24]. Chang, Tiao, and Chen (1988) revealed in 1988 that incorrectly detected outliers lead to a loss of efficacy of test methods. Monte-Carlo simulations were employed to detect the C value which is the critical value used in outlier detection.

Time series is a serious area of research in statistics. It is also an analytical tool frequently used in quality control procedures, genetic algorithm optimization, fuzzy logic studies, import and export forecasting processes, chemical concentration analyses, and all time-dependent optimizations [25]. Also, predictive models like AR, MA, ARMA, ARIMA, and SARIMA through Box-Jenkins methodology have largely evolved in the second half of the 20. Century [26]. These models were used in different fields like health, economy, technology, transportation, and environmental areas [27-31].

Aydın and Karaarslan proposed a digital twin-based health information system in their study. In that study, image and sound data are obtained by mobile phone. Similarly, information such as body temperature and saliva samples are obtained through various sensors. Data from different sources are represented by a digital twin created in the cloud, where it is aimed to determine whether the person is Covid-19 by artificial intelligence and machine learning techniques [32].

The COVID-19 outbreak caused radical changes in public life. Various measures are being tried to prevent the spread of the virus. More than 225,000 deaths and 3.2 million cases occurred during the study by Usman and his friends. Early detection of Covid-19 symptoms will contribute to preventing the outbreak from expanding. In this way, sick individuals can be quarantined and prevented from transmitting the disease to others. Usman et al. looked into the possibility of using speech to detect COVID-19 symptoms at an early stage. With this study, a low-cost and ubiquitous solution is proposed with early diagnosis. No complicated and expensive medical devices or specialized medical professionals are required for the preliminary assessment of symptomatic individuals. COVID-19 symptoms can be detected by an application running on a mobile device. Thus, health authorities can be warned [33].

In the past period, online sites and chatbots have been developed to control the symptoms of COVID-19. Since there is no study that statistically evaluates the accuracy of COVID-19 symptom controllers, Munsch et al. conducted a study. In their study, 10 COVID-19 symptom controllers, which are available online for free between 3-9 April 2020, have been selected. Versions of these symptom controllers in this date range were used. Updates after that date were not considered for analysis. They developed two additional simple symptom controllers as a basis for performance evaluation of 10 online COVID-19 symptom controllers. These two controllers evaluate and weigh the presence of COVID-19 symptom frequencies provided by the World Health Organization based on vector distance (SF-DIST) and cosine similarity (SF-COS) [34].

Mackey et al. aim to identify and characterize user-generated conversations that can be addressed in disease recovery using COVID-19 symptoms, test access experiences, and an unattended approach to machine learning. They collected and examined the tweets between 3-20 March 2020 from Twitter. In these tweets, the words related to COVID-19 were filtered. Subject clusters consisting of these words were analysed using an unsupervised machine learning approach called Biterm Topic Model (BTM). The tweets in these clusters were then removed and manually explained for content analysis and evaluated for their statistical and geographical features. They collected a total of 4,492,954 tweets containing terms that could be related to COVID-19 symptoms. They identified 3465 (<1%) tweets with filtering. They analysed these tweets [35].

An application-based self-reporting tool has been created to identify the distribution pattern and possible unreported symptoms by Zens et al. From April 8 to May 15, 2020, they used an app installed on smartphones by 22327 people. Participants are asked to enter information through questionnaires. In this way, they gathered information on both disease histories and symptoms of COVID-19. With this information reported by the participants, it becomes easier to identify new symptoms of COVID-19 disease and to predict the predictive value of some symptoms. In this way, it helps to develop reliable scanning tools. According to the data obtained in this study, they emphasized the necessity of loss of smell and taste as a cardinal symptom and showed that diabetes is a risk factor for the highly symptomatic course of COVID-19 infection [36].

3. Materials and Methods

In recent years there has been considerable interest in the effect of outlying observations in models. Robust techniques have been developed to reduce the effect of such observations. Hence, a large number of tests and procedures have been developed for outlier detection. Most of these procedures operate in the following ways [14]:

(1) Sequentially, examining the most deviant observation first and considering other observations only when the first is beyond the threshold.

(2) Without regard to the differing influence observations may have on the parameter estimates or predicted values which are often the prime focus of the analysis.

When an outlier is detected, the analyst is faced with a number of questions [14]:

(1) Is the measurement process out of control?

(2) Is the model wrong?

(3) Is some transformation required?

(4) Is there an identifiable subset of the observations that is important in its different behavior?

These issues affect the interpretation and confidence in the resulting estimates and/or predictions.

In this study, outlier values were tried to be determined in the evaluation of the data used in the detection of COVID-19 symptoms. For this chemical concentration value known as Box-Jenkins, A series were used to perform error analysis on the data. All systems and details about the materials and methods used are given in the subsections below.

3.1. Information Systems

A system is defined as a whole and a set of associated elements, with inputs and outputs and with predetermined borders, made up of interacting parts put together to achieve a goal or purpose [6].

The definition of system is a general definition likely to also include information systems. Information systems are specific-purpose systems developed to achieve specific goals. Information systems are featured, special-purpose software that transforms data into tangible information.

Data which refers to the input in information systems represent events occurring in the organization or the physical environment. Data does not have the characteristics and qualities of information by itself and cannot be used as information. However, information obtained by association and processing of data attained as a result of an observation or transaction is usable. Data cannot be used as a reference by itself. For instance, a physician uses large amounts of data in disease diagnoses. Blood values, urine values, blood pressure, and heartbeat counts, and if required, X-ray and similar instruments are evaluated together and an opinion is formed on the diagnosis of the disease. Each value mentioned above is unable to go beyond being data only. An opinion built upon the combination of all data is qualified as information and it supports the decision made by the physician.

3.1.1 Information Systems

Information systems are named in accordance with their functions and purposes. In this sense, six different information systems are available as listed below [6].

These information systems are Transaction Processing System (TPS), Management Information System (MIS), Office Automation Systems (OAS), Decision Support Systems (DSS), Executive Information Systems (EIS), Knowledge Work Systems (KWS), and Health Information Systems (HIS).

It is also possible to come across numerous special-purpose information systems other than those mentioned above and developed outside these categories. The common input of all information systems is data [37].

The healthier the data is, the more reliable and accurate the information will be. Data security depends on the attention and experience of individuals obtaining and entering such data into the system and on the reliability of systems. Since it is not always likely to operate a process with individuals doing their jobs properly, it is essential to test whether the data is healthy and ensure that they are reliable. To this end, it is necessary to ensure the compatibility of data forming the basis for modeling with an appropriate model and to operate an algorithm testing the presence of any outliers on the data. If there are any outliers on the data, then it is required to determine their effects and eliminate such effects (errors) and optimize data. Through this process, the compatibility of data constituting the information can be ensured and consistent results can be achieved.

3.1.2 Strategic Role of Information Systems

Information systems have some significant strategic roles. The first of these roles is to help organizations establish a lasting advantage over their competitors, and the other role is to make technological innovations factors of production in the right place at the right time. Furthermore, the integration of these innovations into the business processes facilitates adaptation to technological changes, enables the transition to automation, and paves the way for obtaining higher quality data at a cheaper price. Although such adaptation is costly and difficult in the initial phase, its gains over time are ample and noteworthy. According to the information systems literature, it is generally regarded that strategic information systems have two types of benefits [38]:

- The first one is information-based gains obtained from a number of technological ideas specific to the organization and earned through creative ideas.
- The second one is already existing benefits that can be acquired by everyone and whose strategic significance is understood through extensive and effective use.

Although the importance of information systems for organizations was not understood in the past, they are becoming increasingly important due to higher competition and unstoppable technological changes.

3.1.3 Advantages of Health Information Systems

Health information systems are software dynamics that are important for an organization at all levels [39]. Given their strategic roles, information systems provide a series of advantages to health institutions. These advantages include:

- Fast access to high-quality information,
- Adaptation to technological changes,
- Cooperation with scientific methods,
- Resistance to compete,
- Opportunity to follow innovations and adapt them to businesses,
- Accurate diagnosis and satisfaction,
- High profits and chance to grow,
- Ease of adaption to changes,
- Remodelling opportunity,
- Setting a basis for research and development activities.

3.2. Statistical Modeling Process

The most frequently used statistical software packages today include SPSS, MINITAB, SAS, STATISTICA, and so on. Model parameters are generated in order to eliminate the potential outliers from data by using these packages which are a kind of information system. In this study, Statistical modeling transactions are performed using MINITAB software package. Data retrieved for this purpose is regarded as time series (ARMA) (Auto Regressive Moving Average Model) data.

3.3. Box-Jenkins and Outlier Forecasting Model

Outlier modeling is a process for eliminating the bias occurring on the data because the result of modeling with biased data is necessarily biased. In this regard, it is necessary to remove the biases occurring on the data and it should be done before modeling. This process is also called data cleaning.

Data Cleaning is the process that consists of detecting and imputing anomalous data1. In this context, two types of outliers are observed in Box-Jenkins time series forecasting models [15]. Type 1 errors (Additive Outlier (AO)) are induced by individuals, devices, or erroneous use of devices. Effects of these outliers which are not data-related must certainly be separated from observations [41]. Type 2 errors (Innovational Outlier (IO)) occurs as a result of natural randomness and affects all observations starting from their emerging with a decreasing trend. There are differing opinions on whether the effect of this second type of error should be separated or not. It is observed that errors in time series can be identified with their sources and reasons [13,15].

In addition, error detection in time series is based on autocorrelation established between residuals (e_t) (ρ: autocorrelation $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t^*$ $(E(\varepsilon_t, \varepsilon_{t-1}) \neq 0)$ $E(\varepsilon_t) = e_t$), and the Type 1 error detection can be done much more easily by data scanning processes [41].On the other hand, in Type 1 error, the effect on parameter estimation is much higher and is defined as a shock effect [40].

Box-Jenkins time series model is defined as follows:

Assume that $\{x_t\}$ is a time series generated with $ARMA(p,q)$ a model containing no data error. $ARMA(p,q)$ Model is described as:

$$\phi(B)x_t = \theta(B)e_t \tag{1}$$

where, $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$, $\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$, $B^k x_t = x_{t-k}$ $E(x_t) = 0$ $\{e_t\} \rightarrow (0, \sigma^2)$. If $\phi_1, ..., \phi_p$ values which are the roots of $\phi(B)$ function defined in the equation model are outside the unit circle, then the model fulfills the stationary condition; if $\theta_1, ..., \theta_p$ values which are the roots of $\theta(B)$ function are outside the unit circle, then it fulfills the nonstationary condition, and thus, reversibility assumptions [42].

The stationarity of time series simply refers to keeping the correlation between observations and error terms within prescribed limits and to the decrease of partial autocorrelation values of observations in parallel with the increase in lags. In addition, if the series is stationary, the assumptions that are prerequisites for time series modeling will be met. In the Box-Jenkins model, error types can be classified as human-induced errors, data-induced errors, and model-induced errors.

3.4. Human-Induced Error Type

If an observation value is calculated differently than expected due to a human or instrument mistake or as a result of a measurement error, these types of outliers are defined as first type or human-induced errors. Such errors cause shock changes on parameter estimation values [40]. It occurs when normal data is incorrectly entered into the system by individuals or in similar cases. Also, parameter bias is sharper and greater in these types of errors. Effects on such observations must definitely be eliminated. They are defined as Additive Outlier (AO) or A-type outlier models in the literature. Such errors can be detected more accurately through data scanning processes. They were first introduced and modeled by Fox in 1972 [15].

Human-induced outlier model is defined as

$$y_t = z_t + \delta x_t \tag{2}$$

where y_t is the observed value, δ is the size of outlier, x_t is variable valued as 1 at the time of outlier (T = t) and 0 in other times [13].

3.5. Data-Induced Error Type

The Data-induced model affects subsequent observations starting from its emerging position. If the outlier has a persistent or permanent effect on the level and variance process of the series, it is called an innovational outlier. This refers to an illness (anomaly) state with decreasing effect. Such effect progressively decreases. It was first introduced by Fox in 1972 and modeled as the Innovational Outlier (IO) model. It is also defined as a B-type model in the literature [15].

The data-induced error model is expressed as

$$y_T = \frac{\theta(B)}{\phi(B)}(e_T - \delta x_T) \tag{3}$$

where θ(B) is the *MA* function, ϕ(B) is the *AR* function, e_t is the residual at the time, x_t is the variable with a value of 1 at T = t and 0 at other times [13].

$$\rho^2 = (1 + \pi_1^2 + \pi_2^2 + ... + \pi_{n-T}^2) \tag{4}$$

$$\varpi_I = e_T \tag{5}$$

$$\varpi_A = \rho^2 \pi(F) e_T \, , \tag{6}$$

$$\pi(F) = (1 - \pi_1 F - \pi_2 F^2 - ... - \pi_{n-T} F^{n-T} \tag{7}$$

$$\lambda_{1.T} = \varpi_I / Var(\varpi_I) \tag{8}$$

$$\lambda_{1.T} = \varpi_A / \rho^2 Var(\varpi_A) \tag{9}$$

In error detection processes, C values are taken as C=3.00 for high-sensitivity detection, as C=3.50 for mid-sensitivity detection, and as C=4.00 for low-sensitivity detection.

3.6. Model-Induced Error Type

Another risk encountered in information systems is the selection of the wrong model equation for the data. In such cases, totally normal observations can become outliers as a result of selecting the wrong model [1]. Furthermore, models established without eliminating the potential error effects on data may be erroneously selected. It is necessary to observe the tendencies of data on the scatter diagram in order to avoid such outliers. Thus, it will be possible to have preliminary information on appropriate models for the data.

3.7. Error Detection Algorithm

Error detection algorithm consists of the following steps:

- Read observations from the defined file.
- Read ARMA parameters obtained by using the program package.
- Calculate π_j 's from the estimated model.
- Use $\hat{\sigma}_a^2$ and obtain $\hat{e}_t$ ’s and find outliers.
- Read critical values C which can be 3.00, 3.50, and 4.00.
- Do;
    1. Calculate $\hat{\sigma}_a^2$ from the $\hat{e}_t$;
    2. Increase the current value by one;
    3. Calculate the $\lambda_{1.T}$, $\lambda_{2.T}$ which define the data effects;
    4. If $\lambda_{1.T} > C$, display outlier position and IO;
    5. If $\lambda_{2.T} > C$, display outlier position and AO.
    6. Otherwise, there is no outlier. Stop.
    7. Calculate the effect of IO and AO and update these effects on observations.

8. Calculate new $\hat{e}_T$'s for updated observations,

- End Do.

Read updated new observations and perform new algorithms again [43].

3.8. Outlier Detection Flowchart

An algorithm flowchart demonstrating the error detection process on datasets is presented in Figure 1. The software for this flowchart is a method called iterative procedure determined to be the most effective method among outlier detection processes. This method is converted into a module with C# programming language and the error outlier detection process is performed [12].

In this method, observational data is read into the statistical program package (MINITAB) and the time series model parameters and residuals for each observation (et) are obtained. Following this transaction, a software program that conducts error detection on data is used and the detection process is initiated. The detection process continues iteratively until no errors are present in the series. When all erroneous observations on the series are eliminated, the software operation is also concluded. As a result of this process, variance and parameter values change and the model is indirectly optimized.

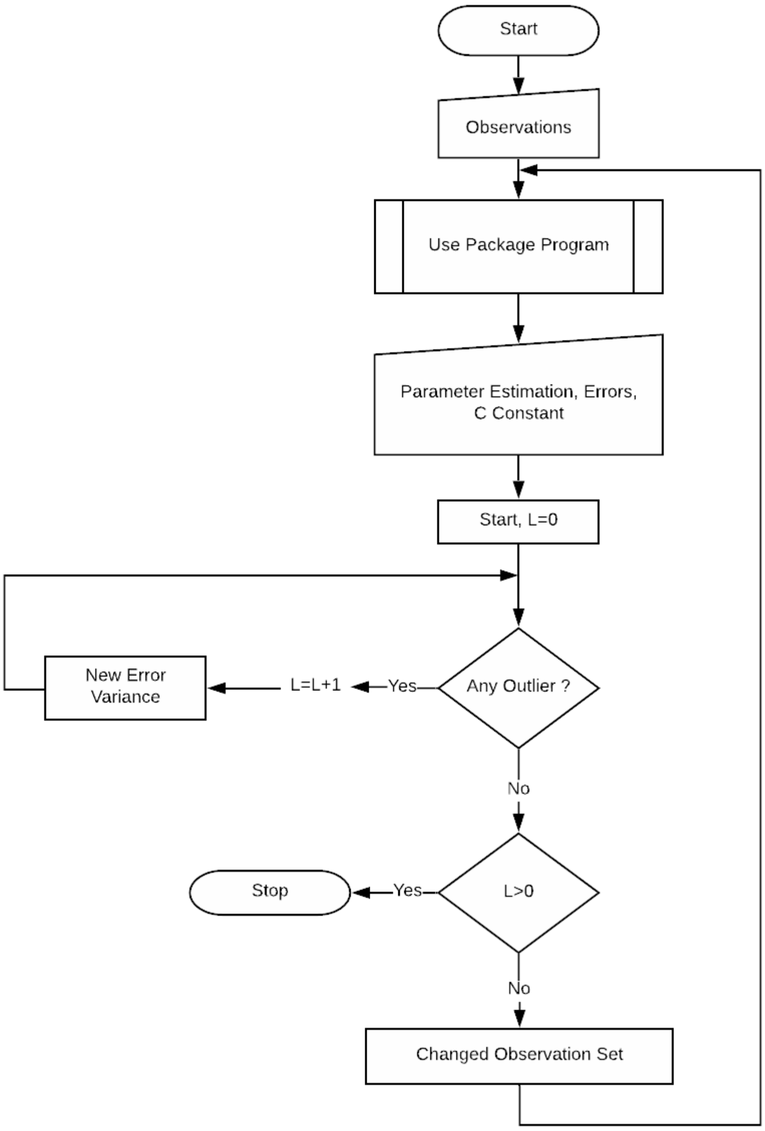

Figure 1. Outlier Detection Algorithm Flowchart

4. Conclusion

By electronic records in hospitals, the most effective and efficient healthcare services are given to the patients within the shortest time. Additionally, hospital staff have less workload and be less likely to make mistakes [44]. On the other hand, there are always risks during the application of information systems. Risks that enable access from abstract datasets to significant information systems are usually data-driven risks. Especially during the COVID-19 pandemic process, the risks that may arise due to the mistakes of health personnel have become very important. Such risks can be eliminated. Thus, a control process on data sets is certainly required. Through these processes, it is possible to attain more minimal variance values, more objective parameter estimations, and more effective models. It will thus be possible to have more effective and accurate datasets by using the results obtained from these models. In addition, even though there are human-induced risks on data, these data can be corrected and rendered more suitable to acquire information. Human-driven or data-driven risks on data may also result in wrong model parameters. Such aberration may be so high that the model format may sometimes entirely change. In this respect, the natural result of human or data-induced errors in data is the selection of wrong models, which appears as model risk. In line with these inferences, it is considered that the following results and suggestions are noteworthy:

- Data-induced errors are natural. The effects of these data are also natural. For example, it occurs when a patient contracts a virus without realizing and some of his/her values are different than expected. Effects on data continue in a decreasing manner until the patient clears the virus from his/her body. According to the scientific literature, it is the researcher's choice to eliminate or not eliminate these effects, because probable natural aberrations in models are inherent in modeling.
- Human-driven risks are easier to detect in data. Such data is called "contaminated", which is "infected", in the literature. They occur when a data entry operator enters the data erroneously while transferring them into the system. Such effects must definitely be estimated and separated from data. Moreover, these effects cause a shock on variance and parameter estimations. In recent studies, outlier detection studies were mostly based on human-induced outliers
- The effect of data-induced risks on the model is normal, while the effect of human-induced risks on the model is abnormal and has to be eliminated.
- Prior to data modeling, data must be observed using graphs called scatter diagrams and an appropriate model must be investigated on graphs. It will thus be possible to obtain healthier results in terms of parameter estimations.
- These kinds of error debugging analyses optimize the parameters and make the models stronger and more dynamic.
- In order to receive the expected benefits from information systems, it must be ensured that data are healthy as it is, so to speak, the nutrients of such systems. It must be remembered that however perfectly a system operates, if it is fed erroneous data, then the results attained will be far from becoming information.
- Informed and qualified personnel must be employed for the achievement of all these goals and the improvement and control of information systems. It is a prerequisite, even though costly, for organizations to achieve a competitive power.

Secure hospital information systems cost high. However, this cost is only for a short period of time. In the long term, their benefits outweigh the costs for organizations.

In the present study, a method based on the statistical time series approach and Box-Jenkins model was implemented to detect the errors in data used in the detection of Covid-19 symptoms. Outlier data was detected with an iterative error detection process which has been developed with C# programming language and is also an information system. An improvement in mean squares of error and a reduction in variance were observed as a result of replacing outliers with estimated data by using the method. These results depict that the proposed method is successful in detecting the errors in time series data and replacing such data with estimated data. It was determined that problems induced by data outliers can be reduced by applying the proposed method to health information systems such as the detection system of Covid-19 symptoms. In future studies, if there is sufficient data on Covid-19 tests, the method here can be applied to detect and correct outliers.

Conflict of interest:

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

The Declaration of Ethics Committee Approval

The author declares that this document does not require an ethics committee approval or any special permission. Our study does not cause any harm to the environment.

Authors' Contributions:

A. K.: Conceptualization, Methodology, Formal analysis, Original draft preparation.
R. G.: Resources, Implementation, Investigation.
Ö. A.: Resources, Implementation, Writing -Original draft preparation, Journal Correspondence.
All authors read and approved the final manuscript.

References

[1] Akouemo, H., Povinelli, R., "Data Improving in Time Series Using ARX and ANN Models". *IEEE PES Transaction on Power Systems*, 32, 3352-3359, 2015.

[2] Ergun, Ü., "Modern Management Accounting Applications by Information Technologies" *Dokuz Eylul University Journal of Faculty of Economics and Administrative Sciences*, 11, 1-17, 1995.

[3] Soyuer, H., İşletmelerde Bilgisayar Destekli Bilgi Sistemi Uygulamaları ve Üretim/İşlemler Yönetiminde Bilgisayara Dayalı Sistemler, Ph. D. thesis, Gazi University Social Sciences Institute, Ankara, TR, 2000

[4] O'Brein, J.A., Introduction to Information Systems. Irwin McGraw Hill, Boston, 1997

[5] Earl, M.J., "Experiences in Strategic Information System Planning", *MIS Quarterly*, 17(1), 1-12, 1993.

[6] Kalıpsız, O., Buharalı, A., Biricik, G., Sistem Analizi ve Tasarımı. [System Analysis and Design], Papatya Yayıncılık Eğitim, İstanbul, TR, 2011

[7] Lucas, H., Information System Concept for Management (5th Edition), McGraw-Hill, New York, NY, 1994

[8] Peppard, J., IT strategy for Business, Pitman Publishing, New York, NY, 1993

[9] Katsikas, S.K., "Health Care Management and Information Systems Security: awareness, training, or education?", *International Journal of Medical Informatics*, 60, 129-135, 2000.

[10] Gritsalis, D.A., "Enhancing Security and Improving Interoperability in Healthcare Information Systems", *Medical Informatics*, 23(4), 309-324, 1998.

[11] Ward, M.J., Self, W.H., Froehle, C.M., "Effects of Common Data Errors in Electronic Health Records on Emergency Department Operational Performance Metrics: A Monte Carlo

Simulation", *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine*, 22(9), 1085-1092, 2015.

[12] Kaya, A., "Modelling for Detection Processes Outliers in Time Series and Its Performance Analysis", *Uludag University Journal of Faculty of Economics and Administrative Sciences*, 22 (1), 271-279, 2003.

[13] Chang, I., Tiao, G.C., Chen, C., "Estimation of Time Series Parameters in the Presence of Outliers", *Technometrics*, 30(2),193-204, 1988. Doi:10.1080/00401706.1988.10488367

[14] Andrews, D., Pregibon, D., "Finding the Outliers that Matter. Journal of the Royal Statistical Society", *Series B (Methodological)*, 40(1), 85-93, 1978.

[15] Fox, A.J., "Outliers in Time Series", *Journal of the Royal Statistical Society Series B*, 34, 350-363, 1972.

[16] Hillmer. S.C., "Monitoring and Adjusting Forecasts in the Presence of Additive Outliers*",* *Journal of Forecasting*, 3, 205-221, 1984.

[17] Tsay, R.S., "Time Series Model Specification in the Presence of Outliers", *Journal of the American Statistical Association*, 81(393), 132-141, 1986.

[18] Pena, D., Measuring the importance of outliers in ARIMA models. New Perspectives in Theoretical and Applied Statistics John Wiley, New York, USA, 1987

[19] Abraham, B., Yatawara, N. A., "Score Test for Detection of Time Series Outliers", *Journal of Time Series Analysis*, 9(2), 109-119, 1988.

[20] Bruce, A.G., Martin, D., "Leave-k-out diagnostics for time series (with discussion)", *Journal of the Royal Statistical Society Series B*, 51, 363-424, 1989.

[21] Abraham. B., Chuang, A., "Outlier Detection and Time Series Modelling", *Technometrics*, 31(2), 241-248, 1989.

[22] Box, G.E.P., Tiao, G.C., "Intervention Analysis with Applications to Economic and Environmental Problems", *Journal of American Statistical Association*, 70(349), 70-79, 1975.

[23] Denby. L., Martin, R.D., "Robust estimation of the first order autoregressive parameter", *The Journal of the American Statistical Association*, 74, 140-146, 1979.

[24] Maronna, R., Martin, R., Yohai, V., Robust Statistics: Theory and Methods, Wiley, New York, USA, 2006.

[25] Firmino, P.R.A., Mattos Neto, P.S.G., "Ferreira TAE. Correcting and combining time series forecasters", *Neural Networks*, 50, 1-11, 2014.

[26] Paulino, J., Gomes, C., Gonçalves Júnior, J., Rodrigues, M., Souza, A., Pimentel, J., Brito, K., Saboia, S., Firmino, P., "Predictive Models and Health Sciences: A Brief Analysis", International Archives of Medicine. 2017. Doi: 10.3823/2487.

[27] Chen, Z., Chen, Y., Li, T., "Port cargo throughput forecasting based on combination model", Proceedings of in Joint International Information Technology. Mechanical and Electronic Engineering Conference (JIMEC 2016). Xi'an, China, 2016,148-154.

[28] Adedia, D., Nanga, S., Appiah, S.K., Lotsi, A., Abaye, D.A., "Box-Jenkins' Methodology in Predicting Maternal Mortality Records from a Public Health Facility in Ghana*", Open Journal of Applied Sciences*, 8, 189-202, 2018. Doi:10.4236/ojapps.2018.86016

[29] Langat, A., Orwa, G., Koima, J., "Cancer Cases in Kenya; Forecasting Incidents Using Box &Jenkins Arima Model", *Biomedical Statistics and Informatics*, 2, 37-48, 2017.

[30] Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F.M., Preen, D.B., Stewart, L.M., Fatovich, D.M., "A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia", *Journal of Biomedical Informatics*, 57, 62-73, 2015. Doi: 10.1016/j.jbi.2015.06.022

[31] Luz, P.M., Mendes, B.V.M., Codeço, C.T., Struchiner, C.J., Galvani, A. P., "Time Series Analysis of Dengue Incidence in Rio de Janeiro, Brazil", *The American Journal of Tropical Medicine and Hygiene*, 79(6), 933-939, 2008. Doi:10.4269/ajtmh.2008.79.933

[32] Aydın, Ö., Karaarslan, E., "Covid-19 Belirtilerinin Tespiti İçin Dijital İkiz Tabanlı Bir Sağlık Bilgi Sistemi". Online International Conference of COVID-19 (CONCOVID), İstanbul, Turkey,2020, pp.8-9.

[33] Usman, M., Wajid, M., Zubair, M., Ahmed, A., "On the possibility of using Speech to detect COVID-19 symptoms: An overview and proof of concept", Researchgate, 2020. Doi:10.13140/RG.2.2.31718.57923

[34] Munsch, N., Martin, A., Gruarin, S., Nateqi, J., Abdarahmane, I., Weingartner-Ortner, R., Knapp, B., "A benchmark of online COVID-19 symptom checkers", *medRxiv* 2020.05.22.20109777, 2020. Doi: 10.1101/2020.05.22.20109777

[35] Mackey, T. K., Purushothaman, V. L., Li, J., Shah, N., Nali, M., Bardier, C., Liang, B., Cai, M., Cuomo, R., "Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated with COVID-19 on Twitter: Retrospective Big Data Infoveillance Study", *JMIR Public Health and Surveillance*, 6(2), e19509, 2020. Doi:10.2196/19509.

[36] Zens, M., Brammertz, A., Herpich, J., Suedkamp, N. P., Hinterseer, M., "App-based tracking of self-reported COVID-19 symptoms (Preprint)". ResearchGate, 2020. Doi:10.2196/preprints.21956

[37] Louden, K.C., Laudon, J.P., Management Information Systems: managing the digital firm (8th. edition), Prentice-Hall, New Jersey, USA, 2004.

[38] Kini, R.B., "Strategic Information Systems", *Information Systems Management*, 10(4), 42-50, 1993.

[39] Ammenwerth, E., Graber, S., Herrmann, G., Bürkle, T., König, J., "Evaluation of Health Information Systems-Problems and Challenges". *International Journal of Medical Informatics*, 71, 125-135, 2003.

[40] Ljung, G.M., "On Outlier Detection in Time Series". *Journal of the Royal Statistical Society: Series B*, 55, 559-567, 1993.

[41] Kaya, A., "A Type Outlier in AR (1) Model", *The Journal of Statisticians*, 3, 1-7, 2010.

[42] Box, G.E.P., Jenkins, G.M., Time series analysis: Forecasting and control, Holden-Day, San Francisco, USA, 1976.

[43] Kaya, A., “Outlier Effects on Databases, LNCS 3261, Advancing Information Systems.” Proceedings of Third International Conference, ADVIS 2004, İzmir, Turkey, Springer, 2004. p. 88-96.

[44] Kılıç, T., “Digital Hospital; An Example of Best Practice”, *International Journal of Health Services Research and Policy*, 1(2), 52-58, 2016. Doi:10.23884/ijhsrp.2016.1.2.04