DOI: https://doi.org/10.53486/cike2023.44

UDC: 004.6:61

THE IMPACT OF DATA PRE-PROCESSING ON THE ASSESSMENT OF THE SIMILARITY OF TREND FUNCTIONS

ILIE COANDA
PhD, Associate Professor
Department of Information Technology and Information Management
Academy of Economic Studies of Moldova
Chisinau, Republic of Moldova
Email account: ildirosv1@gmail.com
ORCID ID: 0000-0002-0010-1202

Abstract. An approach to the way, the technologies of cleaning, completing, smoothing of large volumes of data to be subjected to analysis is proposed. As a rule, depending on the field and the method of data collection / recording on various supports, they could be classified at least in two categories: precise data (recorded by automated techniques, without any influence of the human factor) and data, with a level of approximation (when collecting / recording, to some extent, at a certain stage of the activity, the "man" (human) participates). If, in the case of the same activity, relatively, many people participate, then, and the quality level of the records will be at a different level of precision than the records performed in an automated way. This work aims to highlight the importance / impact of the influence of the quality of the preliminary processing (smoothing, cleaning, etc.) of the primary data used in the analysis process. In case studies, the object of the research is considered to be a set of time series corresponding to data collected regarding the phenomenon of the spread of an epidemic. The data recording of such a phenomenon fits perfectly in the studied case when the data collection is carried out with the intense participation of the "human", who is characterized by frequent deviations from the regulations prescribed by the situation. Consequently, some data could be fixed with a delay or / and people affected by the disease signal the doctor in a different period of time. Such phenomena can create anomalies in the data structure. In order to highlight the impact of the application of different smoothing methods, the completion of the primary data, the approximating functions for each time series were obtained, having previously been "corrected" by: a) averaging the neighboring data; b) "suspicious" data were excluded. As a result, two sets of approximating functions are obtained (approximating functions can be obtained by involving non-linear regressions). By applying the technologies for evaluating the similarity of the functions, the distance (similarity level) between the functions of each set of approximating functions is calculated. Next, the hierarchical clusters of the sets of approximating functions (two sets of approximating functions) can be obtained. By comparing the hierarchical clusters, the level of impact of the "correction" methodology approach a) and b) can be evaluated.

Keywords: cleaning, smoothing, impact, similarity, functions, regression

JEL CLASSIFICATION: C63, I21, I23, I25, I29

Introduction

There are many phenomena in society and nature on the basis of which data can be collected, forming data series, including time series (chronological, temporal). Such data can be used in various ways and purposes (medicine, economics, technical technologies, automated processes, etc.). In particular, an interest could be: predicting the near future, or following some phenomena over time; the evolution of the values of the parameters kept in the research; the classification elements by comparing them according to defined metrics. Thus, time series data is becoming more ubiquitous and important as the data ecosystem expands. Sensors and tracking mechanisms are everywhere, and as a result, unprecedented amounts of data of various quality levels in time series format are available for

research for various purposes. Time series are particularly interesting because they can address questions of causality, trends, and the likelihood of future outcomes. Time series data and its analysis are increasingly important due to the massive production of such data through, for example, information on the Internet, the digitization of healthcare and the growth of smart cities. In the immediate future, we can expect a rapid increase in the level of data quality, quantity, and importance of time series data. As continuous monitoring activities and data collection become more common, it is only natural that there will also be an increase in the need for competent time series analysis, both with statistical and machine learning techniques. We will therefore use various time series processing techniques useful for analyzing and predicting human behavior, scientific phenomena and private sector information, as all these fields provide a rich set of time series data. Time series analysis is the effort to extract meaningful summary and statistical information from points arranged in chronological order. This is done in order to research past behavior as well as to predict future behavior. Thus a variety of approaches are used, ranging from hundred-year-old statistical models to newly developed neural network architectures. Innovations in time series analysis result from new ways of collecting, recording and visualizing data.

Data quality in time series

Above, the issue of data quality was tangentially touched upon, a very important property in the process of exploration, research, extracting information as truthful as possible, to involve in decision-making processes. Whatever the activities that produce the time series, the procedures, the techniques, the ways of data collection, it is logical to assume inaccuracies, deviations from the prescribed standards. Therefore, the use of these data requires a preliminary analysis, in order to prepare them, to the extent necessary, to be involved in decision-making. As with any data analysis task, proper data cleaning and processing is often the most important step of a time stamping process. Fancy techniques can't fix messy data. Most data analysts will need to find techniques for aligning, cleaning and/or smoothing the data involved in the study. As the data is being prepared, a variety of tasks will need to be applied in order to pre-process the information.

Among the most obvious common problems in chronological data sets can be mentioned: missing data; Changing the frequency of a time series (i.e. oversampling and downsampling; smoothing data; dealing with seasonality in data). In the following, only ways of handling missing data and smoothing data will be addressed.

Missing data is surprisingly common. For example, in healthcare, missing data from medical time series can have different causes: the patient did not comply with a desired measurement; The patient's health statistics were in good condition, so there was no need to take a specific measurement; the patient was forgotten or undertreated; a technical component had an unforeseen failure; there were errors in data recording. In order to complete, eliminate this deficiency, various interpolation techniques or the exclusion of these periods from the time series may be involved.

Data smoothing can be done for a number of reasons, and given that real-world time series data is typically smoothed before analysis, especially for visualizations that aim to understand something about the data under investigation. It is necessary to argue why smoothing is necessary, and what would be the most appropriate method of smoothing time series, smoothing can serve several purposes. While outlier detection is a topic in itself, if there are compelling reasons that the data should be smoothed, a so-called "moving average" can be used to remove measurement peaks. Even if the peaks are accurate, they may not reflect the underlying process and may be more of a data

collection-recording problem;

Time series pre-processing model

Figure.1 shows the time series (without any changes) over a period of several days. The data were extracted by the author from the Internet (open sources at that time), (from day to day) during the spread of the pandemic in the fall of 2020 year(data relating to two concrete territorial administrative units: Loc1, Loc2). An analysis, even a superficial one, convinces us of the existence of a periodic law (one week long).

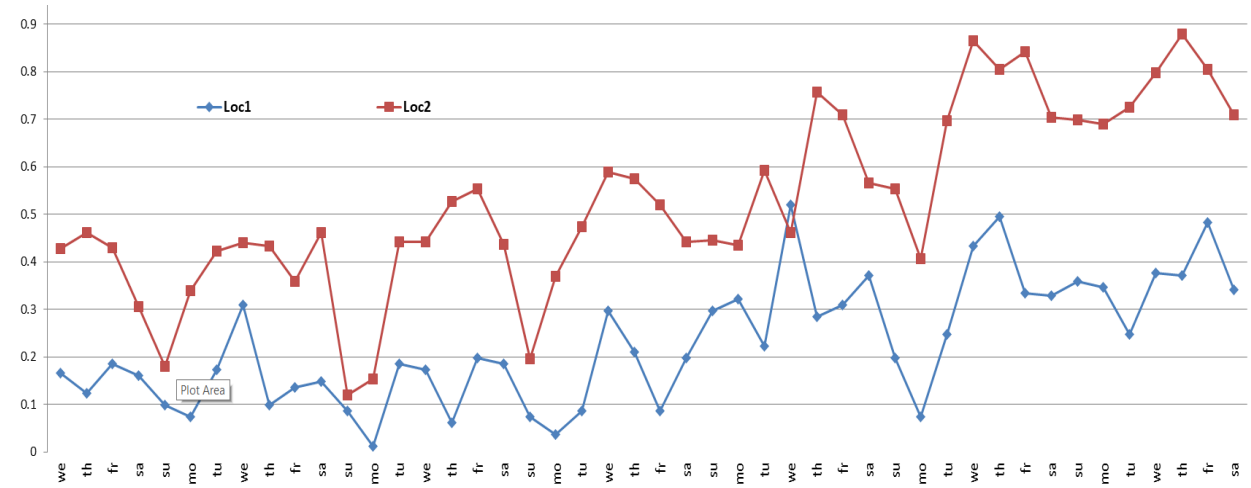

Figure 1 A real example of primary data Source:

author's own study

Figure 2 shows the data extracted from the data set shown in Figure 1, for two different periods of 7 (seven) days each (Loc2). A simple visual analysis of the values can easily guide us to the periods (values) of the corresponding days.

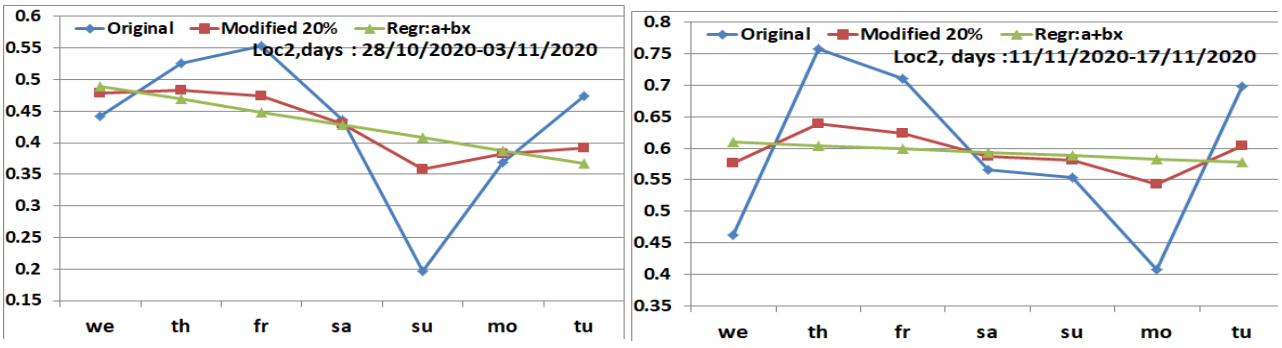

Figure 2 Data extracted from the Time Series (showed in Fig.1 in two periods of days)

Source: author's own study

In both images in Figure 2, the "Original" line represents the time series before the modification, "Modified 20%" - after the modification, and "Regr:a+bx" - the linear regression. The number 20% represents the level of deviation from the regression-line values.

The essence of the model: given the fact that the spread of the pandemic is a continuous process, it means that the intensity of the diseases does not depend on the days of the week in the assumption that on all days of the week the communication between people was about the same intensity. Therefore, values those are out of the ordinary, or too close, or too high, are reasonable to be considered as a defect of the data recording process, or of late detections of the disease, or other causes. In this context, it is proposed to add up all the values that exceed the regression-line by more than (in the given case) 20%. The summary value obtained corresponds to the "number" of erroneously registered "diseases". The "number" of "diseases" obtained **is to be distributed direct proportionally ("moving distributed direct proportionally ")**in all the days of the short period (for example, a week)(values further from the average will undergo more substantial changes).

Figure 3 shows the approximating functions for the primary time series corresponding to the localities and Loc1(Anen) Loc2(Chiș) as well as the respective functions for the modified series Loc1(AnenM) Loc2(ChișM) according to the approaches described above. In the image on the left (fig.3) the non-linear regression is calculated and built over an interval of 21 days (3 weeks), and in the one on the right the non-linear regression is built only for the values of the first week. The length of the interval was chosen from the concrete considerations for the phenomenon of infection, that an infected person could not be registered with a delay of more than 5 (five) – 7 (seven) days.

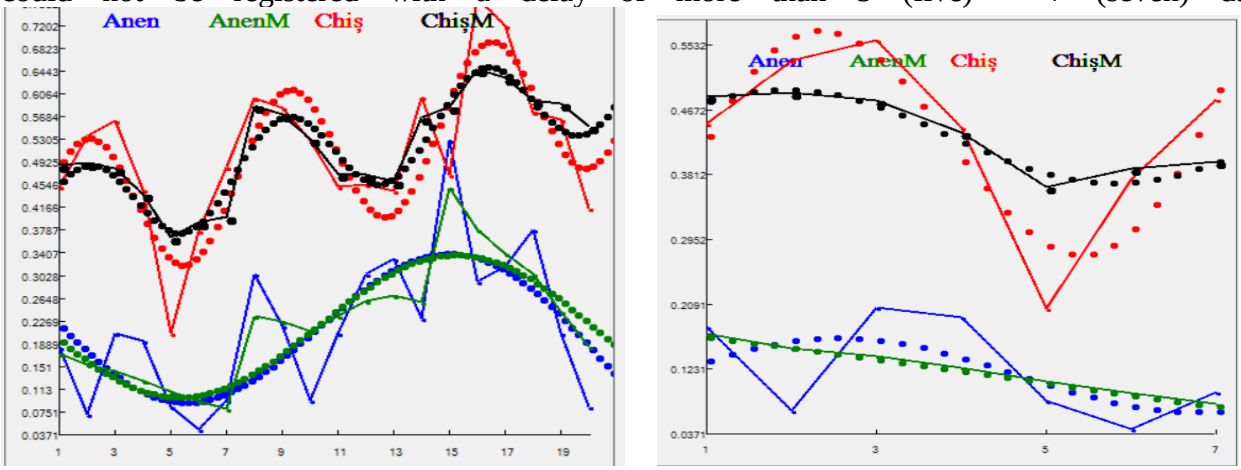

Figure 3 Approximation functions (showed in Fig.1 in two periods of days)

Source: author's own study

An analysis of the behavior of the approximating functions for AnenM and ChișM demonstrates the existence of the impact of neglecting the need for a preliminary analysis of the time series in order to annihilate the significant influence of some foreign values to the context of the studied phenomenon, without applying an adequate process, taking into account the specifics of the phenomenon, of removing non-compliant data values. In (COANDĂ, Ilie, 2022) the essence of the notion of "approximating functions" was discussed, as well as the definition of a technique / tools / model for evaluating the level of similarity between two functions. According to these techniques, the distances between the approximating functions for "Anen", "AnenM", "Chiș", "ChișM" corresponds (see the image on the left of Figure 3) to: d12=0.04; d13=0.95; d14=0.98; d23=0.96; d24=0.94: d34=0.02. (d34 =0.02 represents the "distance" between "Chiș" and ChișM", and d13 0.95- between "Anen" and ChișM". The values of the level of similarity between the approximating functions open up many possibilities for applying algorithms in the field ML, AI.

Conclusions

The proportional distribution methodology ("**moving distributed direct proportionally"**) presented in this paper represents an essential argument in favor of a differentiated approach in the preliminary data processing process. In the case of the presented example, a simple cutting of the vertices, or an averaging of neighboring data, can substantially amplify the veracity of the data. Cutting the peaks would mean decreasing or increasing the values in the numerical time series, which will certainly lead us to erroneous predictive conclusions.

REFERENCES

1. COANDA, Ilie. Evaluation of similarity of trend functions. In: Competitiveness and Innovation in the Knowledge Economy [online]: 26th International Scientific Conference: Conference Proceeding, September 23-24, 2022. Chișinău: ASEM, 2022, pp. 309-312. ISBN 978-9975-3590-6-1 (PDF). https://irek.ase.md/xmlui/handle/123456789/2607