Annual International Scientific Conference
"Competitiveness and Innovation in the Knowledge Economy"
September 26-27, 2025
Chisinau, Republic of Moldova

DOI: https://doi.org/10.24818/cike2025.61

THE IMPACT FACTORS ON THE ASSESSMENT OF SIMILARITY BETWEEN FUNCTIONS

ILIE COANDA
Department of Information Technology and Information Management
Academy of Economic Studies of Moldova
Chisinau, Republic of Moldova
e-mail: coanda.ilie@ase.md
ORCID ID: 0000-0002-0010-1202

Abstract. Data processing, in any field, especially in the case of accessibility of relatively large volumes of data, becomes appropriate to involve more specific techniques, procedures, which at the initial stage could be considered as universal. An important factor in the development of algorithms for assessing the level of similarity between functions, in certain situations, depending on the nature of the phenomenon under research, may be the size of the variation interval of the independent variable. In this context, in this paper, certain suggestions, techniques for obtaining numerical characteristics, obtained based on methodologies for varying the lengths subintervals of the integral definition interval of approximating functions, deduced from the data set involved in the research, will be discussed. One of the main suggestions could be the division of the entire interval into several subintervals. The number of subintervals is supposed to be deduced depending on the nature of the phenomenon under investigation, thus assessing the level of similarity in each subinterval, and then building a synthesis algorithm for the integral interval. Such an algorithm - methodology - is to be presented in this paper, by presenting examples - case studies based on primary data similar to some real data. Another question that may arise is the nature of a possible real factor that could have a significant impact on the results of the similarity assessment. The essence of such factors may be very difficult to deduce from only a single data set. A solution would be to highlight the nature of several data sets related to the "circumstances" of data production in the research process, as well as to their collection methodologies.
Keywords: similarity, evaluation, intervals, subintervals, algorithm, functions.
JEL Classification: C63, I21, I23, I25, I29

1 Introduction

Processing large data sets can be of different complexity from one data collection to another. In this context, a simple approach to processing methodologies, without taking into account other circumstances, circumstances, which at first glance, could not fall into the researcher's attention/horizon, then become factors of significant influence on the argumentation (substantiation) process when making final decisions at a high level of correctness. Data collections subject to the analysis process should be viewed in the widest and most diverse context of the circumstances and the data collection process. Any set (collection) of data is an image of phenomena that can reproduce the respective essence, phenomena, which could be as simple, trivial, and quite complex according to their essence. Another situation may be important, and which is worth taking under control, whether or not there could be a human factor in the collection (registration) of data. The data collection in view, it is necessary to subject it to a cleaning, smoothing procedure as explained in the material (COANDĂ, Ilie, 2022), then to the application of technologies for transition from the discrete form of the data to a continuous form (function). In (COANDĂ, Ilie, 2023) such a methodology is described, thus obtaining the continuous form for the presentation of the data, (approximate functions, which, further, are to be subjected to analysis procedures for the appropriate purposes. It is necessary to specify that the approximate functions represent the image of the essence of the phenomena with a certain margin of error as a result of the application, in the given case, of the procedure characteristic of the regression technique.

2. Applying the procedure for dividing the definition interval of approximating functions into subintervals

Some particularities of the concrete data set on which the data analysis techniques on subintervals obtained from the integral definition interval of the approximating functions will be applied (the division into subintervals has become suggestive in order to increase the level of precision in obtaining the approximating functions:

1. The data refers to the number of people infected with the Corona virus during the period of days 30.09.20 – 19.04.21.

2. The data are public, for each day of the indicated period, the respective data were taken from the Internet, for 40 geographical regions that constitute an entire compact administrative-territorial region (they were taken without being preprocessed).

3. A preliminary analysis of the data revealed the need to apply a certain preprocessing procedure described in (COANDĂ, Ilie, 2023).

4. In (COANDĂ, Ilie, 2024) a numerical evaluation methodology of similarity between two functions, which can be recommended for the classification or clustering algorithms.

Applying the respective techniques for obtaining approximating functions, as well as calculating the level of similarity between the respective functions, it emerged the need to study the problem on subintervals of the entire interval indicated above.

Being aware of external information regarding the incubation period of the virus, which in this case is not greater than 7 (seven) days, (information taken from outside the data collection) it was proposed that the length of a subinterval be greater than the incubation period of the virus (an essential constraint on the length of the subinterval).

Another particularity of the data set, which was involved in the process of selecting the named regions for research, was their proximity. We again have external information, which, most likely, is not attached to the data set, and must be taken from outside. The fact that a similarity is attested (similarity) strong, does not mean at all that these regions are neighboring. Conversely, based on an obvious similarity, it could be concluded that these regions could be neighboring (i.e. intensive social communication is observed, a favorable environment for the spread of the virus).

A preliminary analysis of the data demonstrates the existence of two periods of days, within 30.09.20 - 19.04.21, in which an increase is attested, immediately followed by a significant decrease in the intensity of infection. In this case, it is the interval 20.10.20 - 20.11.20. The second period refers to the interval 10.02.21 - 20.03.21.

Certainly, there must be a explanation. And again, it is necessary to turn to other information that would enhance the level of communication in society. This information, again, most likely was not attached to the data set, so this is external information.

3 Some Case Studies

In the context of the above, the following is a presentation of some results with the respective comments regarding the application of the methodology of dividing the integral interval into several subintervals. Thus, it was first possible to detect some subintervals with substantially different behavior from the others, and then to carry out a broader analysis (study) involving some exogenous information. Such information could be difficult / imposible to extract only from the dataset under research.

Figure 1 presents the intensity functions (number of infected per 1000 people) for certain 4 (four) administrative regions (1 Anen, 2 Chiș, 3 Comr, 4 Vulc) in a certain country.

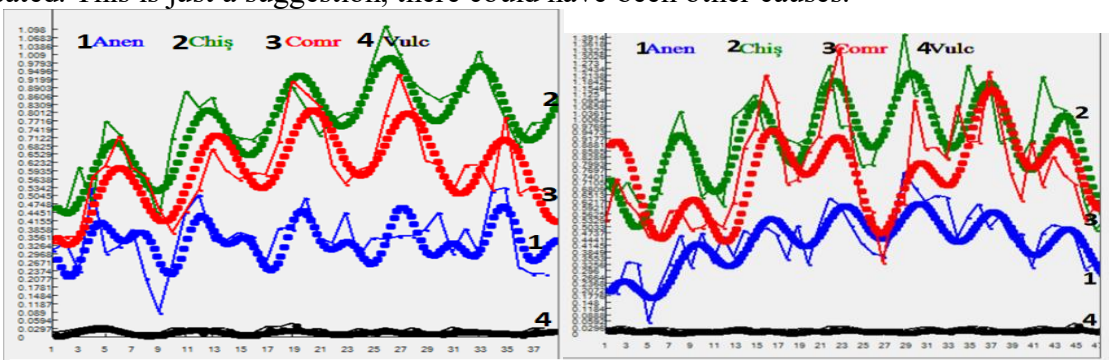
If the oscillatory form of the functions could be quite simple to fix and quite well argued based only on the information presented in the dataset (days of the week), while an adequate explanation, simple to formulate, of a possible cause of the appearance of a vaulted shape, in the period 08.11.20 - 15.12.20 (Figure 1 a)) and in the period 11.02.21 - 30.03.21 (Figure 1 b)), may

be quite difficult to highlight, relying only on the accessible information in the dataset available only in table format.

If an annex is attached to the data set, it is specified that the period 08.11.20 - 15.12.20 includes a good part of the electoral campaign period, and in the period 11.02.21 - 30.03.21 the next semester of studies begins, then things become particularly clear. During these periods, a clearly more intensive intensity of direct contact between individuals is assumed, which essentially favors the more pronounced spread of the virus.

Another question, what would be the reason that on days off, the intensity of the spread of the virus decreases, sometimes to zero, even if it is hard to imagine that on these days the level of direct contact between people decreases dramatically (which is absurd). And if there had also been an annex to the data set with the activity graph of medical enterprises, again, things would have been much clearer. The information presented graphically in Figure 1 a) can also suggest at least one thing: the electoral agitation (i.e. the meetings) in regions 2 Chiș and 3 Comr were qualitatively, clearly similar, with different intensities, which cannot be confirmed for two other regions (1Anen, 4Vulc).

In Figure 2 b), from a qualitative point of view, the suggestion can be extracted that the beginning of the second semester in educational institutions significantly contributed to the intensification of the spread of the virus, because environments favorable to the spread of the virus were created. This is just a suggestion, there could have been other causes.



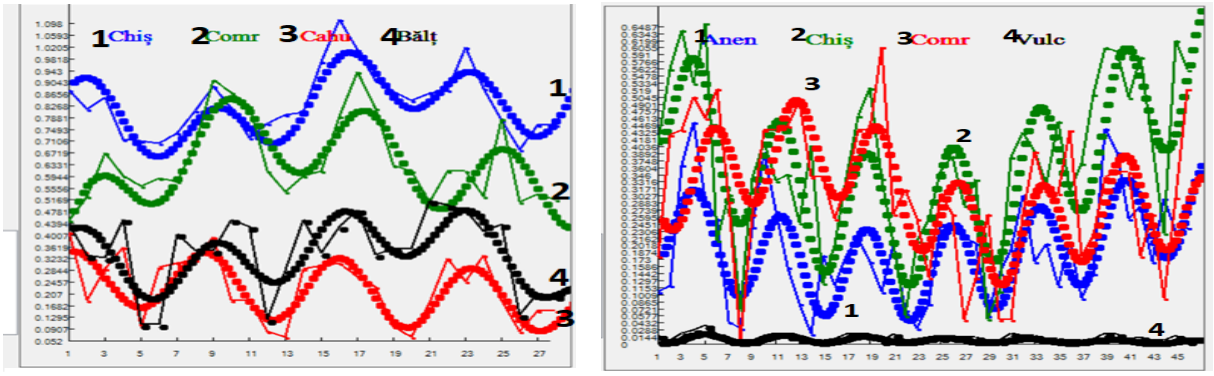
a) 08.11.20 - 15.12.20 b) 11.02.21 – 30.03.21

Figure 1. Approximation's, the same regions, different periods of days:

Source: author's own study

In order to highlight the similar nature of the infection for the most populated centers of social activity (1Chiș,2Comr,3Cahu,4Balț), Figure 2 a) presents representing the intensity of illnesses in the period 08.11.20 - 15.12.20. The graphs of these functions suggest a relatively high similarity, which may argue the suggestion that in these compact regions of social activities, the intensity of direct communications was similar, which demonstrate that, among other things, the electoral agitation was also similar. We specify that the population of the 4Balți region is about 125,000 inhabitants, and that of the 2Comr region is 67,000 inhabitants. Even though the population in the 2Comr region is practically 2 (two) times smaller than that of the 4Balț region, the intensity of infections is almost inversely proportional, which would suggest that the political interest in the elections is more pronounced in the 2Comr region. At the same time, the population of the 3Cahu region is approximately 72,000 inhabitants, very close to that of the 2Comr region, the disease phenomenon is almost twice as weakly registered. This may lead us to believe that the level of interest in the elections is lower, respectively. In Figure 2 b) the behavior of the disease phenomenon is graphically presented in a period of days between the two intervals of days 27.12.20 – 10.02.21, in which the approximating functions begin with a significant increase, then continue with a similar decrease (for both intervals 08.11.20 - 15.12.20 and 11.02.21 – 30.03.21). The image of the functions denotes the fact that the intensity of the disease decreases then

increases. The cause could be due to holidays vacations, which significantly reduce direct contact between people.



a) 08.11.20 - 15.12.20 b) 27.12.20 – 10.02.21

Figure 2. Approximation's, different regions, different periods of days::

Source: author's own study

4 Conclusions

In the examples presented above, in paragraph 3, accompanied by the respective comments, the content of which explains, fully demonstrates the need to apply the information research method by dividing the interval of the independent variable integrally into several subintervals. The method of choosing the different lengths of the subintervals is commented on and argued. Thus, achieving the highlighting of subintervals with different behavior of the phenomenon studied. Special attention was paid to the involvement of external data - information, which can have an essential impact in the process of arguing for the most appropriate decisions and at a higher level of credibility. The need to apply a specific processing and analysis technique was explained, which involves the involvement of several sources of information from the environment in which the respective data were collected, including the environment of social economic activities relating to the essence of the content of the data set. Processing a data set without taking into account the environment and characteristics of social economic activities can reduce the level of efficiency and credibility.

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. Chişinău: ASEM, 2022, pp. 309-312. ISBN 978-9975-3590-6-1 (PDF). https://irek.ase.md/xmlui/handle/123456789/2607
2. COANDA, Ilie. The impact of data pre-processing on the assessment of the similarity of trend functions. In Annual international scientific conference "Competitiveness and Innovation in the Knowledge Economy", [online]: September 22nd-23th, 2023, Chisinau, Republic of Moldova. DOI: https://doi.org/10.53486/cike2023.44, https://irek.ase.md:443/xmlui/handle/123456789/3096
3. COANDA, Ilie. the level of similarity as a functions classification measure chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://irek.ase.md:8080/xmlui/bitstream/handle/123456789/4407/Proceedings%20of%20the%2028th%20International%20Scientific%20Conference_september_2024_p340-343.pdf?sequence=1&isAllowed=y2, pp. 309-312. ISBN 978-9975-3590-6-1 (PDF). https://irek.ase.md/xmlui/handle/123456789/2607