ISBN 978-9975-3590-6-1 (PDF). DOI: 10.5281/zenodo.7563976

DOI: https://doi.org/10.53486/cike2022.37
CZU: 004.421.2

EVALUATION OF SIMILARITY OF TREND FUNCTIONS

ILIE COANDA
Department of Information Technology and Information Management
Academy of Economic Studies of Moldova
Chisinau, Republic of Moldova
Email account: ildirosv1@gmail.com

Abstract. *An approach to the evaluation of the similarity of the functions - approximating trend is proposed. The evaluation process consists of two stages: approximation techniques specific to non-linear regressions are applied, then certain procedures are used - algorithms for comparing the trend-functions obtained. Approximating functions are made up of components of polynomial form as well as terms - parameterized trigonometric functions sine and cosine. A function of this form allows us to obtain approximating functions at an acceptable level of accuracy for each individual case. Beforehand, the primary data sets are subjected to a smoothing process, which also provides for the inclusion of some parameters for the purposes of qualitative monitoring of operations to exclude exceptional values, values that, in some cases, can have a significantly negative impact. Varying the parameters of the approximating functions, in particular, of the trigonometric functions, can provide us with an approximation at a proper level of precision. In some cases, a high level of approximation accuracy can also have a negative impact. Having already obtained the trend functions for the respective data sets, we continue with the process of calculating the parameters that determine the basic fundamental properties of the obtained trend functions. For this purpose, the techniques of researching functions according to theories in the field of applied mathematics are used. Then, the domain of the independent variable is to be divided into several intervals, not necessarily of the same length, then, for each of them, the values corresponding to monotony effects, inflection points, extremes, etc. are calculated. The obtained values are to be included in the distance calculation formula.*

Keywords: similarity, trend, functions, parameters, regression, applied, mathematics.

JEL Classification: C63, I21, I23, I25, I29.

1 Introduction

In the process of data analysis, the need to research the impact of a phenomenon on a group of study objects in the same period of time may often arise. For example, the spread of a particularly contagious disease in several geographical regions relatively close to each other. With the collected data, the trend functions can be determined for each locality, then the research can be continued from the point of view of the similarities between the approximating functions (between trends).The issue of similarities between functions is always at the center of discussions, in particular, when approaching the definition of the "distance" between them, on the basis of which the level of similarity between functions can be evaluated. Without coming up with more details about the advantages and disadvantages of some well-known methodologies for calculating "distances" between functions, this paper will describe a methodology, a similarity evaluation algorithm based on the techniques of studying functions with tools from the field of mathematics.

ISBN 978-9975-3590-6-1 (PDF). DOI: 10.5281/zenodo.7563976

2 Model for evaluating the similarity between two functions

In the process of calculating distances, it is proposed to involve some of the most fundamental notions in the field of function research, such as monotony intervals, convexity and concavity, critical points, extreme values, inflection points, etc. In order to be more explicit, in Figure 1 are presented the graphs of three functions, which describe a certain phenomenon that occurs in three situations in the same period of time. In Figure 1a) the primary statistical data are presented, in Figure 1b) – the approximating functions are added to the three primary dependencies and in Figure 1c) – only the graphs of the approximating functions are displayed. The most effective method of qualitative assessment of the similarity between the three functions presented as an example is the visual one. An attempt at visual evaluation, based on the image in Figure 1a) can create quite significant discomfort, therefore it can be considered ineffective, even if, in some simple situations, at a qualitative level, such evaluation, which is impractical for high volume data cases. Thus, it is necessary to implement approaches that involve certain techniques, methodologies, algorithms that involve quantitative parameters, and that are quite efficient in the data analysis process.

In this context, an analysis of the similarity of approximating functions instead of statistical - primary ones can be considered as a "gateway" to the respective approximating functions to be subjected to research. In the image in Figure 1b) it is possible to perform a comparative visual analysis of the substitution of functions - examples of primary data with approximate ones. A comparative analysis of the images in Figure 1a) and the one in Figure 1c) highlights the openness towards a much more accessible research of the similarity between the functions.

The most important question is the one related to the argumentation of the effect of preserving the level of accuracy in the process of approximating functions. In this context, it is necessary to point out the following:

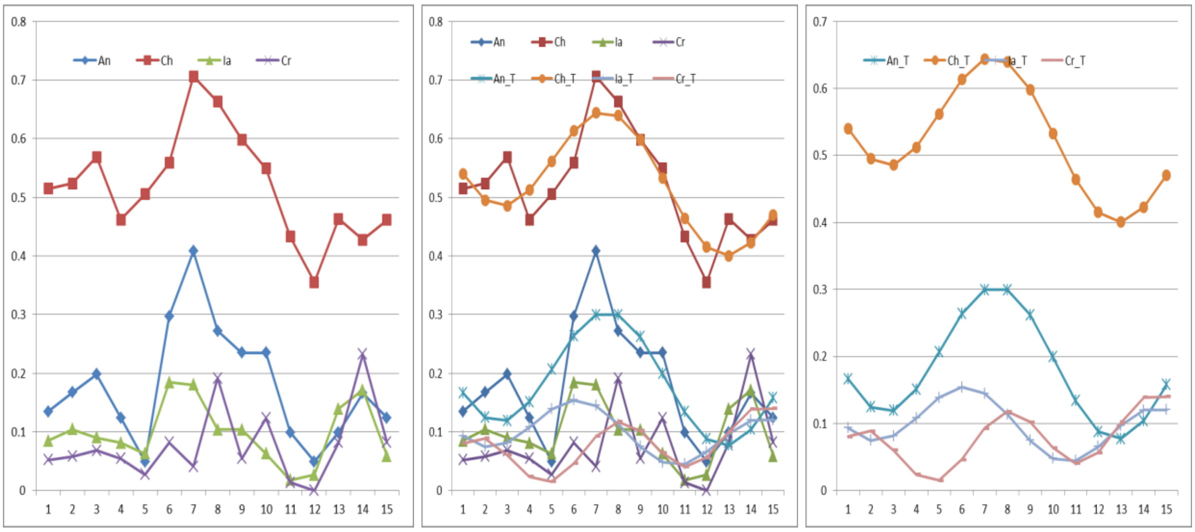
a) Statistical - primary data have some margin of errors, it depends on the field, "tools" and collection technology. It is recommended to take this into account in the process of arguing the veracity of the results obtained in the approximation processes.
b) As a rule, among the values collected there are also out-of-the-ordinary data, so they could have a significantly negative impact on the integral information. Therefore, depending on certain specific situations, including the essence of the information content, a certain "smoothing" of the peaks is necessary. The smoothing technology should preserve the essence of the underlying content of the information exposed through the primary data. In this sense, in some cases it is recommended to carry out case studies in relation to a parameter that would regulate the level of smoothing. That parameter could be defined by the relative number of suspect values as well as their opposite placement in trends. In the case of the functions-examples from Figure 1a) it is possible to determine, with certainty, a single suspicious value for each of the four functions.

Taking into account the arguments in favor of "smoothing", a smoothing which, by the way, is a rather important stage in the process of defining approximating functions, follows the description of the approach to techniques for obtaining their analytical forms. Let us consider a function **y = f(x)** on the interval **[a, b]** and ask to evaluate the error when this function **f(x)** is replaced by another function **φ(x).** To calculate the error, you can use, for example, the formula

ISBN 978-9975-3590-6-1 (PDF). DOI: 10.5281/zenodo.7563976

max | **f (x) - φ(x)** | on the interval **[a, b],** which is the so-called ***maximum deviation*** of **φ(x)** from **f(x).** However, in some cases it is more natural to take as a measure of the error the so-called ***mean square deviation*** **δ**, which is defined by the equation

$$\delta^2 = \frac{1}{(a-b)} \int_a^b [f(x) - \varphi(x)]^2 dx, \tag{1}$$



a) primary functions **b) primary functions, functions approximating functions** **c) approximating**

Figure 1 Examples of primary data and their approximating functions

Source: author's own study

where the function **φ(x),** the approximating function, has the form

$$\varphi(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx)\,] \tag{2}$$

Taking into account expression (2), from expression (1) we obtain

$$\delta^2 = \frac{1}{(a-b)} \int_a^b \left[f(x) - \frac{a_0}{2} - \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx)\,] \right]^2 dx \tag{3}$$

considering a particular case of the function **φ(x)**

$$\varphi(x) = a + bx + c \sin(p\, x) \tag{4}$$

the expression for $\boldsymbol{\delta^2}$ is obtained

$$\delta^{\,2} = \sum_{i=1}^{n} [f(x_i) - (a + b\, x_i + c\, sin(p\, x_i)]^2 \tag{5}$$

a,b,c are coefficients to be calculated by the least squares method. For different values of **p**, different sets of **a,b,c** values will be obtained. It is necessary to involve an algorithm for

ISBN 978-9975-3590-6-1 (PDF). DOI: 10.5281/zenodo.7563976

determining the value of **p** so that the set of values **a,b,c** ensures the smallest possible value of the value. One of the simplest options is to calculate the values of **a,b,c** for a series of values of **p** on an interval, for example **[0,1]** ,and choose the value of **p** that provides the smallest value **δ**.

The graphs of the three approximating functions obtained by the method of least squares using expression (5) are shown in Figure 1c) (without the functions of the primary data), and which offer a possibility to evaluate the similarity of some differentiable functions of any order. Thus, methods can be applied to research functions by using mathematical methods usable only for continuous and differentiable functions.

3 Steps of the proposed algorithm for evaluating the similarity between functions

1. The necessary operations for smoothing the functions determined on the basis of the primary data are launched.
2. Non-linear regression technology is applied for each set of primary data according to formula (5), thus obtaining the analytical forms of approximating functions.
3. Steps of the algorithm for calculating the level of similarity between functions (in the following, we will only refer to approximating functions). Consider two functions ***f1(x)*** and ***f2(x):***

a) The definition interval of the function ***[a,b]*** is divided into ***n*** segments of the same length;
b) For ***f1(x)*** two vectors of size ***n*** are defined: ***V11*** with the coordinate equal to ***1*** (plus one), if the derivative of the first order in the middle of the interval ***i,(i=1,…,n)*** is positive, and, respectively, equal to ***-1*** (minus one), if the derivative of the first order is negative.
c) Similarly, the vector ***V12*** corresponding to the values of the second-order derivative of the function ***f1(x)*** is defined.
d) Similarly, the vectors ***V21*** and ***V22*** are also defined for the second function ***f2(x)..***
e) "Distance" ***d*** between these two functions ***f1(x)*** and ***f2(x)..***is calculated according to the formula

$$d^2 = \frac{1}{4n}\sum_{i=1}^{n}\left((V11_i - V21_i)^2 + (V12_i - V22_i)^2\right) \quad (6)$$

Algorithm testing: fictitious data for **4** locations: (***An, Ch, Ia, Cr),*** respective values for parameter **p** (5) formula: **p**(0.62,0.62,0.75,0.98), number of intervals **n**=20, distances obtained: (d12=0 ; d13=0.54 ; d14=0.7 ; d23=0.54 ; d24=0.7 ; d34=0.69) , d13 - calculated distance from ***An*** to ***Ia*** location.

4 Conclusions

Case studies have demonstrated the effectiveness of the algorithm described in this paper. Some of the most important elements of the algorithm are exposed. However, some key elements need to be further explored. This is the smoothing technology, the analytical model form of the approximating function, the determination of the values for the parameter **p**, the number of intervals **n**, as well as the inclusion of the **cos(qx)** term.