Statistical Analysis of the Popularity of Programming Language Libraries Based on StackOverflow Queries

Ihor Rishnyak[1], Yurii Matseliukh[1], Taras Batiuk[1], Lyubomyr Chyrun[2], Oleksandra Strembitska[1], Oksana Mlynko[1], Viktoriia Liashenko[1] and Andrii Lema[1]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Ivan Franko National University of Lviv, University Street, 1, Lviv, 79000, Ukraine*

Abstract

This paper presents a statistical analysis of existing trends in the spread of programming language libraries based on data set studies. The various problems that arise when using specific libraries of different programming languages for certain periods, the most common is the month, are studied and analyzed. The results of the study of existing trends in the spread of programming language libraries, collected in the studied dataset, are presented graphically, set key descriptive characteristics, taking into account the correlation of data. Trends in the behavior of the studied indicators using the methods of smoothing time series are determined. A cluster analysis of programming language libraries was performed, making it possible to group data by clusters and form appropriate data groups for ranking programming language libraries.

Keywords

Statistical analysis, information technologies, business analysis, programming language libraries, StackOverflow queries, data processing

1. Introduction

The rapid growth in popularity of programming language libraries based on Stack Overflow queries has not yet solved the problem of solving complex technical questions that cannot be answered through queries on the Internet. A typical problem is when developers, looking for answers by submitting queries in search engines, get all kinds of results, often spam or incorrect, outdated, and sometimes off-topic. You often have to look for a blog post and then sit with the source for a long time (more than ten minutes) to identify a way to solve a technical problem in a particular post. Stack Overflow is a place where developers ask and get a reliable answer. Stack Overflow allows developers to improve their level as a programmer, using the experience of others. It increases the code experience even for those already experienced, helping others who have not been able to figure it out themselves. Stack Overflow is the formation of future technologies as the world's future. The above proves the relevance of the study of the popularity of libraries of programming languages, where it is crucial to analyze the composition, structure, and issues of various queries in specific libraries each month. This study is especially relevant for beginners who are now trying to choose a language. The problem's urgency is no less for experienced developers to expand their knowledge in studying each subsequent programming language.

From the point of view of business analysts, this analysis can be considered the creation of a library rating system, i.e., identifying the most significant number of queries and determining the most popular languages. Based on the analyzed data, the business analyst will be able to assess the decline, growth,

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: ihor.v.rishnyak@lpnu.ua (I. Rishnyak); indeed.post@gmail.com (Y. Matseliukh); taras.batiuk.mnsa.2020@lpnu.ua (T. Batiuk); Lyubomyr.Chyrun@lnu.edu.ua (L. Chyrun); oleksandra.strembitska.sa.2019@lpnu.ua (O. Strembitska); oxanamlunko@gmail.com (O. Mlynko); viktoriia.liashenko.sa.2019@lpnu.ua (V. Liashenko); andrii.lema.sa.2019@lpnu.ua (A. Lema)
ORCID: 0000-0001-5727-3438 (I. Rishnyak); 0000-0002-1721-7703 (Y. Matseliukh); 0000-0001-5797-594X (T. Batiuk); 0000-0002-9448-1751 (L. Chyrun); 0000-0003-2754-7076 (O. Strembitska); 0000-0001-9878-6846 (O. Mlynko);0000-0003-0966-7912 (V. Liashenko); 0000-0001-6490-6221 (A. Lema)

and invariability of the popularity of languages in 2009-2019 and offer his vision of the possible development of specific languages.

The work aims to use the main methods of visualization, graphical display, and primary statistical processing of numerical data presented by a sample or time series to identify trends in the studied indicators of programming language libraries, present the nature of their trends, apply time series smoothing methods and tabulation MS Excel. Determination by methods of correlation analysis of experimental data presented by time sequences.

2. Related Works

Research on the popularity of programming languages, according to scientists [1-3], is one of the components of the problem of human capital development. Having hard skills employees and acquiring soft skills is an important task, as it allows you to solve important social [4-5], economic [6-8] and technical [9-23] issues. Since its inception, the system we consider in this paper has provided an opportunity to ask questions about programming and get answers to them for 12 years [24-36].

Confectioners discuss recipes in culinary forums; students discuss their questions in help groups in telegrams; parents of these children have joint chats on Viber, where they solve problems. Older people gather under the porches to discuss neighbors or world news, i.e., every branch of people or professionals should have a place where he can ask his question, hear an expert's opinion, discuss a topic or give advice. Therefore, the importance of using the StackOverflow system is beyond doubt.

Its relevance has been described in many articles [24-36] and videos on YouTube and other social networks. Also, if you practice programming, you are interested in specific questions and decided to enter them in the Google search. One of the first results will be the site Stack Overflow. One example of relevance is the Wikipedia site. He reports that a 2016 study by Android developers using Stack Overflow generated ten times more functional code (but less secure, which is a disadvantage) than developers using official documentation [30, 37].

In researching the chosen topic, we considered the HABR website [31], which was created to publish news and opinions related to IT and business. These libraries of programming languages will be our attributes.

- Month is here are the day-month data on the library in the StackOverflow program;
- NLTK is the number of queries about the NLTK library (a set of libraries and programs for symbolic and statistical processing of natural languages for English, written in the Python programming language);
- spaCy is the number of requests for the spacy library (open-source library for advanced natural language processing (NLP) in Python);
- Stanford-NLP is number of queries about Stanford - NLP library;
- Python is the number of queries about the Python library;
- R is number of requests for r library;
- NumPy is the number of queries about the NumPy library (python language extension);
- SciPy - the number of queries about the spicy library;
- MATLAB is the number of queries about the MATLAB library;
- Machine-Learning - the number of requests for machine learning.

In these works [31], the user described his story. For seven years, he used the system we are discussing, and during this time, he "answered 3516 questions, asked 58, entered the hall of fame in several languages, met many wise people, and actively used all the site's features.

The issues of the most popular programming languages under discussion and their libraries are already discussed on the website StackOverflow [30, 37].

Is it possible to trust the answers to your questions on the site and actively use them? - After all, users are interested in each question and will quickly correct it in case of error. The HABR website [31] also explains that the average programmer cannot write code without a break of several hours. Therefore, to avoid unnecessary distractions and overloads, you can have a great time with like-minded people on StackOverflow [30, 37]. The user is set a rating by answering the question, and his

"reputation" can rise exponentially, depending on the site activity. After a reputation mark of 25,000, the user gets access to all SO statistics and permission to store queries in the user database.

Thus, the SO system is one of the most popular among professional software developers, system administrators, and programmers. All questions are marked with a specific topic tag (or multiple tags, depending on the topics involved) to which the question relates. By clicking on the label, you can view their list to select the topic that interests you. In our case, these are the themes of libraries of different programming languages.

3. Methods

To solve the problems in this work, we will use standard methods [38-45].

The correlation field is a graph that establishes a relationship between variables, where X of each corresponds to the value of the factor feature (abscissa), and Y - the value of the resultant feature (ordinate) of a particular unit of observation. The number of points on the Graph corresponds to the number of observation units. The location of the points indicates the presence and direction of communication [38-45].

Building a correlation field is carried out mainly in the following steps: choose two variables that change over time. Then measure the value of the dependent variable and enter the result in the table. Then construct a coordinate plane on the X-axis to indicate the value of the independent variable and on the Y-axis - the dependent. Then you need to mark the points of the correlation field on the Graph. On the X-axis for the first value of the independent variable, mark a point on the Y-axis corresponding to the value of the dependent variable. The resulting set of points is called the correlation field [38-45]. We analyze the received schedule and conclude the presence of communication or its absence.

Correlation coefficient is an indicator used to measure the density of the relationship between traits in the correlation-regression model of linear dependence [46-52]. The absolute value of the correlation coefficient ranges from -1 to +1.

The correlation ratio determines the correlation in any of its forms, namely in, straight or curved. The correlation ratio can be determined to estimate the curvilinear relationship between the values of X and Y. It always has a positive value and is in the range from 0 to + 1. The value of the zero ratios is taken when the relationship between the features is absent [38-52].

Autocorrelation is the correlation of a function with itself shifted by a certain amount of independent variable. The autocorrelation function graph can be obtained by plotting the correlation coefficient of two functions along the ordinate axis and the value along the abscissa axis [38-45]. The autocorrelation function measures the linearity of the relationship between the elements of the time series spaced apart at x points in time. The Graph of an autocorrelation function is called a correlogram.

The correlation matrix is a table that represents the values of the correlation coefficients for different variables. It shows the numerical value of the correlation coefficient for all combinations of variables. It is generally used when we need to determine the relationship between more than two variables. It consists of rows and columns that contain variables, and each cell contains coefficient values that inform the degree of association and linear relationship between two variables [38-45]. In addition, it can be used in specific statistical analyzes. Multiple linear regression, where we have several independent variables and a correlation matrix, helps determine the degree of association.

The multiple correlation coefficient describes the correlation's intensity, or the relationship's degree of closeness, between a dependent variable and several independent variables [38-45]. Its value cannot be less than the absolute value of any partial or straightforward correlation coefficient. The primary indicator of the closeness of the connection in multiple correlations is the coefficient of multiple correlations, which has a value from 0 to +1.

4. Experiments

The structure of the dataset [24] is presented in the Table 1 and has ten fields, among which month, NLTK (Natural Language Toolkit), spaCy, Stanford-NLP, Python, R, NumPy, SciPy, MATLAB, Machine-Learning, and 132 rows for 12 years.

Table 1
The dataset structure of the programming language libraries based on StackOverflow queries

month	NLTK	spaCy	Stanford-NLP	Python	R	NumPy	SciPy	MATLAB	Machine-Learning
09-Jan	0	0	0	631	8	6	2	19	8
09-Feb	1	0	0	633	9	7	3	27	4
…	…	…	…	…	…	…	…	…	…
19-Nov	72	79	14	23602	4883	1297	199	479	918
19-Dec	82	72	13	20058	4150	1118	159	349	983

Charts are used to represent data on a sheet graphically. There are several standard chart types in Excel. Charts can be ¬placed directly on the sheet next to the data used to build the chart. Such charts are called embedded. In addition, the chart can occupy a separate sheet in the book, which is called a chart sheet. No matter how the chart was created, it is always linked to the sheet data. If the data changes, the chart will be updated automatically [33]. The graphical form of data representation is called a chart. In the form of a chart, you can provide sets of numbers, sums of money, percentages, dates, and time values. Chart is created using the Chart Wizard, launched by the Chart Wizard button on the Standard toolbar (Fig. 1). Output Range is a range of spreadsheet cells ¬that contains data that will be displayed graphically or in the form of textual explanatory elements. A graphic representation of a single value is called a data element in the chart. A row of data is a sequence of data arranged in a single row or column of a spreadsheet and displayed graphically on a chart (Fig. 2). Typically, the value shown in the diagram depends on another value or set of text values. Such independent values and text values are called data categories [33].

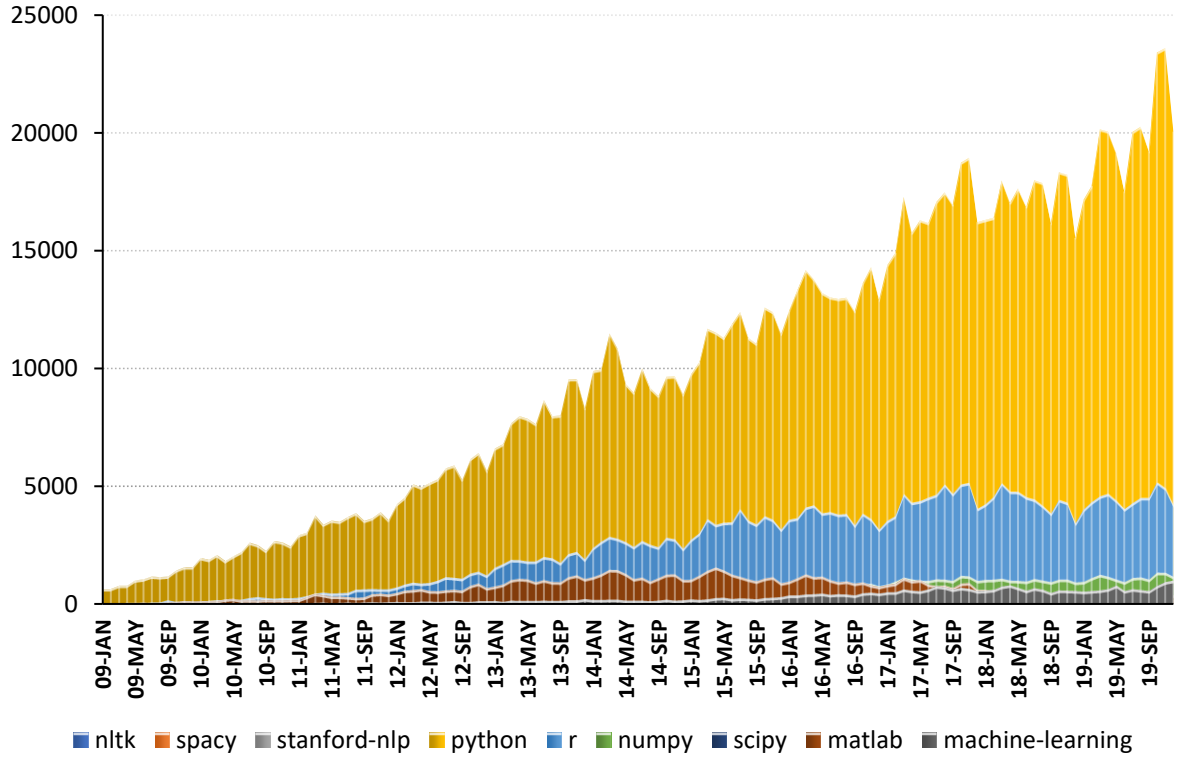


Figure 1: Graphical data representation of queries by date in the Cartesian coordinate system

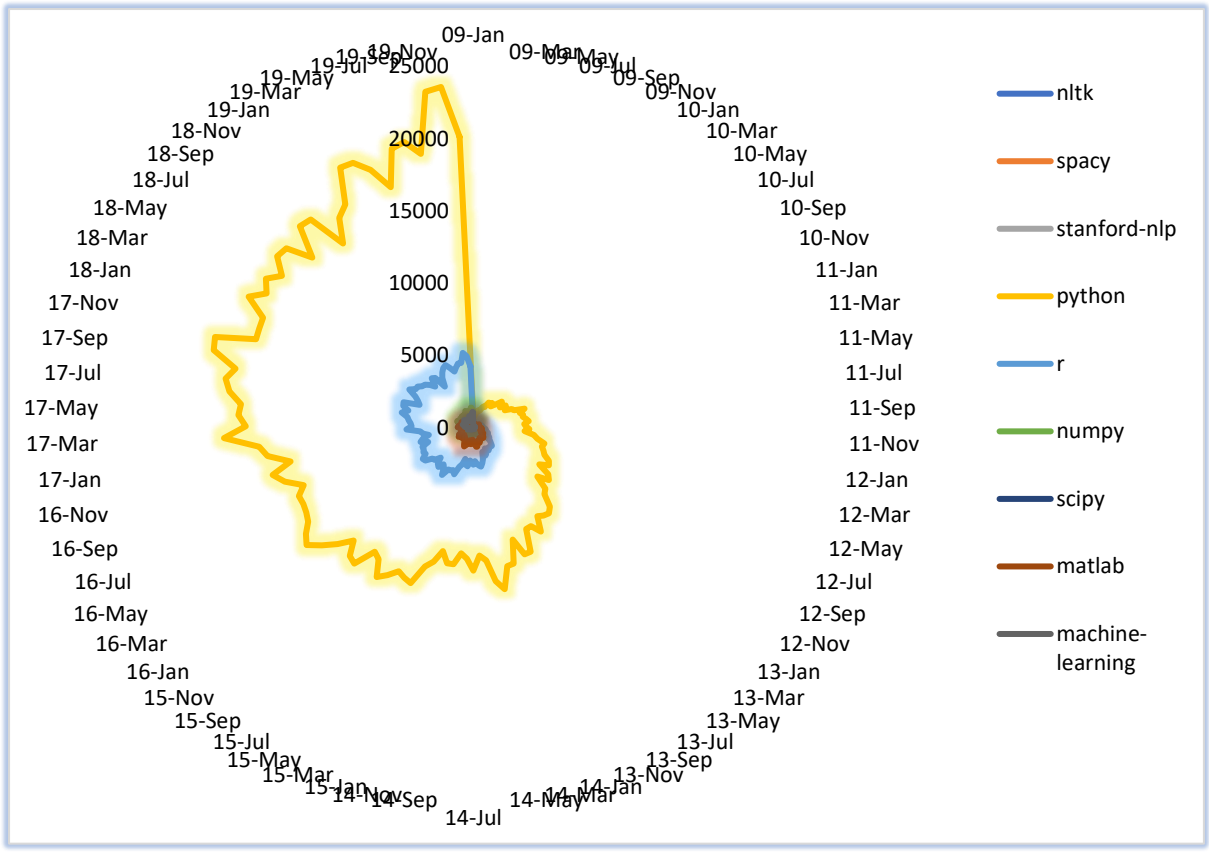


Figure 2: Graphical data representation of queries by date in the polar coordinate system

Descriptive statistics [25-29, 32-36] provide the basis for the formation of competencies for choosing a measurement scale, automation of data processing using different formats at the stage of their collection, presentation of results in various forms, graphical presentation of results, calculation of statistical distribution parameters, and evaluation of general population parameters using information technology. It selects quantitative information necessary (or interesting) for different people. Large data sets must be generalized or collapsed before humans can study them. It is what descriptive statistics do, which describes, summarizes, or reduces the properties of data sets to the desired type. Descriptive statistics are used to analyze and interpret statistical data, construct statistical distributions and calculate the relevant numerical parameters that characterize the study population. It is used to organize information collection, check the quality of data and their interpretation, and the image of statistical material [25-29, 32-37]. A result of descriptive statistics shows in the Table 2.

The construction of histograms interprets the distribution data more apparent [32]. It involves dividing the entire range of possible values of X into a finite number of intervals (in the multidimensional case - rectangular) and counting the number of implementations that fall into each of them (Fig. 3).

Cumulate is the curve of the interval variation series's accumulated frequencies [34]. The Graph of the integral distribution function $F(x)$ *is* compared with the cumulative and is also considered in probability theory [34]. The concepts of histograms and cumulates are associated with continuous data and their interval variation series [34]. Their graphs are empirical estimates of the probability density and distribution function (Fig. 3).

The methods of smoothing time series are the method of moving average, exponential smoothing, adaptive smoothing, and their modifications [25-29, 32-36]. They are used to reduce the influence of a random component (random fluctuations) in time series. They make it possible to obtain more "pure" values, which consist only of deterministic components. Some of the methods aim to highlight some components, such as trends [25-29, 32-36]. Smoothing methods can be divided into two classes based on analytical and algorithmic approaches.

Table 2
The dataset structure of the programming language libraries based on Stack Overflow queries

Index	NLTK	spaCy	Stan-ford-NLP	Python	R	NumPy	SciPy	MAT-LAB	Machi ne-Learni ng
Mean	42,70	11,85	25,54	9856,70	2411,86	514,20	112,45	651,68	264,40
Standar d Error	2,53	1,83	1,99	541,47	149,25	34,20	6,06	34,46	21,73
Median	44,50	0,00	17,50	9651,50	2613,50	486,00	130,50	581,00	154,50
Mode	0,00	0,00	0,00	-	139,00	6,00	2,00	99,00	8,00
Standar d Deviatio n	29,02	21,07	22,82	6221,07	1714,76	392,88	69,68	395,95	249,66
Sample Variance	842,4 2	443,8 1	520,80	38701728 ,16	294039 9,25	154357, 03	4855,4 1	156776, 11	62327, 85
Kurtosis	-1,23	2,15	-0,79	-1,15	-1,51	-1,32	-1,33	-1,04	-0,57
Skewnes s	0,05	1,80	0,66	0,17	-0,06	0,22	-0,28	0,13	0,80
Range	106,0 0	79,00	79,00	22971,00	5136,00	1306,00	227,00	1516,00	981,00
Minimu m	0,00	0,00	0,00	631,00	2,00	4,00	2,00	19,00	2,00
Maximu m	106,0 0	79,00	79,00	23602,00	5138,00	1310,00	229,00	1535,00	983,00
Sum	5637, 00	1564, 00	3371,0 0	1301085, 00	318365, 00	67875,0 0	14844, 00	86022,0 0	34901, 00
Count	132,0 0	132,0 0	132,00	132,00	132,00	132,00	132,00	132,00	132,00
Largest (2)	94,00	79,00	79,00	23414,00	5117,00	1297,00	223,00	1433,00	918,00
Smallest (2)	0,00	0,00	0,00	633,00	4,00	6,00	2,00	24,00	3,00
Confide nce Level (95.0%)	5,00	3,63	3,93	1071,17	295,25	67,65	12,00	68,18	42,99

The simplest way of forecasting is considered to be the approach that determines the forecast estimate from the achieved level using the average level, average growth, and average growth rate—extrapolation based on the average level of the series [25-29, 32-36]. When extrapolating socio-economic processes based on the average level of the series, the predicted value is taken as the arithmetic mean of the previous levels of the series. The reliability interval considers the uncertainty hidden in the estimate of the mean. However, the projected indicator is assumed to be equal to the average sample value. The approach doesn't consider that individual indicator values fluctuated around the average in the past [25-29, 32-36]. It will also happen in the future.

Methods of analytical smoothing include regression analysis and the method of least squares and its modifications [25-29, 32-36]. To identify the primary trend by the analytical method means to give the studied process the same development throughout the observation period. Therefore, for 4 of these methods, choosing the optimal function of the deterministic trend (growth curve) is essential, which smooths out several observations.

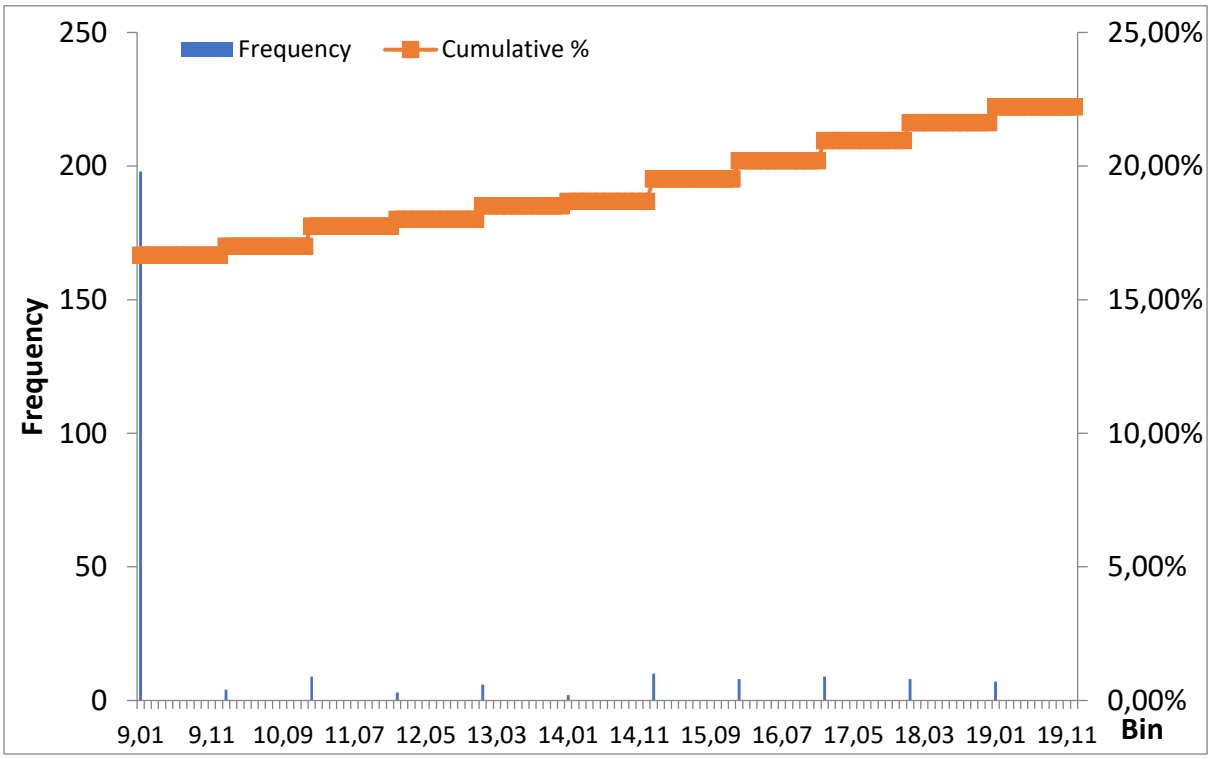


Figure 3: The diagrams of the distribution data of queries – frequency and cumulate

Forecasting methods based on regression methods are used for short-term and medium-term forecasting. They do not allow adaptation: the forecasting procedure must be repeated first with the receipt of new data. The optimal length of the lead period is determined separately for each economic process, taking into account its statistical instability.

5. Results

The most commonly used method is smoothing time series using moving averages [25-29, 32-36]. The algorithm for calculating the moving average is as follows [25-29, 32-36].

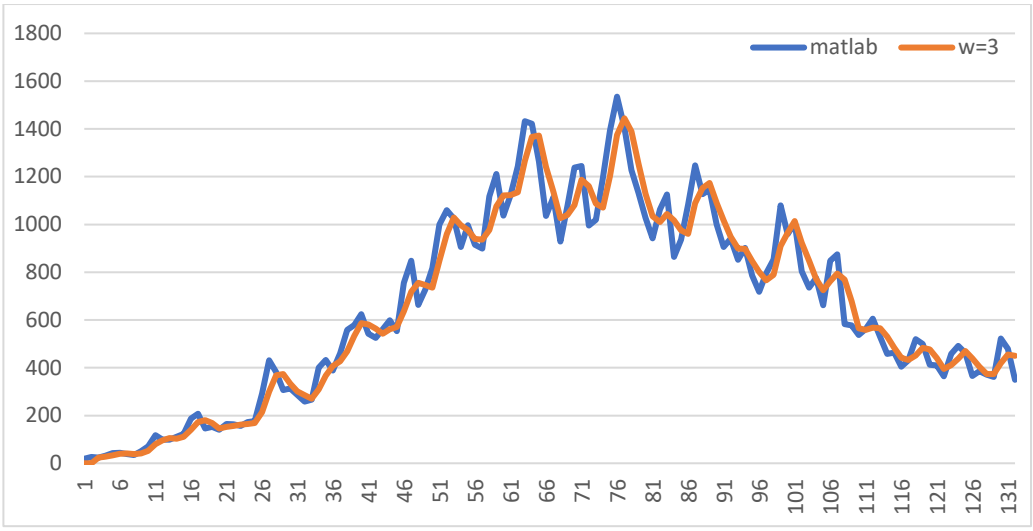
$$\tilde{y}_i = y_1^* + y_2^* + \ldots + y_k^* + \sum_{j=k+1}^{N-2k}\left[\frac{1}{w}\sum_{i=j}^{j+2k+1} y_i\right] + y_{N-k}^* + \ldots + y_{N-1}^* + y_N^* \quad (1)$$

Algorithm for calculating the weighted average is as follows [25-29, 32-36].

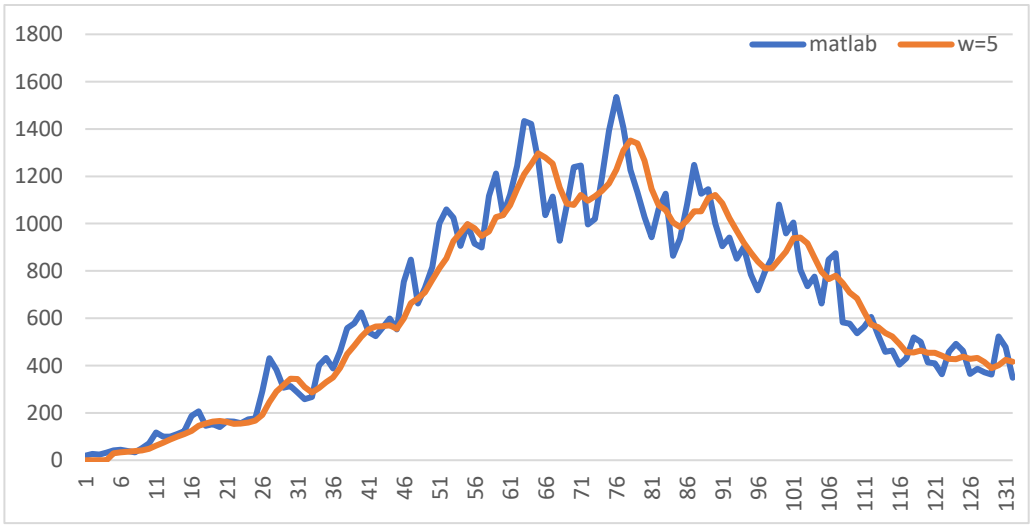
$$\tilde{y}_i = y_1^* + y_2^* + \ldots + y_k^* + \sum_{j=k+1}^{N-2k}\left[\frac{1}{w}\sum_{i=j}^{j+2k+1} \alpha_i y_i\right] + y_{N-k}^* + \ldots + y_{N-1}^* + y_N^* \quad (2)$$

5.1. Smoothing according to Kendel formulas - simple moving average

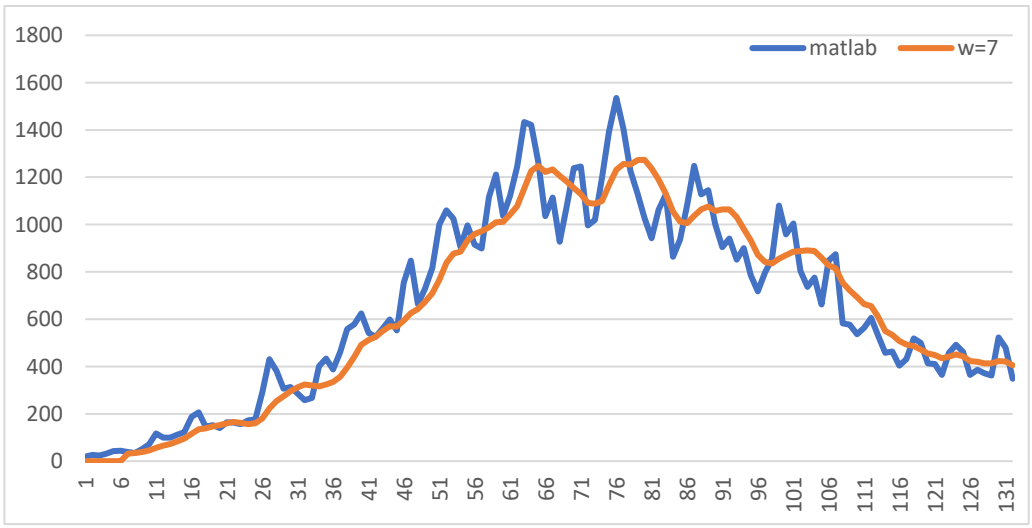
Smooth the data using the dimensions of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 are presented in Fig. 4-Fig. 6. The smoothed data for queries about MatLab are calculated using to Kendel formulas for the smoothing interval w = 3 (Fig. 4, a), w = 5 (Fig. 4, b), w = 7 (Fig. 4, c), w = 9 (Fig. 5, a), w = 11 (Fig. 5, b), w = 13 (Fig. 5, c), w = 15 (Fig. 6).



a



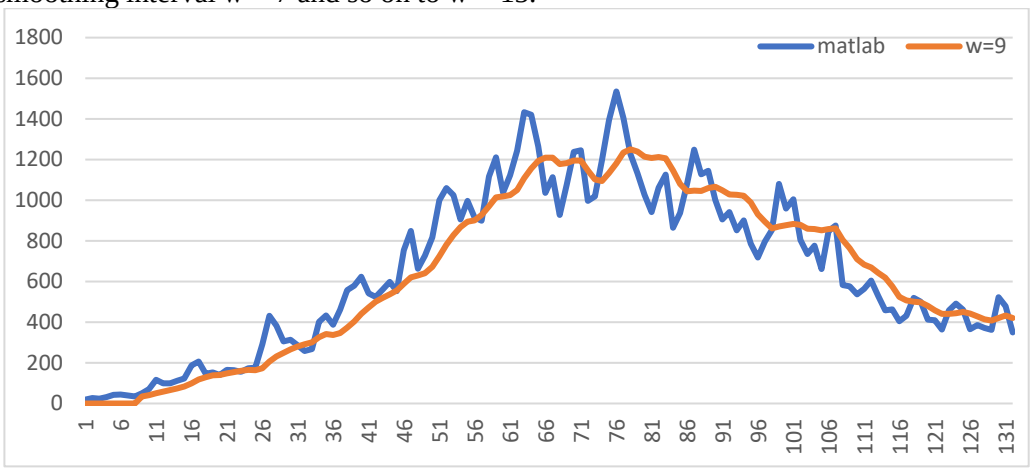
b



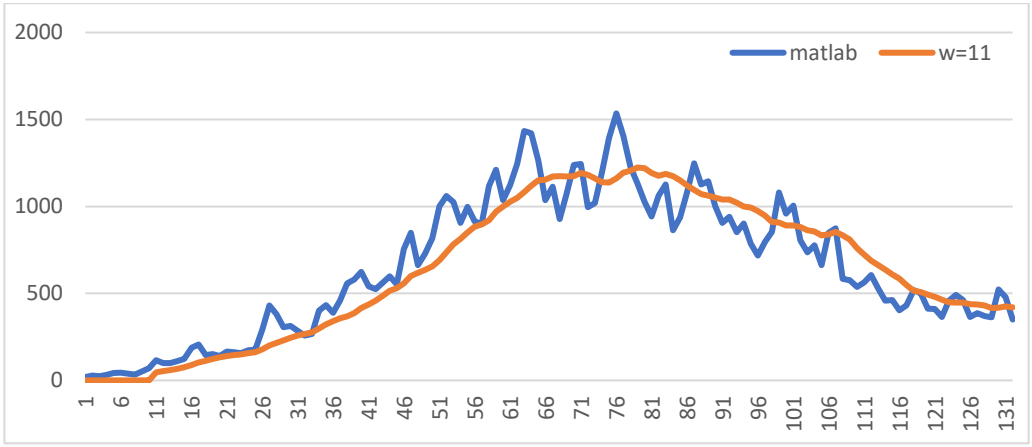
c

Figure 4: The smoothed data for queries about MatLab using the smoothing interval w = 3 (a), w = 5 (b), w = 7 (c)

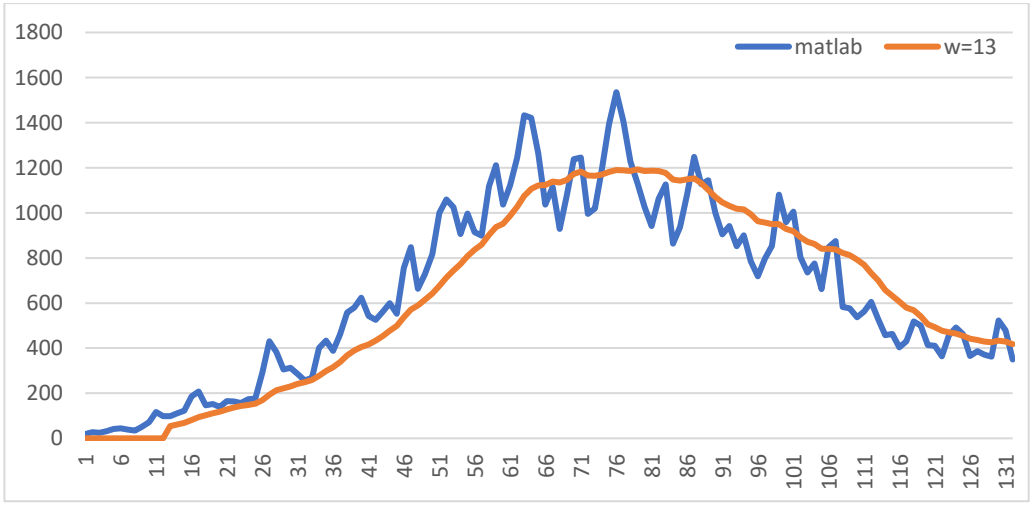
We smoothed the data using the smoothing interval w = 3, then we smoothed the obtained smoothed data again, but use the size of the smoothing interval w = 5. We continued smoothing the obtained data with a smoothing interval w = 7 and so on to w = 15.



a



b



c

Figure 5: The smoothed data for queries about MatLab using the smoothing interval w = 9 (a), w = 11 (b), w = 13 (c)

The smoothed data for queries about MatLab obtained by the smoothing data again. In Fig. 7 there are presented the smoothed data for queries about MatLab using the smoothing interval w = 5 (w = 3) (a), w = 7 (w = 5) (b). Fig. 8 shown the smoothed data for queries about MatLab for w = 9 (w = 7) (a), w = 11 (w = 9) (b), w = 13 (w = 11) (c), w = 15 (w = 13) (d) according to Kendel formulas.

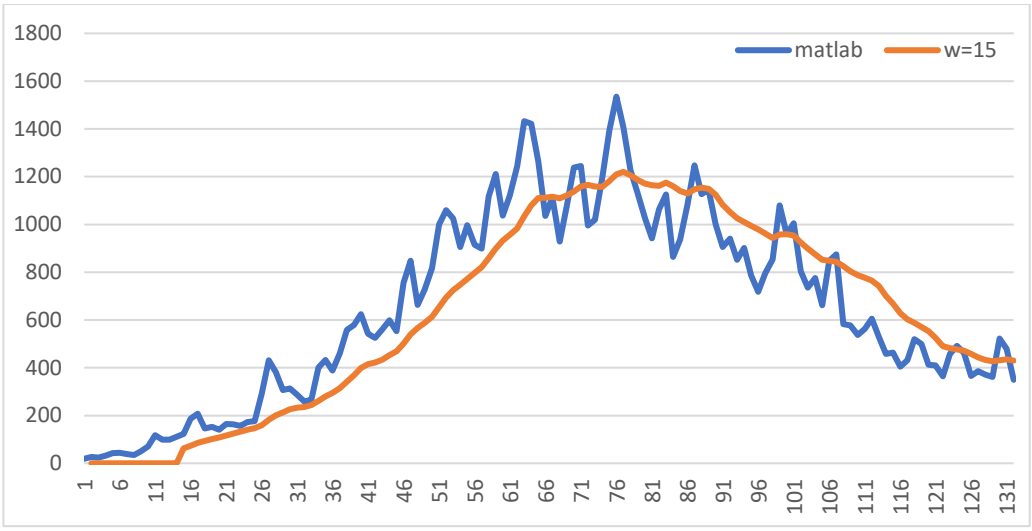
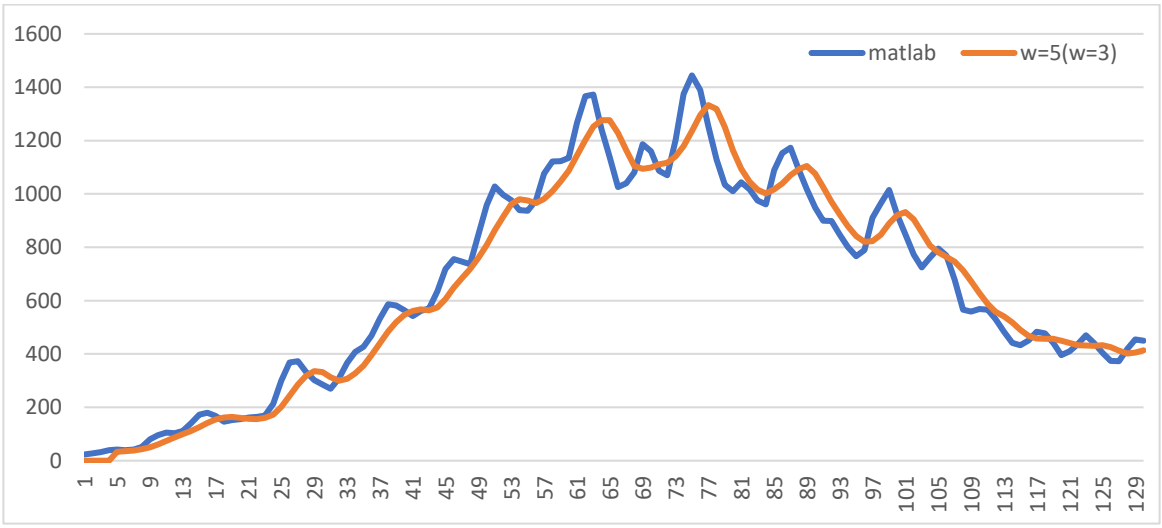
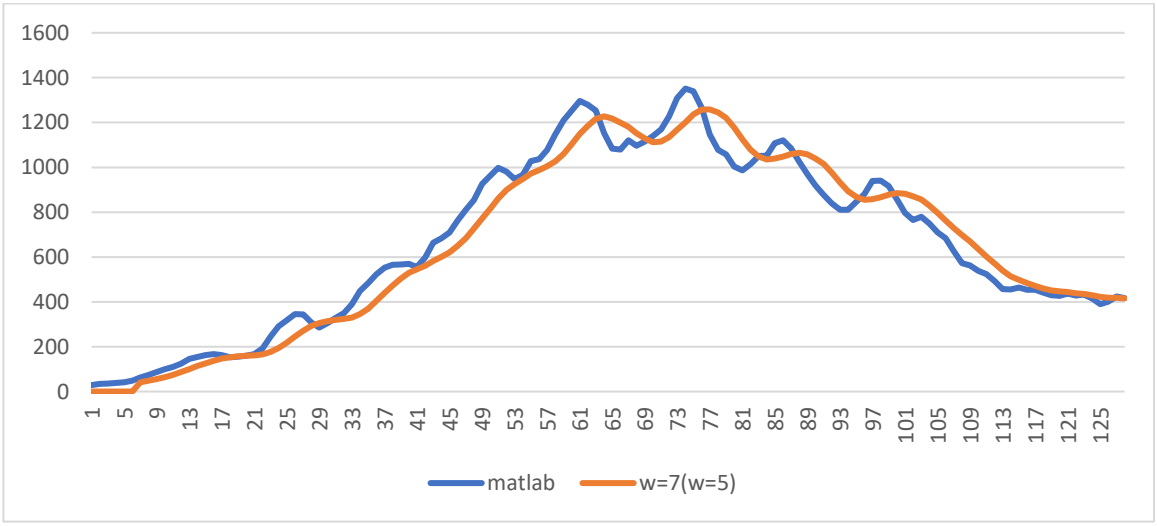


Figure 6: The smoothed data for queries about MatLab using the smoothing interval w = 15 according to Kendel formulas



a



b

Figure 7: The smoothed data for queries about MatLab using the smoothing interval w = 5 (w = 3) (a), w = 7 (w = 5) (b)

In both cases, we find for each smoothing the number of turning points and correlation coefficients between the original values and the smoothed ones.

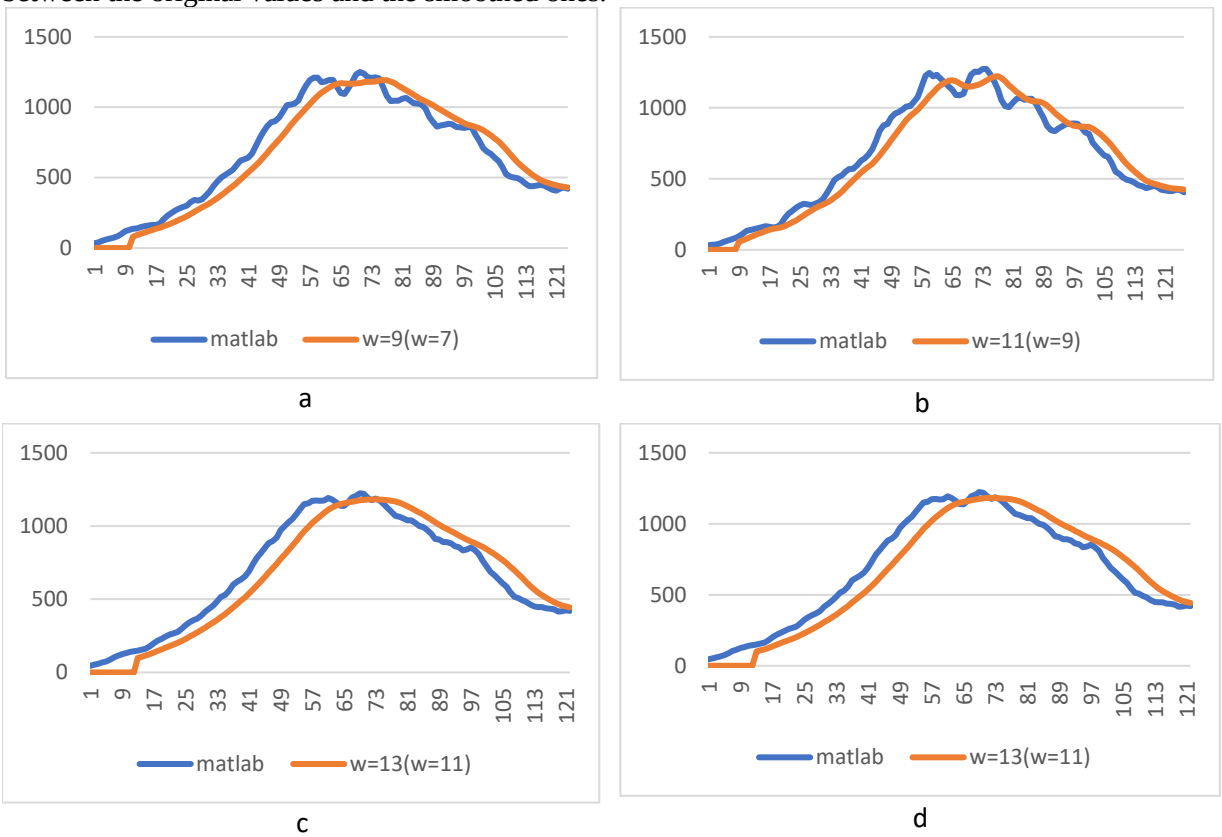


Figure 8: The smoothed data for queries about MatLab using the smoothing interval w = 5 (w = 3) (a), w = 7 (w = 5) (a), w = 9 (w = 7) (a), w = 11 (w = 9) (a), w = 13 (w = 11) (a), w = 15 (w = 13) (a) according to Kendel formulas

The correlation coefficients between the original values and the smoothed ones are calculated in Table 3.

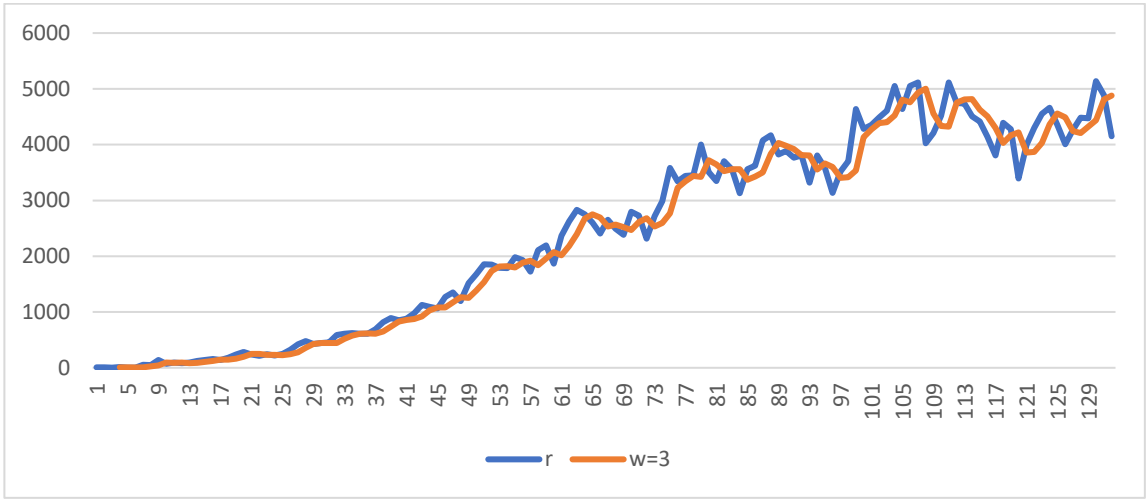
Table 3
The correlation coefficients between the original values and the smoothed ones

Interval w	3	5	7	9	11	13	15	5 (3)	7 (5)	9 (7)	11 (9)	13 (11)	15 (13)
Correlat. coeffic.	0,980	0,962	0,953	0,939	0,925	0,916	-	0,977	0,971	0,965	0,958	0,953	0,950
Number of correct turning points	36	30	24	23	16	14	14	20	8	4	4	2	2

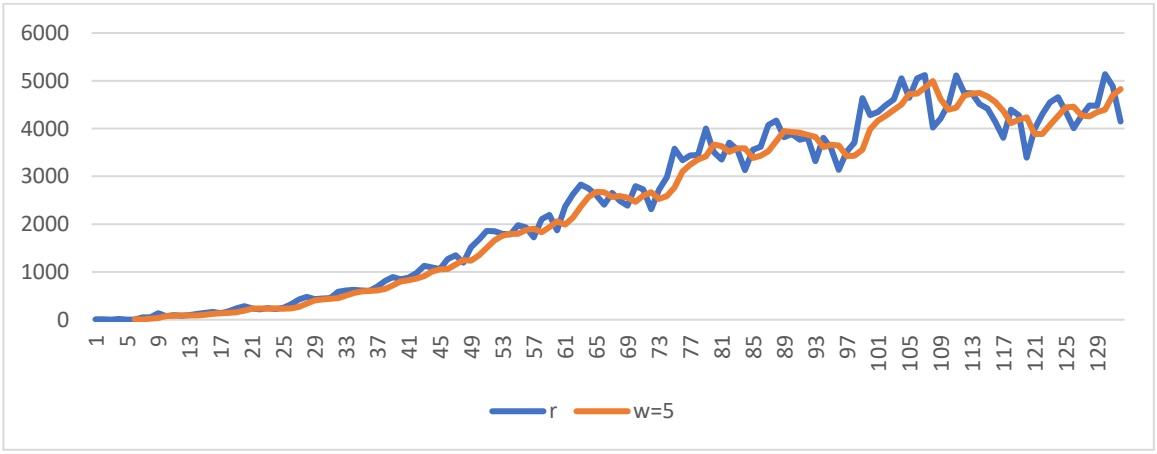
5.2. Smoothing according to Pollard formulas

John Pollard's algorithm, proposed by him in 1975, is used to factorize integers [28]. It is based on Floyd's algorithm for finding the length of the cycle in the sequence and some consequences of the paradox of birthdays. The algorithm most effectively factored composite numbers with relatively minor factors in the decomposition. All of Pollard's ρ-methods construct a numerical sequence, the elements of which form a loop, starting with some number n, which can be illustrated by the arrangement of numbers in the Greek letter ρ. It was the name for a family of methods [28].

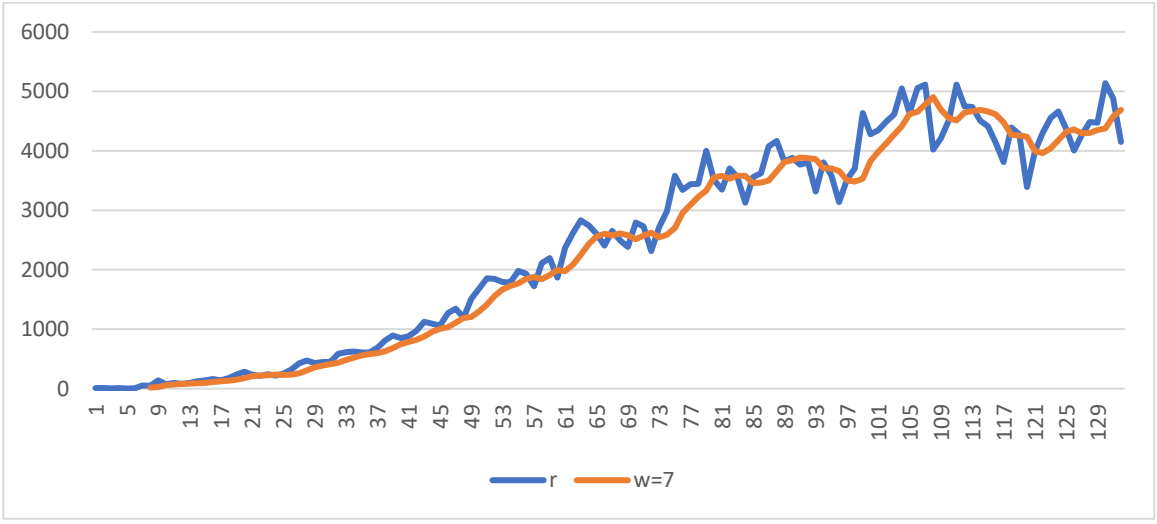
We smooth the data for queries about R using the same dimensions of the smoothing interval (w = 3, 5, 7, 9, 11, 13, 15). It is presented in Fig. 9-Fig. 11. The smoothed data for queries about R are calculated using Pollard formulas for the smoothing interval w = 3 (Fig. 9, a), w = 5 (Fig. 9, b), w = 7 (Fig. 9, c), w = 9 (Fig. 10, a), w = 11 (Fig. 10, b), w = 13 (Fig. 10, c), w = 15 (Fig. 11).



a

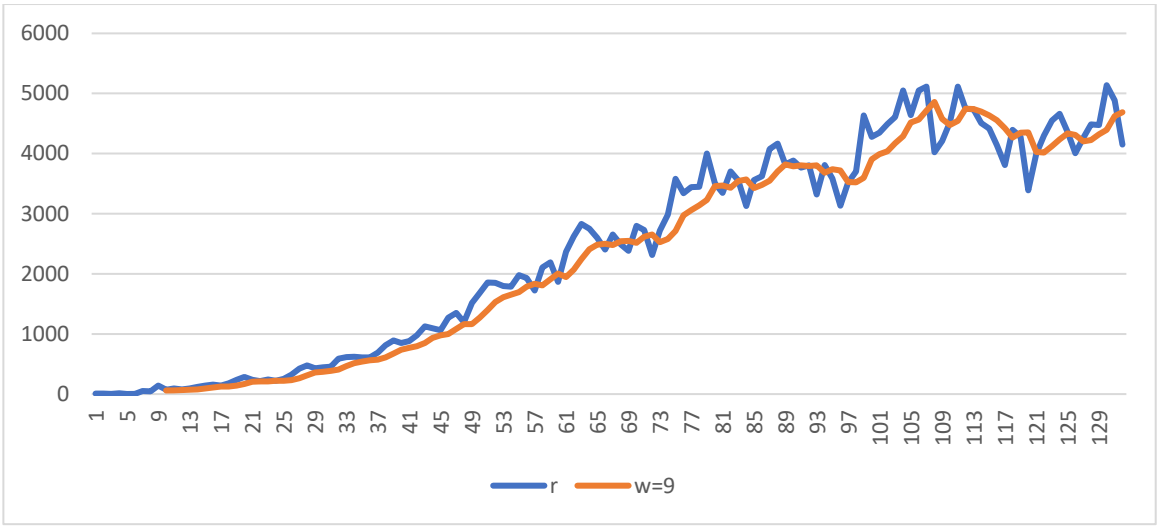


b

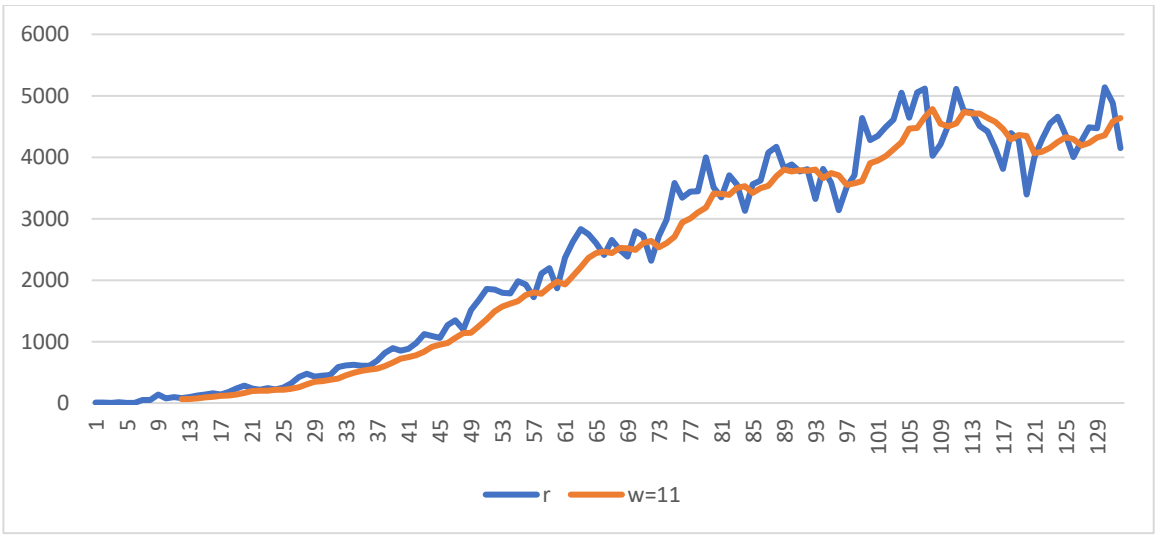


c

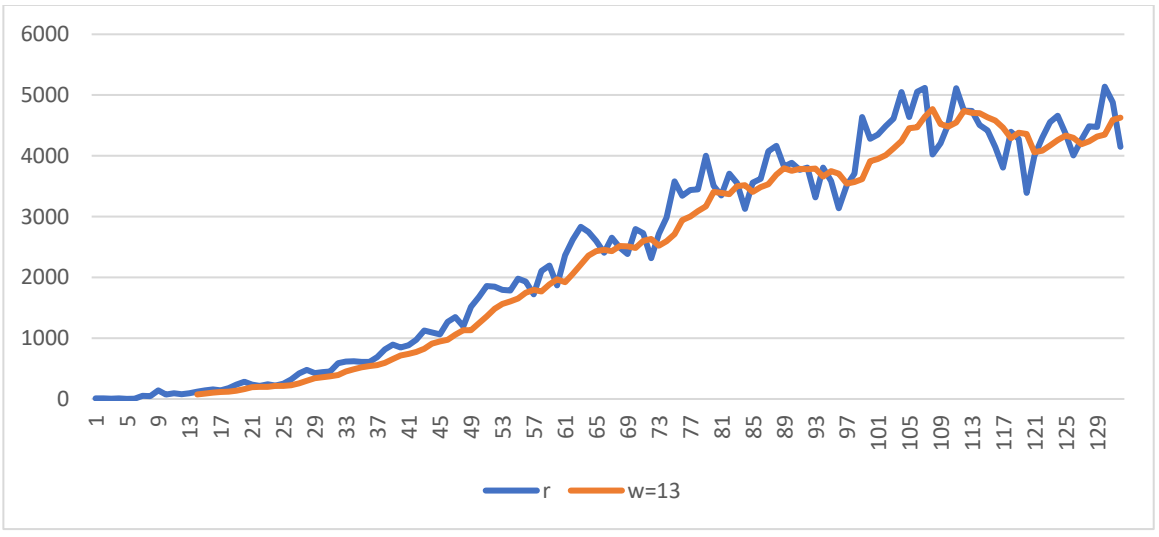
Figure 9: The smoothed data for queries about R using the smoothing interval w = 3 (a), w = 5 (b), w = 7 (c)



a



b



c

Figure 10: The smoothed data for queries about R using the smoothing interval w = 9 (a), w = 11 (b), w = 13 (c) according to Pollard formulas

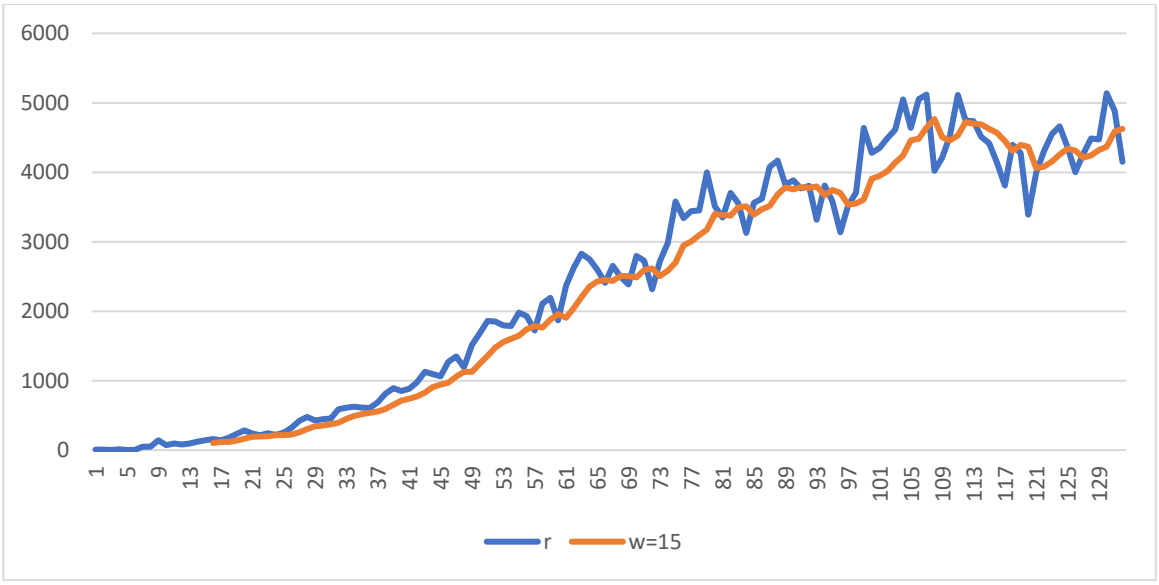
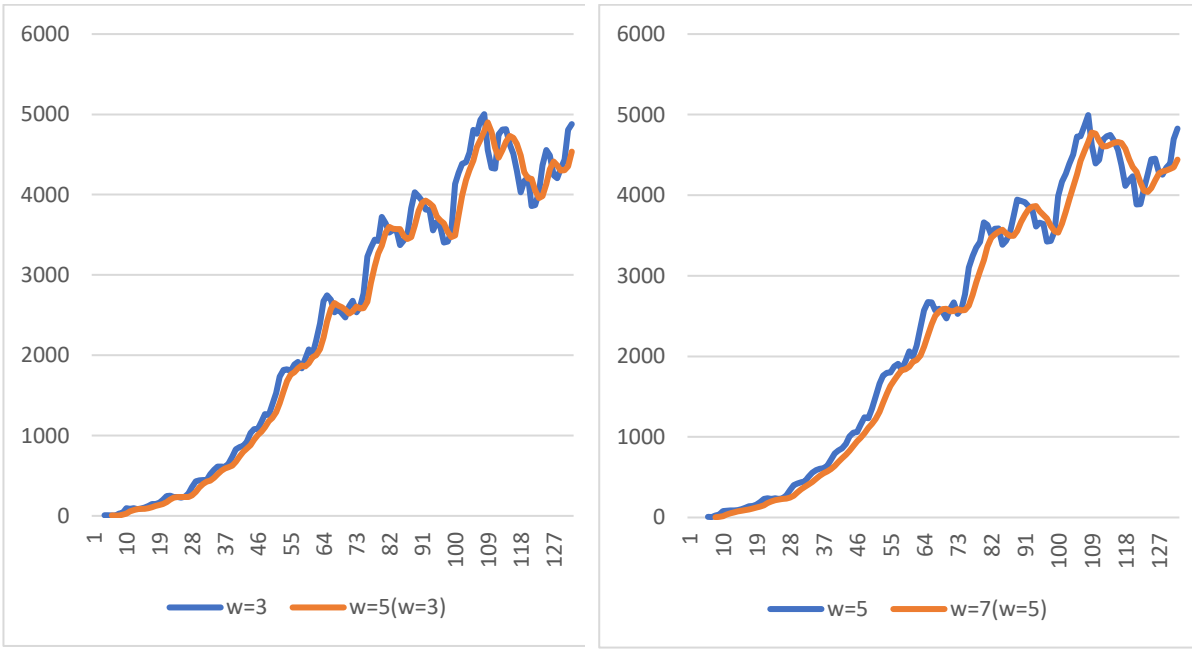


Figure 11: The smoothed data for queries about R using the smoothing interval w = 15 according to Pollard formulas

We smooth the data using the size of the smoothing interval w = 3, then we smooth the obtained smoothed data again, but we use the size of the smoothing interval w = 5.

The smoothed data for queries about R obtained by the smoothing data again. In Fig. 12 there are presented the smoothed data for queries about R using the smoothing interval w = 5 (w = 3) (a), w = 7 (w = 5) (b). Fig. 13 shown the smoothed data for queries about R for w = 9 (w = 7) (a), w = 11 (w = 9) (b), w = 13 (w = 11) (c), w = 15 (w = 13) (d) according to Pollard formulas.



a b

Figure 12: The smoothed data for queries about R using the smoothing interval w = 5 (w = 3) (a), w = 7 (w = 5) (b) according to Pollard formulas

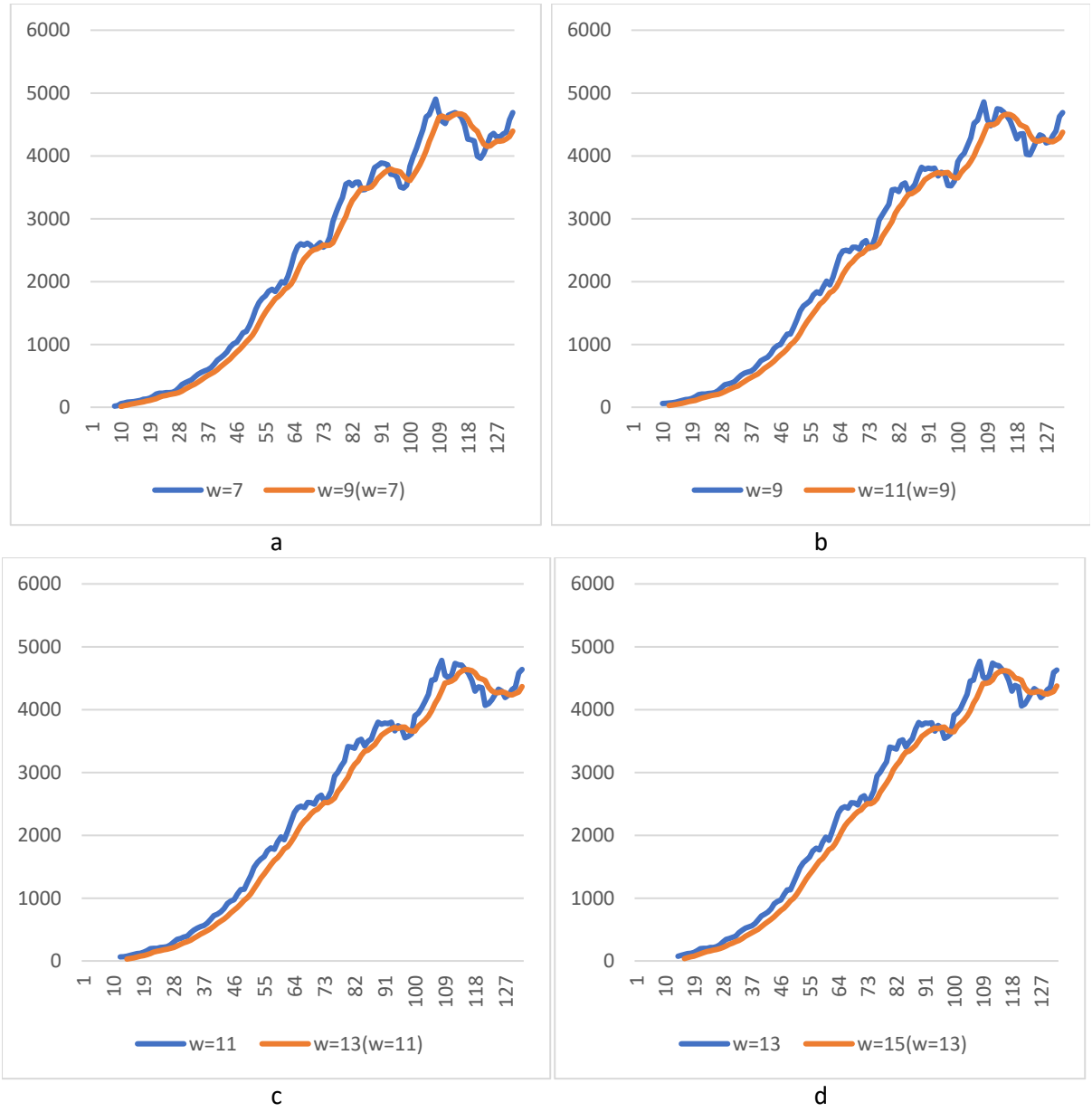


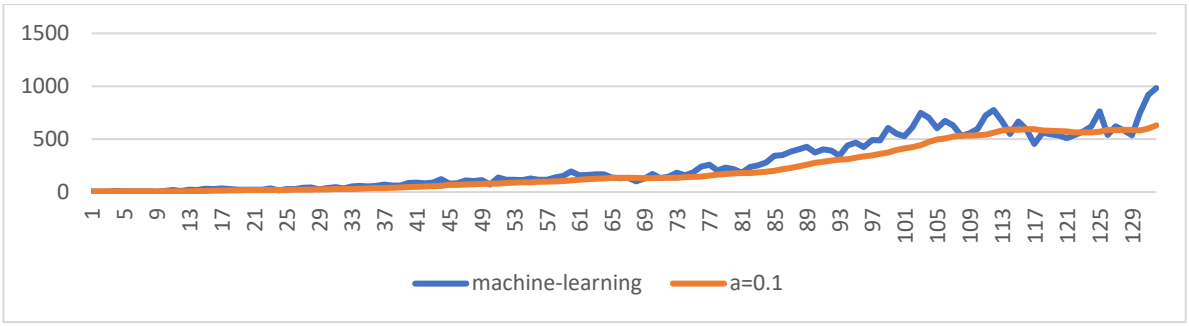
Figure 13: The smoothed data for queries about R using the smoothing interval w = 9 (w = 7) (a), w = 11 (w = 9) (b), w = 13 (w = 11) (c), w = 15 (w = 13) (d) according to Pollard formulas

5.3. Exponential smoothing

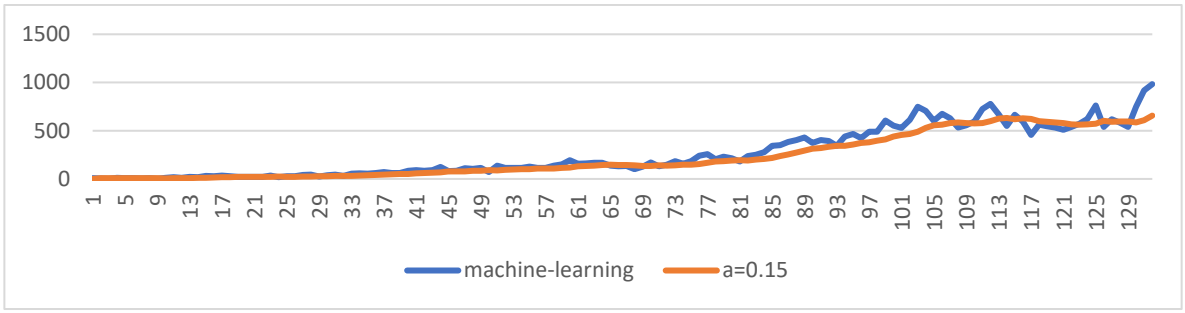
We add all sample elements to construct exponential smoothing and multiply by a factor (1 - α). The α takes values from zero to one, and the last element of the already created table of values for a certain α is multiplied by α (the Sum of coefficients should be equal to 1). The following is a graph of exponential smoothing for all required α.

Exponential smoothing queries about Machine Learning for a=0.1 (a), a=0.15 (b), a=0.2 (c), a=0.25 (d), a=0.3 (e) are presented in the Fig. 14.

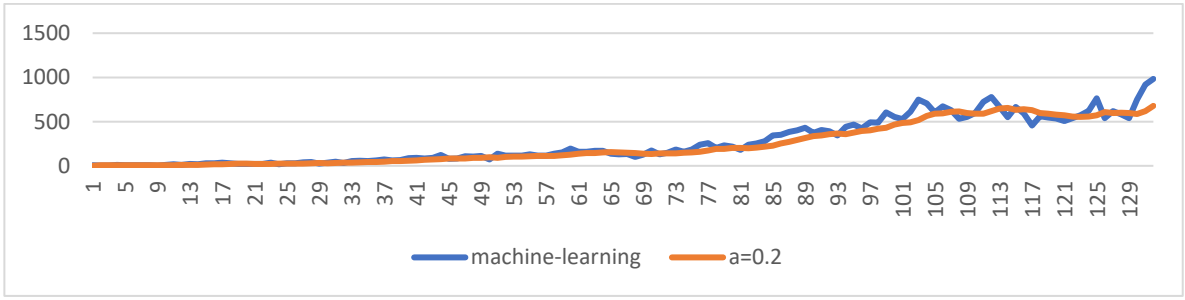
We find the number of turning points and coefficients for each smoothing correlation between original and smoothed values. The correlation coefficients between the original values and the smoothed ones are calculated in the Table 4.



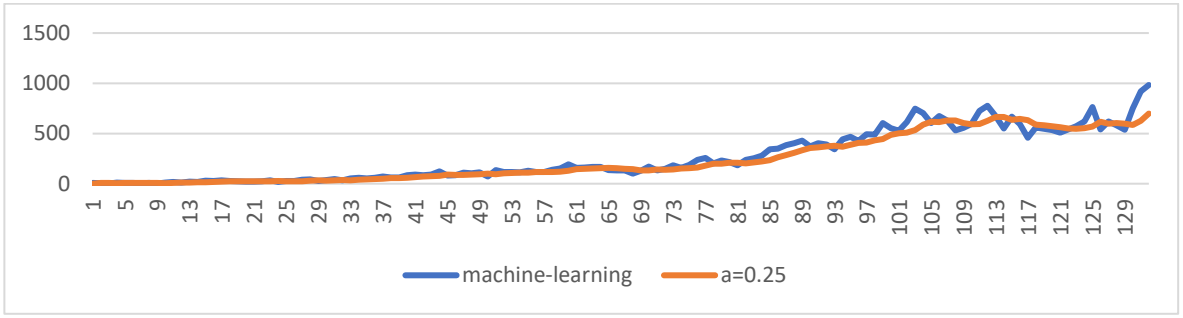
a



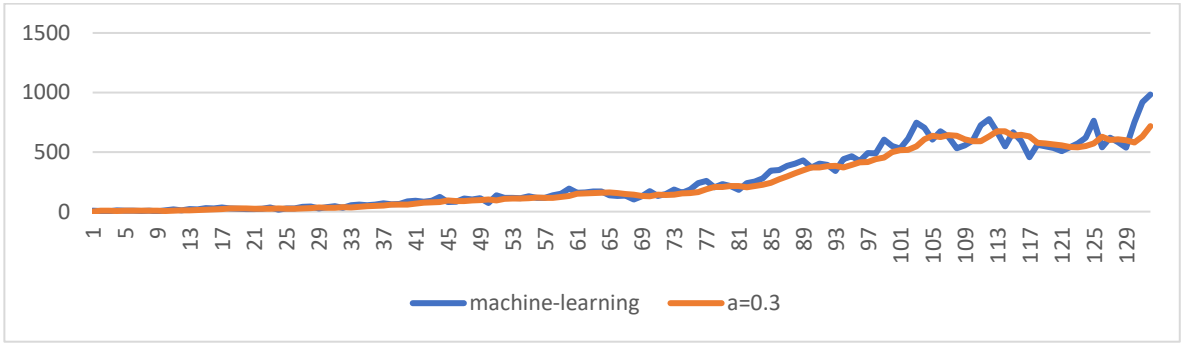
b



c



d



e

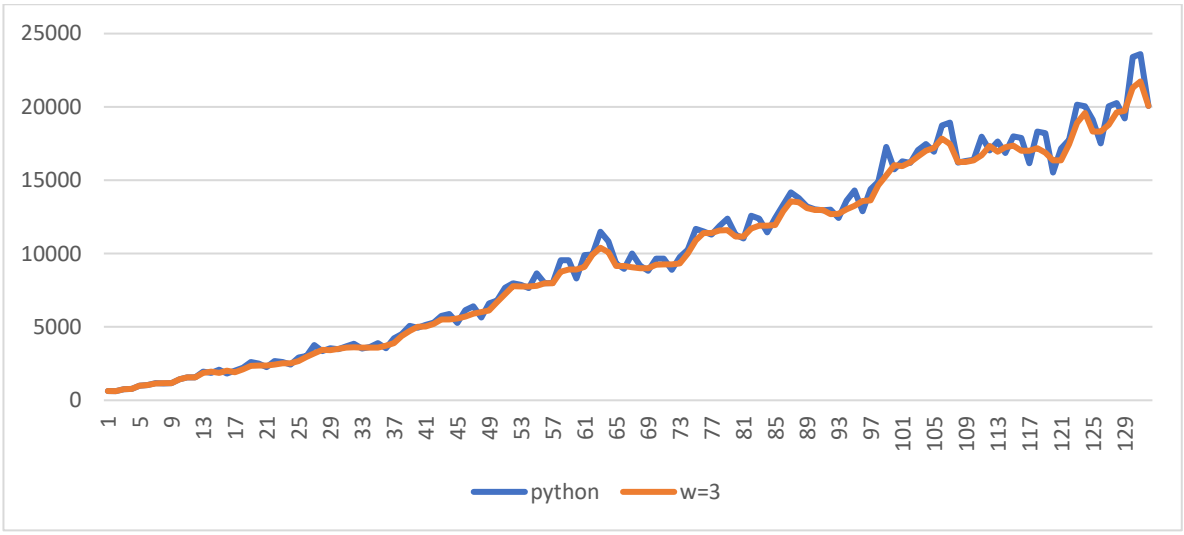
Figure 14: Exponential smoothing queries about Machine Learning for a=0.1 (a), a=0.15 (b), a=0.2 (c), a=0.25 (d), a=0.3 (e)

Table 4
The correlation coefficients between the original values and the smoothed ones

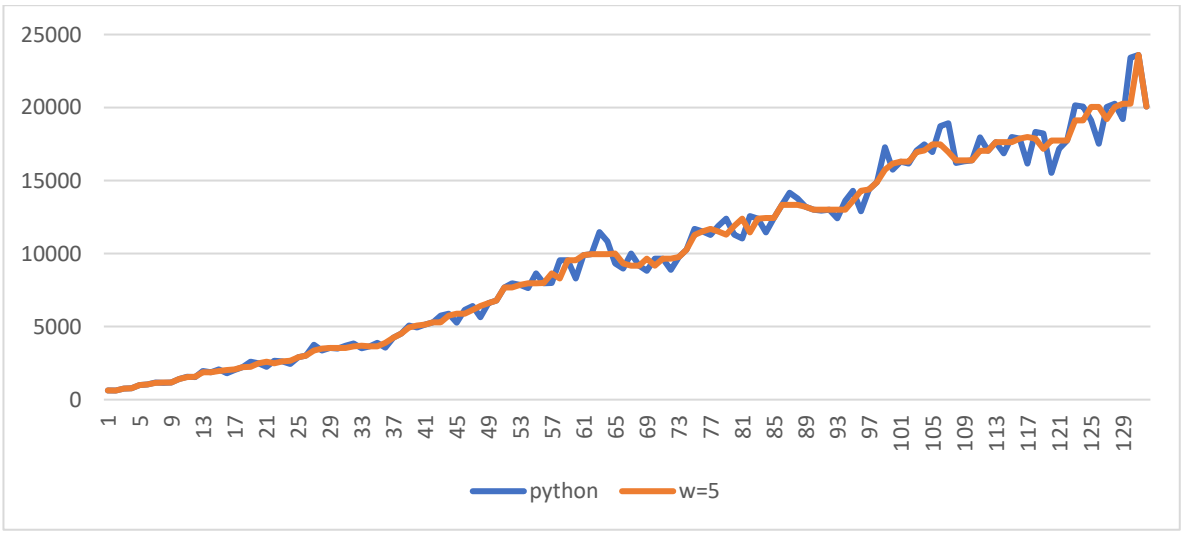
Factor α	0.1	0.15	0.2	0.25	0.3
Correlation coefficient	0,958867	0,964152	0,96739	0,969568	0,971129
Number of correct turning points	26	32	38	38	42

5.4. Median smoothing

Median smoothing queries about Python for w=3 (a), w=5 (b), w=7 (a), w=9 (a), w=11 (a), w=13 (a), w=15 (a) are presented in the Fig. 15-Fig. 17.

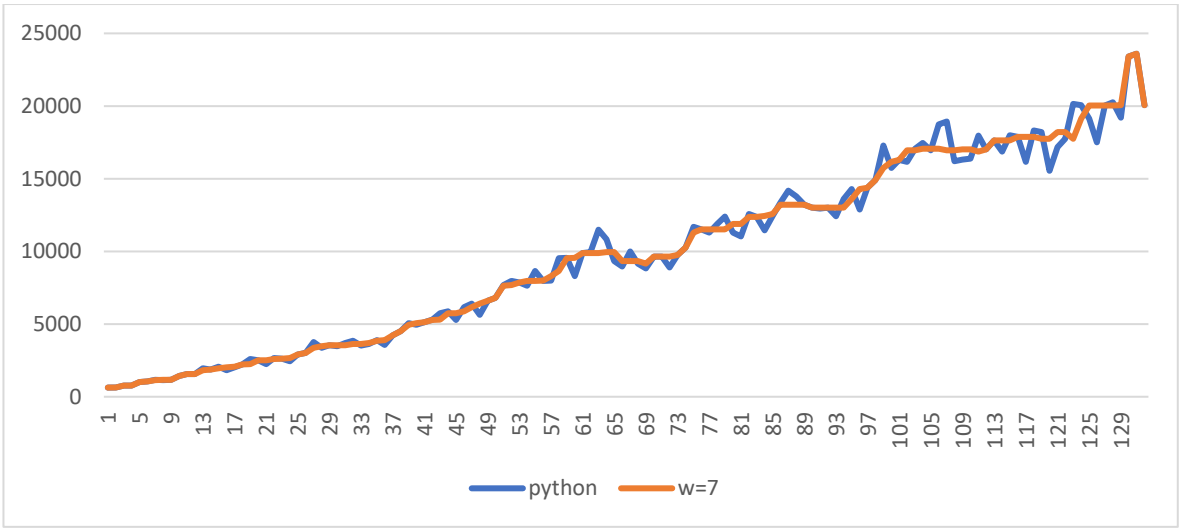


a

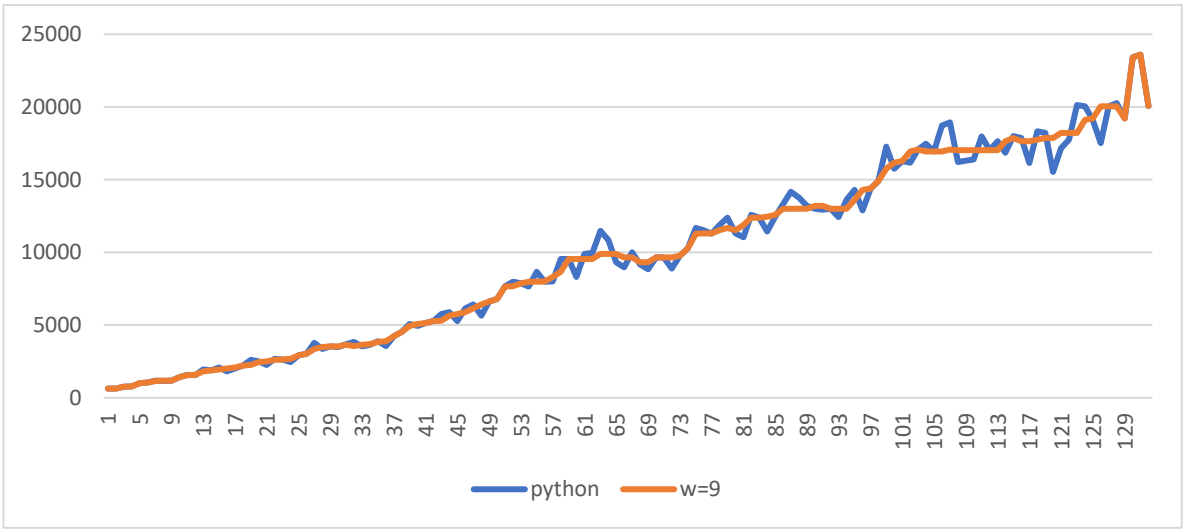


b

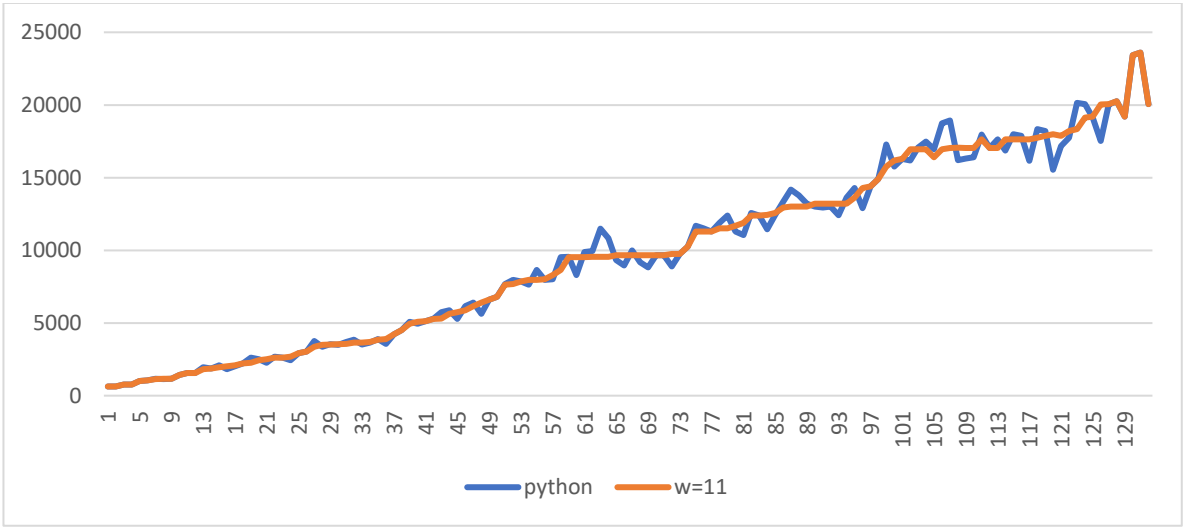
Figure 15: Median smoothing queries about Python for w=3 (a), w=5(b)



a

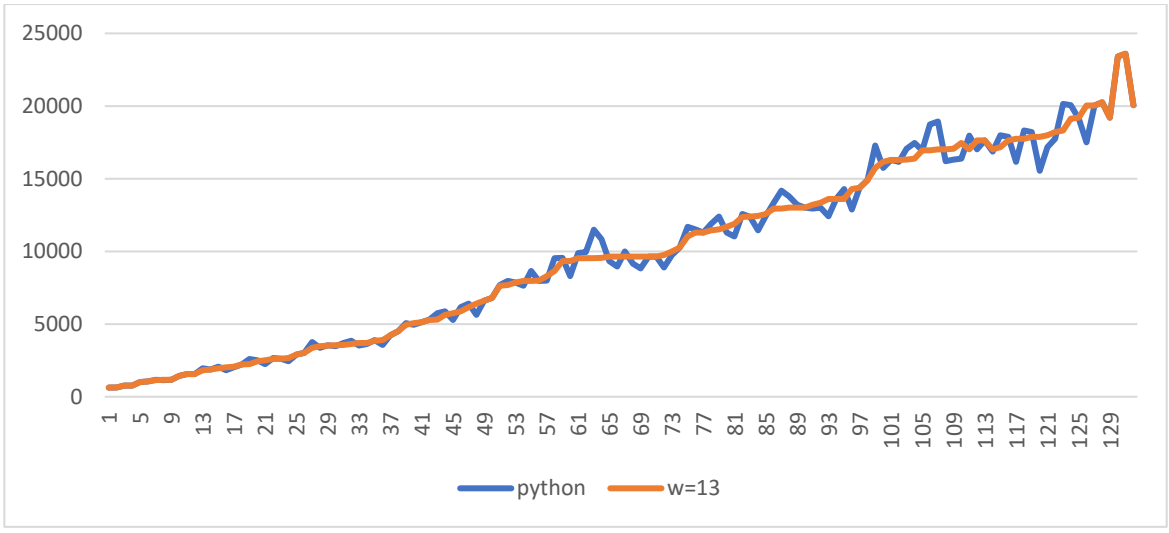


b

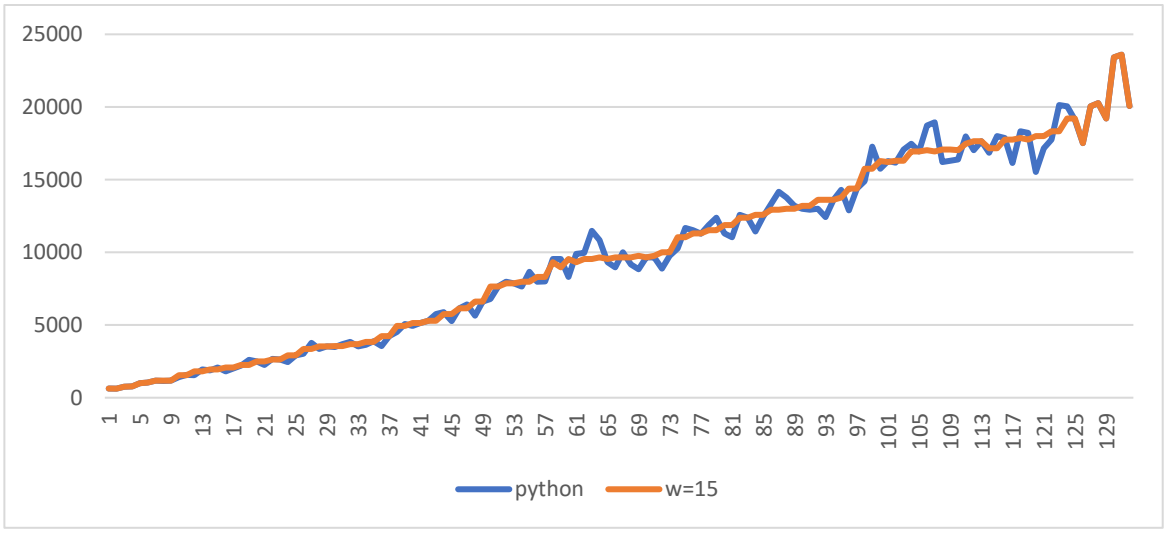


c

Figure 16: Median smoothing queries about Python for w=7 (a), w=9 (b), w=11 (c)



a



b

Figure 17: Median smoothing queries about Python for w=13 (a), w=15 (b)

6. Discussions

6.1. Data correlation

Correlation analysis is a group of methods that can detect the presence and degree of relationship between several parameters that change randomly [24]. Two samples (data sets) are studied in the simplest case. Their multidimensional complexes (groups) are studied in the general case. The purpose of correlation analysis is to determine whether one variable has a significant dependence on another [25]. The main tasks of correlation analysis are the definition and expression of the form of analytical dependence of the resultant trait y on the factor traits x_i.

There are the following stages of correlation analysis [24, 25].

- Identifying the relationship between the signs;
- Determining the form of communication;
- Determination of strength (tightness) and direction of communication.

Advantages of correlation analysis are as following.

- Ability to create a new rule of the interaction of functions with each other;
- Estimation of the interaction of functions received strangely.

Disadvantages are as following. The results obtained using the technique can be used only in the field of this study or close to it.

A correlation occurs when a series of values of a function (dependent variable) corresponds to the same value of an argument (independent variable) [24].

To construct a correlation field, we considered the definition of the concept of correlation field. The correlation field (scatter plot) is a graphical representation of the relationship between the two studied sequences [24, 25]. Thus, it is a set of points in a rectangular coordinate system, the abscissa of each of which corresponds to the value of the factor feature (x), and the ordinate - the value of the resultant feature (y) of a particular unit of observation. The number of points on the Graph corresponds to the number of observation units. The location of points on the correlation field allows you to judge the nature of the dependence, for example, linear, parabolic, hyperbolic, logistical, logarithmic, exponential, exponential, or no dependence [24].

Fig. 18 shows the behavior of the correlation field for queries about the python programming language for only one month for each day. From Fig. 18 it is seen that the nature of the dependence is linear. The dependence is described by an equation $y = 1932x - 17317$ with a high coefficient of determination $R^2 = 0{,}972$.

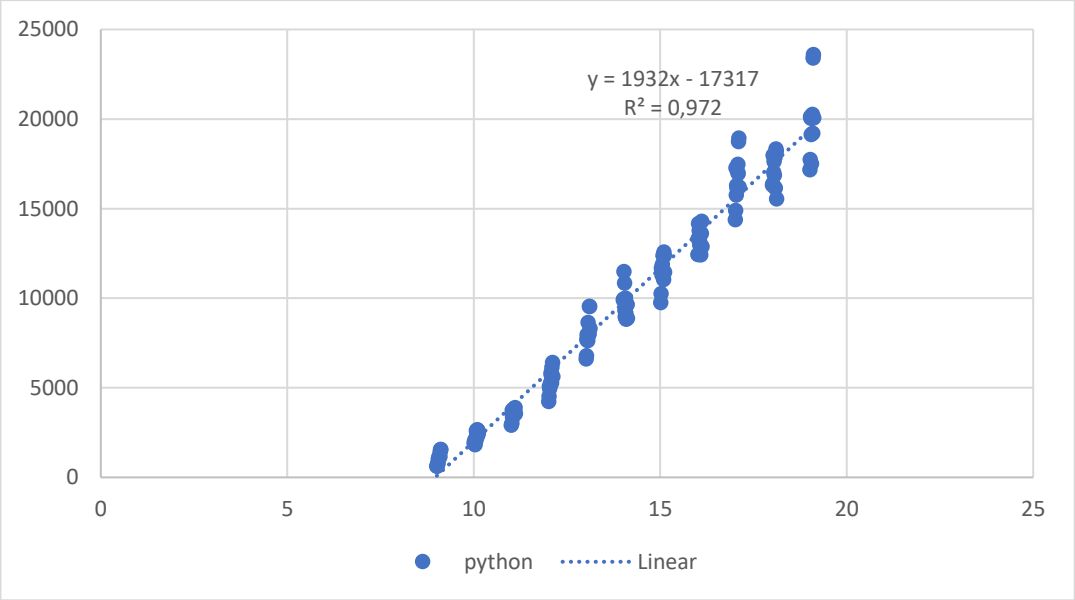


Figure 18: The correlation field for queries about Python to days during one month

The correlation field is built from the input data (x and game) in the form of a scatter plot. Analyzing the location of points on the correlation field, we can judge the nature of the dependence, namely that it is linear. Request dates start from 2009 and are collected until 2019 inclusive, broken down by all months. The lowest number of requests for Python was one month of 2009 and increased with each passing month, indicating the language's growing popularity and increased number of users. Data from 2019 to 2021 are not collected in the network date. However, analyzing the statistics, we can predict even more significant growth in the popularity of the programming language, as there are requests for its library. That is, the data has a growing trend.

We are determining the value of the correlation coefficient. A sample correlation coefficient is used to quantify the closeness of the connection. The correlation coefficient characterizes the degree of closeness of the linear dependence. In general, when some stochastic dependence relates the X and Y values, the correlation coefficient may have a value in the range of $-1 \le r \le +1$ [24].

The formula for calculating the correlation coefficient is as following.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i-1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i-1}^{n} (y_i - \bar{y})^2}}. \quad (3)$$

The statistical scientific sources [24-29] are recommended to use the following expression to calculate the correlation coefficient.

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}. \quad (4)$$

The calculated correlation coefficient for queries about Python is equal to R^2=0,98588536. It is a good correlation coefficient, and it shows that there is a dependence; it is linear and quite close.

The correlation ratio is used in cases where there are following case [24-29].

- Between a pair of studied features, there is a nonlinear relationship;
- The nature of the sample data (number, density of location on the correlation field) allows their grouping on the y-axis, and secondly, the ability to calculate "individual" mathematical expectations within each grouping interval.

According to the preliminary construction of the correlation field, we see that the Graph is linear, so it is impractical to calculate the correlation ratio.

To divide one of the sequences into three equal parts we divide the sequence, corresponding to the number of queries about the python in the programming language library (Table 5).

Table 5

The divided sequence of the queries about python in the programming language libraries into three equal intervals

	1st part	2nd part	3rd part
Interval	(1; 45)	(45; 89)	[89; 132]
Number of sample items	44	44	44

As we can see the partition is performed so that the number of sample elements at each interval is the same, and it is equal 44. In the case of many observations, when the correlation coefficients need to be calculated sequentially for several samples, for convenience, the obtained coefficients are summarized in tables, which are called correlation matrices.

The correlation matrix is a square table where the correlation coefficient between the corresponding parameters is located at the corresponding row and column intersection [24-29].

Dividing the sample into three equal parts, we build a correlation matrix (Table 6).

Table 6

The correlation matrix of the queries about python

	1st part	2nd part	3rd part
1st	1		
2nd	0,92230619	1	
3rd	0,8602376	0,86988678	1

The formula for calculating the autocorrelation coefficient is as following [24-29].

$$r(\tau) = \frac{(n-\tau)\sum_{t=1}^{n-\tau} y_t y_{t+\tau} - \sum_{t=1}^{n-\tau} y_t \sum_{t=1}^{n-\tau} y_{t+\tau}}{\sqrt{[(n-\tau)\sum_{t=1}^{n-\tau} y_t^2 - (\sum_{t=1}^{n-\tau} y_t)^2][(n-\tau)\sum_{t=1}^{n-\tau} y_{t+\tau}^2 - (\sum_{t=1}^{n-\tau} y_{t+\tau})^2]}} \quad (5)$$

To calculate the autocorrelation coefficient according to the formula (5). We used the CORREL function. The autocorrelation coefficient for queries about Python is presented in Table 7. The sequence of autocorrelation coefficients of the levels of the first, second, third, etc. orders is called the autocorrelation function. The Graph of the autocorrelation function is called the correlogram [25-29].

Table 7
Multiple correlation coefficients for the queries about python

Lag	Autocorrelation coefficients
1	0,98701089
2	0,98096642
3	0,98308662
4	0,9783225
5	0,9783225
6	0,97742187
7	0,97709153

The correlogram for the queries about python is presented in Fig. 19. The pattern of the correlogram shows that the studied series is not stationary because in the case of a stationary time series, the correlogram must decline rapidly.

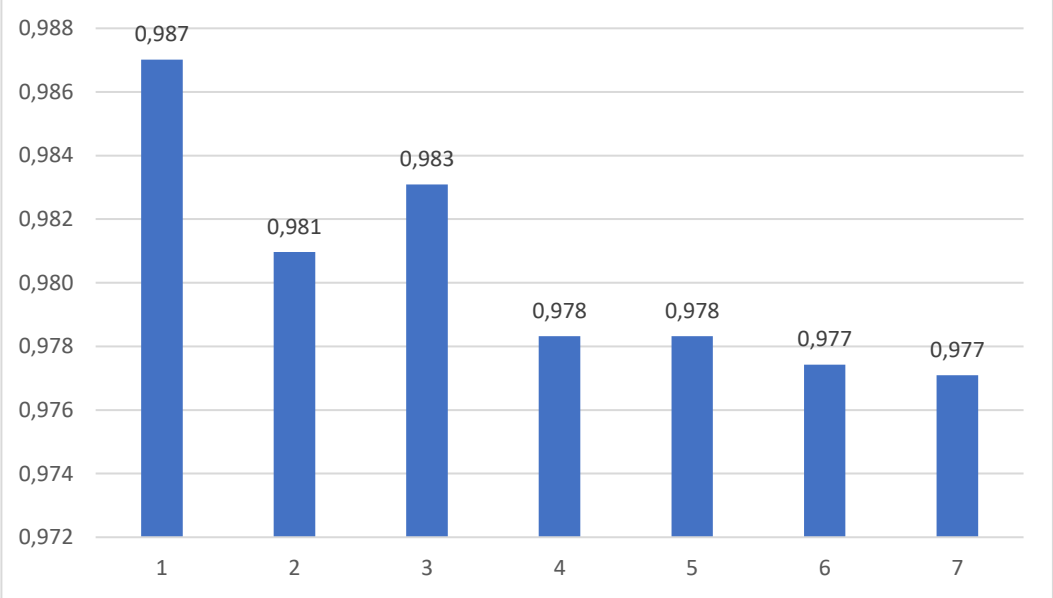


Figure 19: The correlogram for queries about Python vs each lag

6.2. The cluster data analysis

To form an "object-property" table from our data, let's split the data so that the 2nd, 3rd, 4th, and 5th columns can be considered objects. The first column will then be considered a property. To calculate each of the properties, we use the standard formulas [53-62]. To calculate the properties in column 2016, we used only the data for queries collected from this 2016 (Table 8). The term "average" in the Table 8 means the average number of queries in the NumPy library overall 12 months. Accordingly, "minimum" shows the lowest number of requests during the year (for a month), and "maximum" - the most. "Volume" - the number of lines for a given year. There are 12 of them every year, because of 12 months a year. "Fashion" is the value of a certain quantity, which occurs most often in all observations. Since the statistics on queries changed every month and there was never one repeated for at least two months, cannot talk about fashion, it is impossible to determine. "Median" is a number that divides the list of attribute values into two equal parts so that there is the same number of units on both sides. "Standard error" is the approximate standard deviation of the statistical sample. The more data points involved in calculating the mean, the smaller the standard error [63-79]. "Standard deviation" is the deviation of all characteristic values from their average value.

Table 8
The Normalized table "object-property"

Index	2016	2017	2018	2019
Average	13259.25	16678.92	17191.67	19861.33
Standard error	0.037773	0.026568	0.037599	0.027778
Median	13108	16620	17334.5	20047.5
Fashion	# N / A	# N / A	# N / A	# N / A
Standard deviation	580.6382	1304,687	894.8652	1939,741
Sampling variance	367789.8	1856955	873582.2	4104650
Kurtosis	-0.63691	-0.25216	-1.25289	0.25504
Asymmetry	0.396782	0.09637	-0.39425	0.7177
Interval	328,5209	738.1827	506.3083	1097,492
Minimum	12424	14388	15537	17167
Maximum	14296	18935	18329	23602
Sum	159111	200147	206300	238336
Amount	12	12	12	12
Reliability level (95%)	0.083137	0.058476	0.082755	0.061138

It is one of the essential methods to help determine how much a particular value change [74-79]. The larger the standard deviation, the more comprehensive the range of changes in the values of this value "Amount" - the total number of requests to the library for twelve months for each described year. The "level of reliability" is the ability to reject the null hypothesis when it is correct. It is a good possibility of error of the first kind for this task. "Sampling variance" - allows you to measure how far random values are distributed from their average value. Larger variance values indicate more significant deviations of the values of the random variable from the center of the distribution. "Excess" is a numerical characteristic of the probability distribution of an objective random variable. The excess coefficient characterizes the "steepness," i.e., the rate of increase of the distribution curve compared to the standard curve. "Asymmetry" measures how asymmetric the distribution (skew) can be. If we talk about the opposite concept of symmetry, the distribution relative to the center on the right and left is ideal mirror images of each other. "Interval" - the interval between the extreme values of the feature in the group of units. To construct a matrix of similarities (Table 9) we used formula (6) by analogy with the previous Table 8 [53-73].

$$d_E = \sqrt{\sum_{p=1}^{q} (x_{ip} - x_{jp})^2} \quad (6)$$

Table 9
The proximity matrix for four clusters

Cluster	1	2	3	4
1	0	1489747	508047.4	3737727
2	1489747	0	983393.4	2248031
3	508047.4	983393.4	0	3231234
4	3737727	2248031	3231234	0

The resulting proximity matrix (Table 9) is a symmetric diagonal matrix that indicates the amount of proximity between objects. Agglomerative hierarchical cluster analysis is performed based on such a matrix. The choice of integration strategy is determined by the approach. We chose the strategy of the nearest neighbor. In it, the distance between two groups is defined as the distance between the two closest elements of these groups.

After performing the cluster analysis procedure sequentially, we obtained proximity matrices for 3 (Table 10) and 2 clusters (Table 11).

Table 10
The proximity matrix for 3 clusters

Cluster	1.3	2	4
1.3	0	983393,4	3231234
2	983393,4	0	2248031
4	3231234	2248031	0

Table 11
The proximity matrix for 2 clusters

Cluster	1.3.2	4
1.3.2	0	2248031
4	2248031	0

The cluster analysis procedure starts with the proximity matrix. In it, we determine the smallest number. It is 508047.4, located at the 1st and 3rd objects intersection. Therefore, we group the 1st and 3rd objects and create a new table. Now determine the minimum number again. This time it is at the intersection of objects (1.3) and (2). We are grouping them again. We built a table "union-node-metric" (Table 12).

Table 12
The union-node-metric table for programming language libraries

Step	Association	Node	Metrics
1	1 + 3	d5	508047.4
2	1 + 3 + 2	d6	983393.4
3	1 + 3 + 2 + 4	d7	2248031

Our union-node-metric table is formed in 3 steps. In the first, there is a union of objects 1 and 3. In the second step of objects (1,3) and 2. In the third (1,3,2) and 4. According to the steps, nodes are formed, named d 5, d 6, and d 7, because there are four objects, and the next numbering begins after the 4th. And the representation of the metric is the minimum value at each stage of the construction of the table.

The constructed dendrogram for programming language libraries can help us to visualize the results of cluster analysis in the Fig.20. We construct the dendrogram of clustering several objects manually in the draft version and then implement it in a graphical environment. Indicators on the dendrogram on the left represent the metric, the bottom objects, and the top point to each node separately. Drawing horizontal lines in the plane of the dendrogram at a given height, in this case, allows you to select individual clusters.

When interpreting the results of cluster analysis, we observe 3 clusters at level 2248031, among which cluster 1 includes objects 1.3, the second cluster - only one object 2, and the third - only object 4. At level 983393,4 we observe 2 clusters, among which the first cluster includes three objects 1,3,2, and the second - only one object 4. At the level of 508047,4 we observe one cluster of all elements.

7. Conclusions

In this work, we learned the basic visualization methods, graphical display, and primary statistical processing of numerical data represented by a sample of time series.

We got acquainted with the main methods of highlighting the trend of the behavior of the studied indicator, which is represented by the nature of its trend, using methods of smoothing time series and presenting the results using an MS Excel spreadsheet.

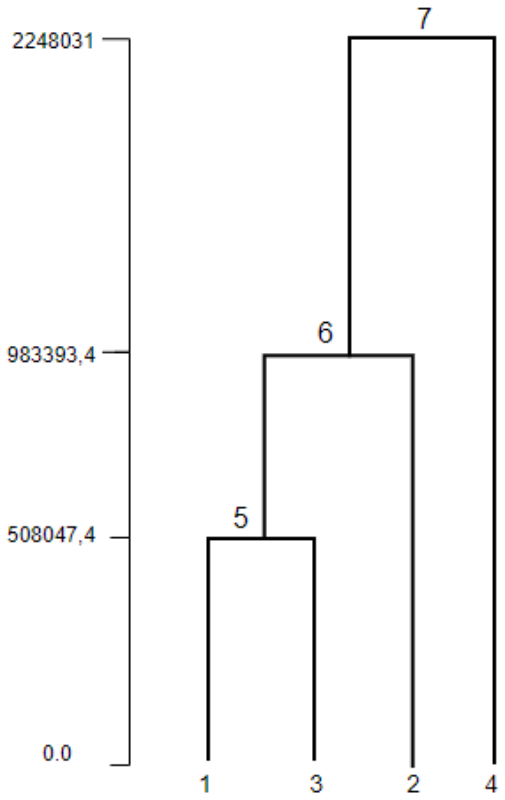


Figure 20: The constructed dendrogram of programming language libraries

We also got acquainted with the methods of correlation analysis of experimental data presented by time sequences. We learned to build a correlation field, determine the value of the correlation coefficient, calculate the correlation ratio, plot autocorrelation functions, divide one of the sequences into three equal parts, build a correlation matrix for them and find multiple correlation coefficients. We also divided a given set of objects, each characterized by the same set of specific features, into separate groups using hierarchical agglomerative cluster analysis.

A library rating system has been created, i.e., the most significant number of queries has been identified, and the most popular language has been identified. In ranking queries in language libraries, where the first is Python, the least popular - is spacy. The tendency of the growing popularity of all language libraries characterizes the active development of programming and, most importantly, people's interest in the work. The obtained data will allow experts to assess the decline, growth, and invariability of the popularity of languages in the recent period (2009-2019) and offer their vision of the possible development of specific programming languages.

8. References

[1] O. Kuzmin, M. Bublyk, A. Shakhno, O. Korolenko, H. Lashkun, Innovative development of human capital in the conditions of globalization, E3S Web of Conferences 166 (2020) 13011.

[2] I. Bodnar, M. Bublyk, O. Veres, O. Lozynska, I. Karpov, Y. Burov, P. Kravets, I. Peleshchak, O. Vovk, O. Maslak, Forecasting the risk of cervical cancer in women in the human capital development context using machine learning, CEUR workshop proceedings Vol-2631 (2020) 491-501.

[3] M. Bublyk, V. Vysotska, Y. Matseliukh, V. Mayik, M. Nashkerska, Assessing losses of human capital due to man-made pollution caused by emergencies, CEUR Workshop Proceedings Vol-2805 (2020) 74-86.

[4] D. Koshtura, M. Bublyk, Y. Matseliukh, D. Dosyn, L. Chyrun, O. Lozynska, I. Karpov, I. Peleshchak, M. Maslak, O. Sachenko, Analysis of the demand for bicycle use in a smart city based on machine learning, CEUR workshop proceedings Vol-2631 (2020) 172-183.

[5] M. Bublyk, Y. Matseliukh, U. Motorniuk, M. Terebukh, Intelligent system of passenger transportation by autopiloted electric buses in Smart City, CEUR workshop proceedings Vol-2604 (2020) 1280-1294.
[6] I. Rishnyak, O. Veres, V. Lytvyn, M. Bublyk, I. Karpov, V. Vysotska, V. Panasyuk, Implementation models application for IT project risk management, CEUR Workshop Proceedings Vol-2805 (2020) 102-117.
[7] V. Vysotska, A. Berko, M. Bublyk, L. Chyrun, A. Vysotsky, K. Doroshkevych, Methods and tools for web resources processing in e-commercial content systems, in: Proceedings of 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 1, 2020, pp. 114-118. doi: 10.1109/CSIT49958.2020.9321950.
[8] M. Bublyk, A. Kowalska-Styczen, V. Lytvyn, V. Vysotska, The Ukrainian Economy Transformation into the Circular Based on Fuzzy-Logic Cluster Analysis, Energies 2021 (14) 5951. doi: 10.3390/en14185951.
[9] A. Berko, I. Pelekh, L. Chyrun, M. Bublyk, I. Bobyk, Y. Matseliukh, L. Chyrun, Application of ontologies and meta-models for dynamic integration of weakly structured data, in: Proceedings of the IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP, 2020, pp. 432-437. doi: 10.1109/DSMP47368.2020.9204321.
[10] V.-A. Oliinyk, V. Vysotska, Y. Burov, K. Mykich, V. Basto-Fernandes, Propaganda Detection in Text Data Based on NLP and Machine Learning, CEUR workshop proceedings Vol-2631 (2020) 132-144.
[11] R. Lynnyk, V. Vysotska, Y. Matseliukh, Y. Burov, L. Demkiv, A. Zaverbnyj, A. Sachenko, I. Shylinska, I. Yevseyeva, O. Bihun, DDOS Attacks Analysis Based on Machine Learning in Challenges of Global Changes, CEUR workshop proceedings Vol-2631 (2020) 159-171.
[12] V. Vysotska, Linguistic Analysis of Textual Commercial Content for Information Resources Processing, in: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 2016, pp. 709-713. doi: 10.1109/TCSET.2016.7452160.
[13] V. Lytvyn, V. Vysotska, A. Rzheuskyi, Technology for the Psychological Portraits Formation of Social Networks Users for the IT Specialists Recruitment Based on Big Five, NLP and Big Data Analysis, CEUR Workshop Proceedings Vol-2392 (2019) 147-171.
[14] Lytvyn Vasyl, Vysotska Victoria, Dosyn Dmytro, Holoschuk Roman, Rybchak Zoriana, Application of Sentence Parsing for Determining Keywords in Ukrainian Texts, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2017, pp. 326-331. doi: 10.1109/STC-CSIT.2017.8098797.
[15] Y. Burov, V. Vysotska, P. Kravets, Ontological approach to plot analysis and modeling, CEUR Workshop Proceedings Vol-2362 (2019) 22-31.
[16] V. Vysotska, O. Kanishcheva, Y. Hlavcheva, Authorship Identification of the Scientific Text in Ukrainian with Using the Lingvometry Methods, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 34-38. doi: 10.1109/STC-CSIT.2018.8526735.
[17] A. Gozhyj, I. Kalinina, V. Gozhyj, V. Vysotska, Web service interaction modeling with colored petri nets, in: Proceedings of the International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS, 1, 2019, pp. 319-323. doi: 10.1109/IDAACS.2019.8924400.
[18] A. Gozhyj, I. Kalinina, V. Vysotska, S. Sachenko, R. Kovalchuk, Qualitative and Quantitative Characteristics Analysis for Information Security Risk Assessment in E-Commerce Systems, CEUR Workshop Proceedings Vol-2762 (2020) 177-190.
[19] L. Podlesna, M. Bublyk, I. Grybyk, Y. Matseliukh, Y. Burov, P. Kravets, O. Lozynska, I. Karpov, I. Peleshchak, R. Peleshchak, Optimization model of the buses number on the route based on queueing theory in a smart city, CEUR workshop proceedings Vol-2631 (2020) 502 - 515.
[20] O. Bisikalo, O. Kovtun, V. Kovtun, V. Vysotska, Research of Pareto-Optimal Schemes of Control of Availability of the Information System for Critical Use, CEUR Workshop Proceedings Vol-2623 (2020) 174-193.
[21] V. Vysotska, Ukrainian Participles Formation by the Generative Grammars Use, CEUR workshop proceedings Vol-2604 (2020) 407-427.

[22] V. Vysotska, S. Holoshchuk, R. Holoshchuk, A comparative analysis for English and Ukrainian texts processing based on semantics and syntax approach, CEUR Workshop Proceedings Vol-2870 (2021) 311-356.
[23] K. Tymoshenko, V. Vysotska, O. Kovtun, R. Holoshchuk, S. Holoshchuk, Real-time Ukrainian text recognition and voicing, CEUR Workshop Proceedings Vol-2870 (2021) 357-387.
[24] Data Set, 2022. URL: https://www.kaggle.com/aishu200023/stackindex.
[25] M. Bublyk, Y. Matseliukh, Small-batteries utilization analysis based on mathematical statistics methods in challenges of circular economy, CEUR workshop proceedings Vol-2870 (2021) 1594-1603.
[26] Standard error, 2022. URL: https://ua.nesrakonk.ru/standard-error/.
[27] Standard deviation, 2022. URL: https://studopedia.su/10_11382_standartne-vidhilennya.html.
[28] Statistical models of marketing decisions taking into account the uncertainty factor, 2022. URL: https://excel2.ru/articles/uroven-znachimosti-i-uroven-nadezhnosti-v-ms-excel.
[29] Grouping of statistical data - BukLib.net Library, 2022. URL: https://buklib.net/books/35946/.
[30] Stack Overflow, 2022. URL: https://en.wikipedia.org/wiki/Stack_Overflow.
[31] StackOverflow is more than just a repository of answers to stupid questions, 2022. URL: https://habr.com/ru/post/482232/.
[32] TechTrend, 2022. URL: http://techtrend.com.ua/index.php?newsid=20844.
[33] Graphic presentation of information, 2022. URL: https://studopedia.com.ua/1_132145_grafichne-podannya-informatsii.html.
[34] Construction of an interval variable sequence of continuous quantitative data, 2022. URL: https://stud.com.ua/93314/statistika/pobudova_intervalnogo_variatsiynogo_ryadu_bezperernih_kilkisnih_danih.
[35] Forecasting the trend of the time series by algorithmic methods, 2022. URL: http://ubooks.com.ua/books/000269/inx42.php.
[36] Wikideck, 2022. URL: https://wp-uk.wikideck.com/.
[37] StackOverflow, 2022. URL: https://ru.stackoverflow.com.
[38] P. Bidyuk, A. Gozhyj, I. Kalinina, V. Vysotska, Methods for Forecasting Nonlinear Non-Stationary Processes in Machine Learning, Communications in Computer and Information Science 1158 (2020) 470-485. doi: 10.1007/978-3-030-61656-4_32.
[39] P. Bidyuk, A. Gozhyj, I. Kalinina, V. Vysotska, M. Vasilev, R. Malets, Forecasting Nonlinear Nonstationary Processes in Machine Learning Task, in: Proceedings of the IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP, 2020, pp. 28-32. doi: 10.1109/DSMP47368.2020.9204077.
[40] A. B. Lozynskyy, I. M. Romanyshyn, B. P. Rusyn, Intensity Estimation of Noise-Like Signal in Presence of Uncorrelated Pulse Interferences, Radioelectronics and Communications Systems 62(5) (2019) 214-222. doi: 10.3103/S0735272719050030.
[41] N. Romanyshyn, Algorithm for Disclosing Artistic Concepts in the Correlation of Explicitness and Implicitness of Their Textual Manifestation, CEUR Workshop Proceedings Vol-2870 (2021) 719-730.
[42] O. Rudenko, O. Bezsonov, Robust Training of ADALINA Based on the Criterion of the Maximum Correntropy in the Presence of Outliers and Correlated Noise, CEUR Workshop Proceedings Vol-2870 (2021) 1694-1705.
[43] Y. Yusyn, T. Zabolotnia, Methods of Acceleration of Term Correlation Matrix Calculation in the Island Text Clustering Method, CEUR workshop proceedings Vol-2604 (2020) 140-150.
[44] B. Rusyn, V. Ostap, O. Ostap, A correlation method for fingerprint image recognition using spectral features, in: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET 2002, 2002, pp. 219–220. doi: 10.1109/TCSET.2002.1015935.
[45] A. Lozynskyy, I. Romanyshyn, B. Rusyn, V. Minialo, Robust Approach to Estimation of the Intensity of Noisy Signal with Additive Uncorrelated Impulse Interference. In: Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018, 2018, pp. 251–254. doi: 10.1109/DSMP.2018.8478625.
[46] N. Boyko, O. Moroz, Comparative Analysis of Regression Regularization Methods for Life Expectancy Prediction, CEUR Workshop Proceedings Vol-2917 (2021) 310-326.

[47] L. Mochurad, Optimization of Regression Analysis by Conducting Parallel Calculations, CEUR Workshop Proceedings Vol-2870 (2021) 982-996.
[48] R. Yurynets, Z. Yurynets, D. Dosyn, Y. Kis, Risk Assessment Technology of Crediting with the Use of Logistic Regression Model, CEUR Workshop Proceedings Vol-2362 (2019) 153-162.
[49] A. Kucher, O. Boyko, K. Ilkanych, A. Fechan, N. Shakhovska, Retrospective analysis by multifactor regression in the evaluation of the results of fine-needle aspiration biopsy of thyroid nodules, CEUR Workshop Proceedings Vol-2753 (2020) 443–447.
[50] O. Murzenko, S. Olszewski, O. Boskin, I. Lurie, N. Savina, M. Voronenko, V. Lytvynenko, Application of a combined approach for predicting a peptide-protein binding affinity using regulatory regression methods with advance reduction of features, in: Proceedings of the 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS , 2019, 1, pp. 431–435, 8924244. doi: 10.1109/IDAACS.2019.8924244.
[51] B. van Stein, H. Wang, W. Kowalczyk, T. Bäck, M. Emmerich, Optimally weighted cluster kriging for big data regression, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9385 (2015) 310–321. doi: 10.1007/978-3-319-24465-5_27.
[52] C. L. M. Belusso, S. Sawicki, V. Basto-Fernandes, R. Z. Frantz, F. Roos-Frantz, Price modeling of laaS providers using multiple regression [Modelagem de Preços de Provedores de IaaS Utilizando Regressão Múltipla], in: Iberian Conference on Information Systems and Technologies, CISTI, 2017. 10.23919/CISTI.2017.7975845.
[53] P. Kravets, Y. Burov, V. Lytvyn, V. Vysotska, Gaming method of ontology clusterization, Webology 16(1) (2019) 55-76.
[54] P. Kravets, Y. Burov, O. Oborska, V. Vysotska, L. Dzyubyk, V. Lytvyn, Stochastic Game Model of Data Clustering, CEUR Workshop Proceedings Vol-2853 (2021) 214-227.
[55] I. Lurie, V. Lytvynenko, S. Olszewski, M. Voronenko, A. Kornelyuk, U. Zhunissova, O. Boskin, The Use of Inductive Methods to Identify Subtypes of Glioblastomas in Gene Clustering, CEUR Workshop Proceedings Vol-2631 (2020) 406-418.
[56] Y. Bodyanskiy, A. Shafronenko, I. Klymova, Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach, CEUR Workshop Proceedings Vol-2870 (2021) 6-15.
[57] Y. Meleshko, M. Yakymenko, S. Semenov, A Method of Detecting Bot Networks Based on Graph Clustering in the Recommendation System of Social Network, CEUR Workshop Proceedings Vol-2870 (2021) 1249-1261.
[58] N. Boyko, S. Hetman, I. Kots, Comparison of Clustering Algorithms for Revenue and Cost Analysis, CEUR Workshop Proceedings Vol-2870 (2021) 1866-1877.
[59] R. J. Kosarevych, B. P. Rusyn, V. V. Korniy, T. I. Kerod, Image Segmentation Based on the Evaluation of the Tendency of Image Elements to form Clusters with the Help of Point Field Characteristics, Cybernetics and Systems Analysis 51(5) (2015) 704-713. doi: 10.1007/s10559-015-9762-5.
[60] S. Babichev, B. Durnyak, I. Pikh, V. Senkivskyy, An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms, Advances in Intelligent Systems and Computing 1020 (2020) 532-553. doi:10.1007/978-3-030-26474-1_37.
[61] S. Babichev, M. A. Taif, V. Lytvynenko, V. Osypenko, Criterial analysis of gene expression sequences to create the objective clustering inductive technology, in: Proceedings of the International Conference on Electronics and Nanotechnology, ELNANO, 2017, pp. 244–248. doi: 10.1109/ELNANO.2017.7939756.
[62] S. Babichev, V. Lytvynenko, V. Osypenko, Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm, in: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2017, 1, pp. 479-484. doi: 10.1109/STC-CSIT.2017.8098832.
[63] S. A. Babichev, A. Gozhyj, A. I. Kornelyuk, V. I. Lytvynenko, Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm, Biopolymers and Cell 33(5) (2017) 379–392. doi: 10.7124/bc.000961.

[64] V. Lytvynenko, I. Lurie, J. Krejci, M. Voronenko, N. Savina, M. A. Taif., Two Step Density-Based Object-Inductive Clustering Algorithm, CEUR Workshop Proceedings Vol-2386 (2019) 117-135.
[65] S. Mashtalir, O. Mikhnova, M. Stolbovyi, Multidimensional Sequence Clustering with Adaptive Iterative Dynamic Time Warping, International Journal of Computing 18(1) (2019) 53-59.
[66] R. Melnyk, R. Tushnytskyy, 4-D pattern structure features by three stages clustering algorithm for image analysis and classification, Pattern Analysis and Applications 16(2) (2013) 201-211. doi: 10.1007/s10044-013-0326-x.
[67] R. Melnyk, R. Tushnytskyy, Circuit board image analysis by clustering, in: Proceeding of the 4th International Conference of Young Scientists on Perspective Technologies and Methods in MEMS Design, MEMSTECH, 2008, pp. 44-45. doi: 10.1109/MEMSTECH.2008.4558732.
[68] N. Shakhovska, V. Yakovyna, N. Kryvinska, An improved software defect prediction algorithm using self-organizing maps combined with hierarchical clustering and data preprocessing, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12391 (2020) 414–424. doi: 10.1007/978-3-030-59003-1_27.
[69] S. Babichev, V. Osypenko, V. Lytvynenko, M. Voronenko, M. Korobchynskyi, Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles, in: Proceeding of the IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO, 2018, pp. 298–304. doi: 10.1109/ELNANO.2018.8477439.
[70] S. Babichev, J. Krejci, J. Bicanek, V. Lytvynenko, Gene expression sequences clustering based on the internal and external clustering quality criteria, Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2017, 1, pp. 91–94. doi: 10.1109/STC-CSIT.2017.8098744.
[71] S. Babichev, V. Lytvynenko, J. Skvor, J. Fiser, Model of the objective clustering inductive technology of gene expression profiles based on SOTA and DBSCAN clustering algorithms, Advances in Intelligent Systems and Computing 689 (2018) 21–39. doi: 10.1007/978-3-319-70581-1_2.
[72] N. Shakhovska, V. Vysotska, L. Chyrun, Features of E-Learning Realization Using Virtual Research Laboratory, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2016, pp. 143–148. doi: 10.1109/STC-CSIT.2016.7589891.
[73] N. Shakhovska, V. Vysotska, L. Chyrun, Intelligent Systems Design of Distance Learning Realization for Modern Youth Promotion and Involvement in Independent Scientific Researches, Advances in Intelligent Systems and Computing 512 (2017) 175-198. doi: 10.1007/978-3-319-45991-2_12.
[74] M. Emmerich, V. Lytvyn, I. Yevseyeva, V. B. Fernandes, D. Dosyn, V. Vysotska, Preface: Modern Machine Learning Technologies and Data Science, CEUR Workshop Proceedings Vol-2386 (2019).
[75] M. Emmerich, V. Lytvyn, V. Vysotska, V. Basto-Fernandes, V. Lytvynenko, Preface: Modern Machine Learning Technologies and Data Science, CEUR Workshop Proceedings Vol-2631 (2020).
[76] M. Emmerich, V. Lytvyn, V. Vysotska, V. B. Fernandes, V. Lytvynenko, Preface: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, CEUR Workshop Proceedings Vol-2917 (2021).
[77] P. S., Malachivskyy, Y. V. Pizyur, V. A. Andrunyk, Chebyshev Approximation by the Sum of the Polynomial and Logarithmic Expression with Hermite Interpolation, Cybernetics and Systems Analysis 54(5), (2018) 765-770. doi: 10.1007/s10559-018-0078-0.
[78] B. van Stein, H. Wang, W. Kowalczyk, M. Emmerich, T. Bäck, Cluster-based Kriging approximation algorithms for complexity reduction, Applied Intelligence 50(3) (2020) 778–791. doi: 10.1007/s10489-019-01549-7.
[79] H. Wang, M. Emmerich, B. Van Stein, T. Back, Time complexity reduction in efficient global optimization using cluster kriging, in: Proceedings of the 2017 Genetic and Evolutionary Computation Conference on GECCO, 2017, pp. 889–896. doi: 10.1145/3071178.3071321.