Intelligent Analysis of Best-Selling Books Statistics on Amazon

Andrii Vasyliuk[1], Yurii Matseliukh[1], Taras Batiuk[1], Mykhailo Luchkevych[1], Iryna Shakleina[2], Halyna Harbuzynska[1], Serhii Kondratiuk[1] and Ksenia Zelenska[1]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Ivan Franko Drohobych State Pedagogical University, I. Franko Street, 24, Drohobych, 82100, Ukraine*

Abstract

A data set of 550 elements was analyzed during the study, containing data on books: title, author, number of reviews, price, and rating. Smoothing methods were used to average local data for further forecasting, in which non-systematic elements replace each other. Clear graphs without sharp peaks were obtained. Based on which the methods of moving average, weighted moving average, exponential smoothing and median filtering, it was found that the books that became bestsellers in 2009-2011 did not have a high response rate. While in the following 2012 to 2019, there was an increase in the number of reviews; the average was 25,000 per book.
Examining the correlation between Reviews and Prices, we found a relationship that indicates that books priced above $ 60 typically have no more than 10,000 reviews. Books priced from $ 40 to $ 60 have an average of no more than 20,000 reviews. Books up to $ 20 have different reviews, but mostly this value does not exceed 40,000, and those books that have more are the exception rather than the rule.
The clustering method was applied to the 12 authors who wrote the bestsellers, and after normalizing the table and building proximity tables, we identified clusters.

Keywords

Cluster analysis, information technologies, intelligent analysis, system analysis, data on books, best-selling books, exponential smoothing, median filtering, data processing

1. Introduction

A system analyst is a specialist whose responsibilities include data collection, processing and visualization. Thus, these data can benefit stakeholders and help test hypotheses or find patterns to help make the right decisions. However, to achieve this goal, the analyst must understand how to work with data, identify important and unimportant, know and apply different visualization methods, and have the skills to work with different means of interpreting data. In today's world, no industry can do without the study of statistics because they can qualitatively track current trends and predict future phenomena or human behaviour [1-5].

This work aimed to use methods of visualization of selected statistics, graphical display to determine trends in the behaviour of various indicators, acquaintance with the methods of correlation analysis, and presentation of the analysis results using an MS Excel spreadsheet.

The objectives of the work are:

- to carry out preliminary statistical processing;
- identify trends in the time series by smoothing methods;
- perform correlation analysis;
- perform cluster analysis.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: andrii.s.vasyliuk@lpnu.ua (A. Vasyliuk); indeed.post@gmail.com (Y. Matseliukh); taras.batiuk.mnsa.2020@lpnu.ua (T. Batiuk); luchkevychmm@gmail.com (M. Luchkevych); ioshakleina@gmail.com (Iryna Shakleina); halyna.harbuzynska.sa.2019@lpnu.ua (H. Harbuzynska); serhii.kondratiuk.sa.2019@lpnu.ua (S. Kondratiuk); ksenia.zelenska.sa.2019@lpnu.ua (K. Zelenska)
ORCID: 0000-0002-3666-7232 (A. Vasyliuk); 0000-0002-1721-7703 (Y. Matseliukh); 0000-0001-5797-594X (T. Batiuk); 0000-0002021960252X (M. Luchkevych); 0000-0003-0809-1480 (Iryna Shakleina); 0000-0001-5695-897X (H. Harbuzynska); 0000-0002-3778-924X (S. Kondratiuk); 0000-0002-4793-5984 (K. Zelenska)

2. Literature review

Analysis of data on various socio-economic indicators has been conducted previously by many scientists [6-10]. Prominent examples of public opinion research include [11-13], which studies which book author received the highest average rating, which author wrote the most bestsellers, which book received the most reviews, which genres became bestsellers more often how their ratings differ, etc. [14-27]. The advantage is that the author [11] analyzed important indicators to conclude specific authors and genres, but the disadvantage is that the author did not use smoothing methods. Also, the author [11] identified the most popular author only based on rating, not taking into account the number of reviews. The fact that the author [11] got rid of repetitions and cleaned the sample is an advantage and indicates the reliability of the results and minimization of deviations. In addition, the author [11] constructed a correlation matrix and visualized the results.

Many authors [12, 13] use various methods of data visualization. Other similar results include the works [28-30]. The authors [28, 29] conducted data research by constructing a histogram based on rating data, reviews, prices, and a bar chart. The author [30], in turn, built histograms and pie charts and used the method of linear regression. In addition, he built a heat map on which he reflected the dependencies of ratings, reviews and prices. Intelligent systems are used each time in new areas [31, 32], allowing successful management decisions.

3. Methods

The histogram is constructed for simple representation of tabular data, using two axes that display specific parameters and columns (rectangles) of the same width but different heights for visual comparison with other columns [33-39]. This comparison allows you to understand the difference between the selected elements simply.

There are two types of time series smoothing approaches: analytical and algorithmic. Analytical is based on assumptions, where the most striking examples may be the visual dependence of data in various functions, such as hyperbola, exponent, parabola, etc. One of the above functions will be suitable for a specific dependence, and the evaluation itself will be performed on this function. However, the algorithmic approach uses other methods, such as moving averages. This method is that at a particular time interval, the average value is chosen instead of the initial ones, respectively as if the angles are smoothed, which can be seen visually on the graph.

Smoothing with Kendall formulas is based on comparing hypotheses [40-43]. Specific indicators are "ranked" - sorted by a certain parameter. Similarities/coincidences are determined, the sum of coincidences is calculated according to certain formulas, and the sum of inversions is calculated. The significance level is determined, and conclusions can be drawn from hypotheses.

Exponential smoothing is used to detect trends over some time. The principle is one-parameter, i.e., its only parameter is the alpha, which is selected under certain conditions. The smoothing method's result strongly depends on the smoothing parameter alpha, where the less alpha - the more smoothed levels on the graph where we analyze them.

We will use correlation analysis as one of the research methods [1, 2, 40-43]. This method is used in various fields where statistics are used. It shows the relationship between two or more parameters. The method can be used to conclude according to statistical data and construct statistical probabilities in the form of diagrams for a visual understanding of the relevant dependencies.

The correlation field is conditionally a graph of the dependence of the x and y axes. It is a built-in correlation analysis, whereafter construction; you can determine the nature of the dependence, its direction and form. It is a visual representation of the correlation, which helps assess the dependencies of certain statistics better.

4. Experiments

At the beginning of our work, we had 550 rows of data. Our team intended to clear the data of duplicate books, as one book could be a bestseller for several years. Still, after researching the available

information, we realized that it is necessary to keep a history for a comprehensive analysis of anti-aliasing methods. That's why we decided to leave the dataset unchanged. So, we have 550 books sorted over 11 years at the group level (Table 1). Then, we sorted the books by year and assigned the authors an ID (Table 2).

Table 1
The beginer of the dataset

ID	Name	Author	User Rating	Reviews	Price	Year	Genre
1	Act Like a Lady Think Like a Man: What Men Think About Love, Relationships, Intimacy and Commitment	Steve Harvey	4,6	5013	17	2009	Non-Fiction
2	Arguing with Idiots: How to Stop Small Minds and Big Government	Glenn Beck	4,6	798	5	2009	Non-Fiction
3	Breaking Dawn (The Twilight Saga Book 4)	Stephenie Meyer	4,6	9769	13	2009	Fiction
4	Crazy Love: Overwhelmed by a Relentless God	Francis Chan	4,7	1542	14	2009	Non-Fiction
5	Dead And Gone: A Sookie Stackhouse Novel (Sookie Stackhouse/True Blood)	Charlaine Harris	4,6	1541	4	2009	Fiction

Table 2
The end of the dataset

ID	Name	Author	User Rating	Reviews	Price	Year	Genre
546	Unicorn Coloring Book: For Kids Ages 4-8 (US Edition) (Silly Bear Coloring Books)	Silly Bear	4,8	6108	4	2019	Non-Fiction
547	What Should Danny Do? (The Power to Choose Series)	Adir Levy	4,8	8170	13	2019	Fiction
548	Where the Crawdads Sing	Delia Owens	4,8	87841	15	2019	Fiction
549	Wrecking Ball (Diary of a Wimpy Kid Book 14)	Jeff Kinney	4,9	9413	8	2019	Fiction
550	You Are a Badass: How to Stop Doubting Your Greatness and Start Living an Awesome Life	Jen Sincero	4,7	14331	8	2019	Non-Fiction

Our dataset consists of seven columns [28-30]:

- ID - normal identifier;
- Name - the title of the book;
- Author - the author of the book;
- User rating - readers' rating, up to 5 points;
- Reviews - readers' reviews of the book;
- Price - the price of the book;
- Genre is the genre of the book.

We chose only two genres: Fiction and Non-fiction. Most of the bestsellers from our dataset were written by authors such as Stephanie Meyer, Stephen King, Rick Riordan, John Grisham, Jeff Kinney, J.K. Rowling, George R.R. Martin, and Dav Pilkey, Bill O'Reilly. If you analyze the column User rating,

the rating ranges from 3.3 to 4.9. "The Casual Vacancy" by J.K. Rowling had the lowest rating among our sample, but 27 books received the highest marks.

The book entitled "Divine Soul Mind Body Healing and Transmission System: The Divine Way to Heal You, Humanity, Mother Earth, and All Universes" by Zhi Gang Sha received as many as 37 reviews. It is interesting to note that the following book has the least number of reviews from the same author. On the other hand, the book from Delia received the most reviews from Owens, entitled "Where the Crawdads Sing», - with an incredible number - 87841.

Surprisingly, the price of the book also affects its success. Thus, in our dataset, prices ranged from $ 1 (10 books) to $ 105 ("Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5" from the American Psychiatric Association).

We had only two genres - Fiction and Non-fiction. One hundred fifty-nine books of the genre are fiction, and one hundred eighty-six books are Non-fiction. Thus, we can say that non-fiction books became bestsellers more often than fiction.

Ivan Naratov [29], using the same dataset, identified the best authors by user rating (Fig. 1)

Author	User Rating
Nathan W. Pyle	4.9
Patrick Thorpe	4.9
Eric Carle	4.9
Emily Winfield Martin	4.9
Chip Gaines	4.9
Jill Twiss	4.9
Rush Limbaugh	4.9
Sherri Duskey Rinker	4.9
Alice Schertle	4.9
Pete Souza	4.9
Sarah Young	4.9
Lin-Manuel Miranda	4.9
Bill Martin Jr.	4.9

Figure 1: The best authors for Ivan Naratov

After analyzing the data, we determined:

- The following 13 authors have the highest rating: Nathan W. Pyle, Patrick Thorpe, Eric Carle, Emily Winfield Martin, Chip Gaines, Jill Twiss, Rush Limbaugh, Sherri Duskey Rinker, Alice Schertle, Pete Souza, Sarah Young, Lin-Manuel Miranda, Bill Martin Jr., Dav Pilkey. Their average rating is 4.9. If you buy a book, you should look at these authors.
- Authors who wrote the most bestsellers: Jeff Kinney - 12 books, Rick Riordan - 10, JK Rowling - 8, Stephenie Meyer - 7, Dav Pilkey - 6, Bill O'Reilly - 6, John Grisham - 5, EL James - 5, Suzanne Collins - 5, Charlaine Harris - 4. These authors always have something worth reading.
- The most reviewed books are the followings: Where the Crawdads Sing - 87841 reviews, The Girl He the Train - 79446, Becoming - 61133, Gone Girl - 57271, The Fault in Our Stars - 50482.

The analysis identified which authors received the highest ratings from readers, which authors wrote the most bestsellers, and which books received the most reader reviews.

In addition, it was found that non-fiction literature is increasingly becoming a bestseller, but users also prefer fiction, which confirms the statistically significant results obtained during testing.

5. Results

Our team took a sample of all elements and two columns - ID and Reviews to conduct descriptive statistics. The main type of data visualization in experimental and scientific-practical research reports is graphs. The graph shows the relationship between two quantities, one of which is an independent variable and its value, and the second variable is dependent and its value.

On the graph "Dynamics of the indicator", we have shown the dynamics of responses to each book. The book id is an independent variable, and the number of reviews is dependent (Fig. 2, Fig.3).

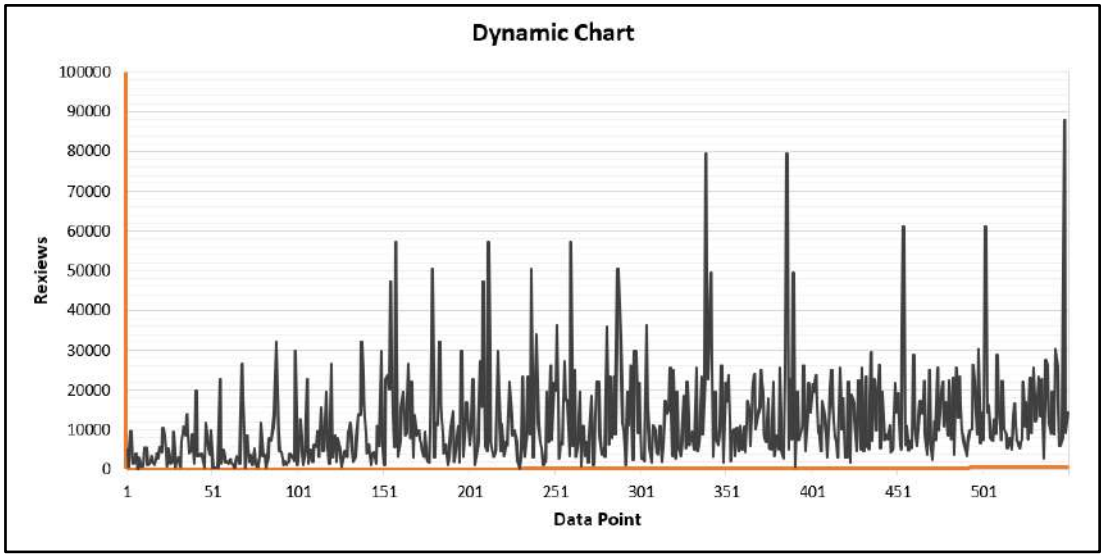


Figure 2: Graphical representation of the dynamics of the indicator in the Cartesian coordinate system

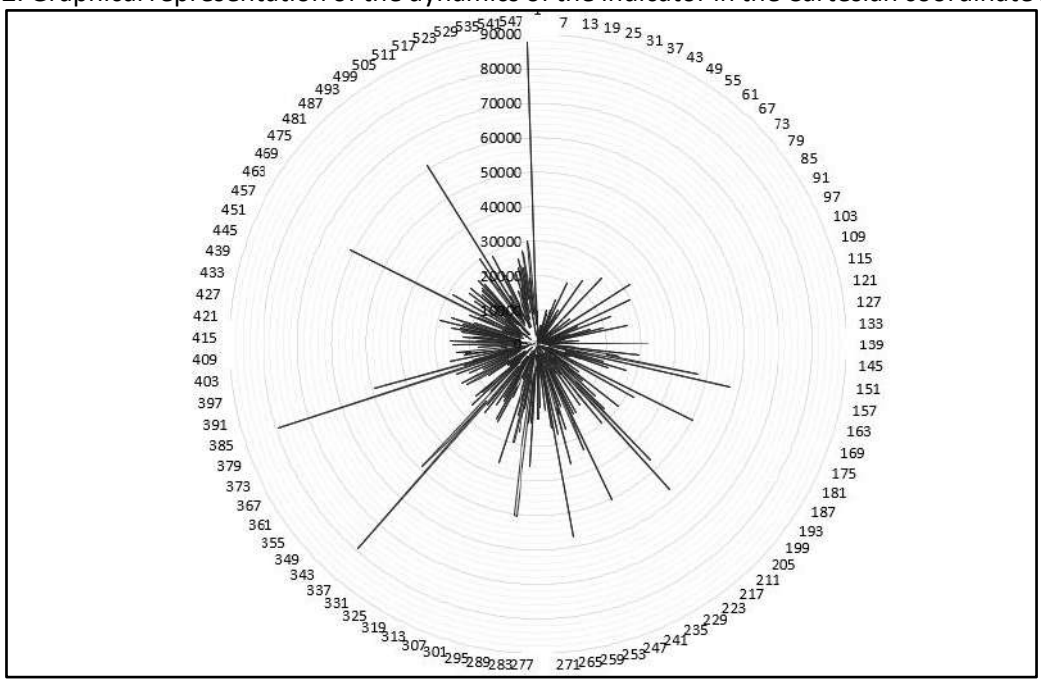


Figure 3: Graphical representation of the dynamics of the indicator in the polar coordinate system

In addition to the tabular and graphical representation of data, they are accompanied by their general numerical and statistical characteristics, which relate to descriptive or descriptive statistics.

The volume of our sample is 550. Let's start with the arithmetic mean, which measures the major trend that reflects this sample's most characteristic value. The formula determines it [33-43]:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad (1)$$

where *n* is the sample size.

Mode (denoted by "Mo") is the most common value among the variables sample.

Median (denoted by "Me") is a value that bisects an ordered set of variables, i.e., to determine the median, you need to sort the data, for example, in ascending order.

The values of mean, mode and median are close to each other. They are equal in an ideal normal distribution because they have the same meaning: the middle of the distribution.

Scope (interval) - an indicator that indicates the width of the range of values. It is equal to the difference between the maximum and minimum values.

Standard deviation (σ) - is a measure of variability (variation) of the trait, which reflects the magnitude of its variance relative to the arithmetic mean [33-43]:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}, \quad (2)$$

where *n* is the sample size.

For a more accurate and clear idea of the variation of the values of the indicator relative to the average, the coefficient of variation is used [19-29]:

$$v = \frac{\sigma}{\bar{x}} \cdot 100\%, \quad (3)$$

where σ is standard deviation.

Coefficient of variation expresses the degree of variability of the sign-in per cent. Asymmetry is an indicator that reflects the skew of the distribution relative to the fashion to the left or right. For left-handed or positive, asymmetry in the distribution, lower values are more common, and for right-handed or negative - higher.

Excess is an indicator that reflects the height of the distribution. In those cases, any reasons promote the emergence of close to average values, the distribution with positive excess is formed. Suppose extreme values dominate the distribution and, at the same time, lower and higher. In that case, such a distribution is characterized by negative excess, and in the center of the distribution may form a depression, which turns it into two vertices. The results of these indicators can be seen in Table 3.

Table 3

Results of descriptive statistics prepared for the report (quantitative data are presented with two digits)

Head 1	Head 2	Head 3
Sample size	$n =$	550,00
Average	$\bar{x} =$	11993,73
Standard error	$\mathrm{e} =$	153,14
Mode	$M_0 =$	8580,00
Median	$M_e =$	8580,00
Sampling range	$P =$	87804,00
Standard deviation	$\sigma =$	11758,31
Sampling variance	$D =$	138257875,10
Coefficient of variation	$v =$	0,98
Asymmetry	$A =$	2,40
Excess	$E =$	8,67
Minimum	$\min =$	37,00
Maximum	$\max =$	87841,00
The sum	$\Sigma =$	6596552,00
Interval	$I =$	87804,00

So, analyzing the results (Table 3), we can say that:

- The average number of reviews for the book is 11993;
- Fashion and median are the same.
- Width of our range - scope - 87804;
- The measure of the variability of responses is the value equal to 11758, and the coefficient of variation = 0.98, which is quite low;
- Asymmetry is equal to 2.39. Lower values are more common in the distribution, so it is left-handed (positive).

If the obtained data form a normal representative sample, i.e., a sample whose volume is sufficient to determine their distribution, traditionally, to establish its type, a histogram is built. It is also a chart that clearly shows the method of grouping data by some essential features.

The most problematic in constructing a histogram is choosing the number of grouping intervals, i.e. the number of groups divided into the sample. There are several dozen formulas for selecting or determining the number of partition intervals, but the two most common are:

Sturges' formula is following [19-29]:

$$k = 1 + \log_2 n, \tag{4}$$

where *k* is number of intervals.

Scott's formula is following [19-29]:

$$h = 3.5 \cdot s \cdot n^{-1/3} = \frac{3.5 \cdot s}{\sqrt[3]{n}}, \tag{5}$$

where *h* is the width of the interval, and *s* is the standard deviation of the values of the series.

The number of intervals k is equal 10,103288 by the Sturges' formula, and the interval width *h* is equal 5022,9488 by the Scott's formula.

Based on the image of the histogram, the analysis of frequencies of values is carried out for approximation of empirical density of the law of distribution of initial data - the received histogram by some analytical function.

When constructing our histogram, based on the results according to the formulas of Sturgess and Scott, we obtained the following histogram (Fig. 4). You can see that based on the histogram image used - hyperbolic distribution law, hyperbolic function.

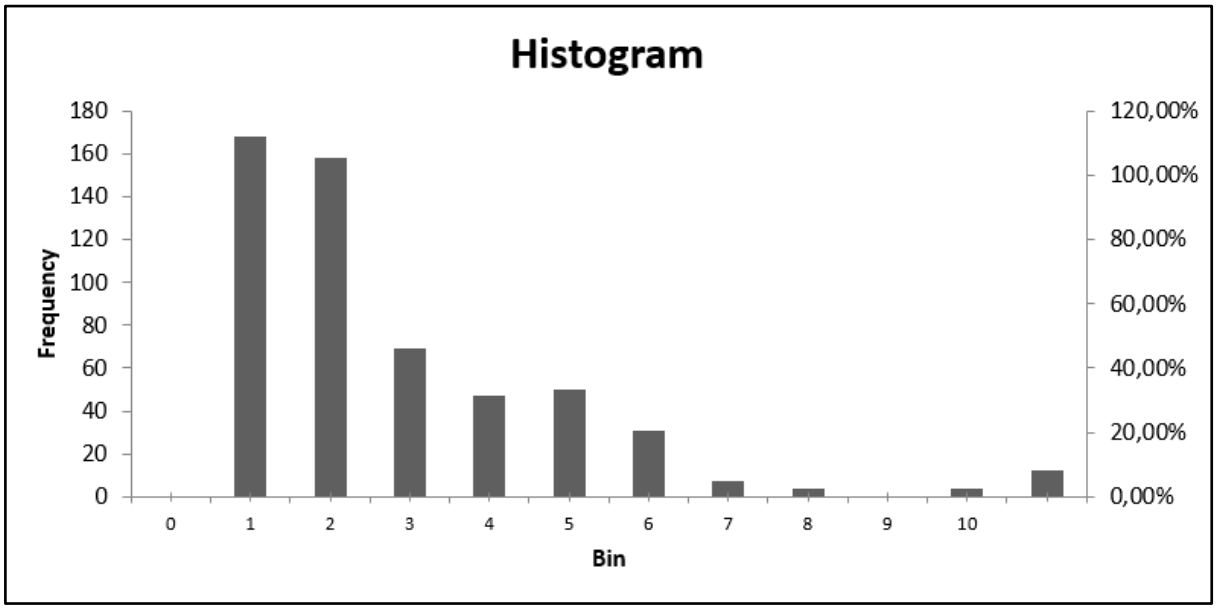


Figure 4: Histogram

The main disadvantages of the histogram are: subjectivity in choosing the number of grouping intervals, the value of the arithmetic mean, modes and medians given to the size of the interval, and the limited number of approximation nodes.

However, some of these shortcomings can be eliminated by using the method of constructing a histogram to construct a cumulative - empirical graph of the distribution law function. The results of which can be seen in Fig. 5.

6. Discussion

Smoothing methods can be divided into two classes based on analytical and algorithmic approaches [19-29]. In the algorithmic approach, the appearance of the trend is obtained due to various algorithms

that practically implement smoothing procedures [44-56]. These procedures provide the researcher only with an algorithm for calculating the new value of the time series at any given time *t*.

These methods can be classified as follows [55-78]:

- simple or ordinary moving average;
- weighted moving average;
- exponential smoothing;
- median smoothing.

For these methods, we used the entire sample size in 2 categories - ID and Reviews.

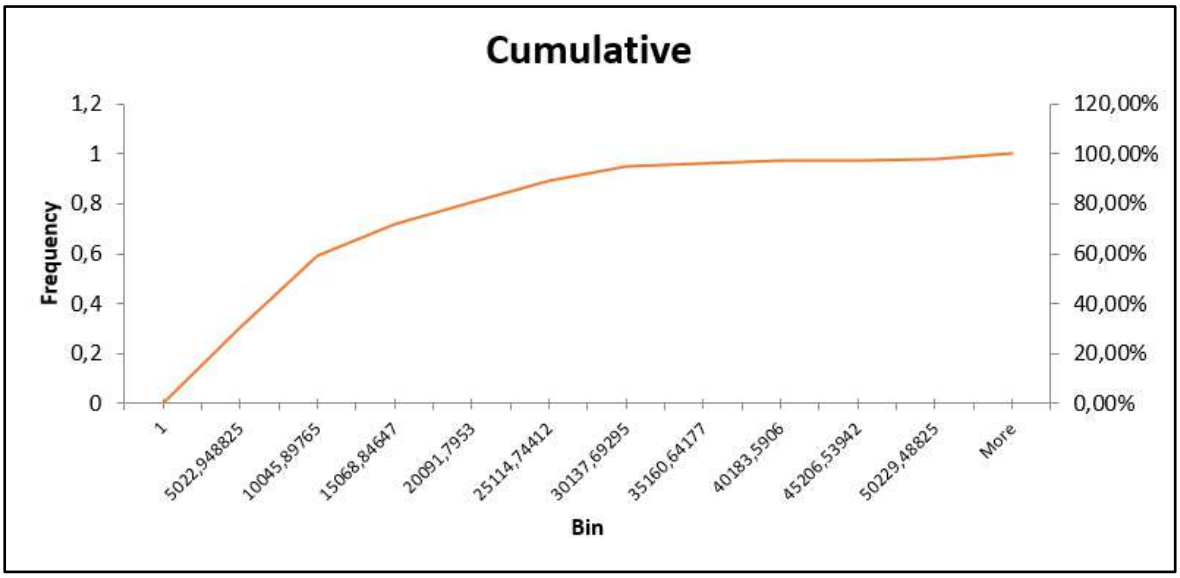


Figure 5: The cumulate is based on histograms

Smoothing, according to Kendall formulas [44], is a simple moving average when smoothing the data, using different sizes of smoothing intervals w = 3, 5, 7, 9, 11, 13, 15, got seven consecutive columns. After graphical construction of each smoothing interval, we obtained the following graphs (Fig. 6-12). According to them, it can be clearly stated that the larger the interval, the smoother the curve. Smoothed the data using the size of the smoothing interval $w = 3$ (Fig. 6), $w = 5$ (Fig. 7), $w = 7$ (Fig. 8), $w = 9$ (Fig. 9), $w = 11$ (Fig. 10), $w = 13$ (Fig. 11), $w = 15$ (Fig. 12).

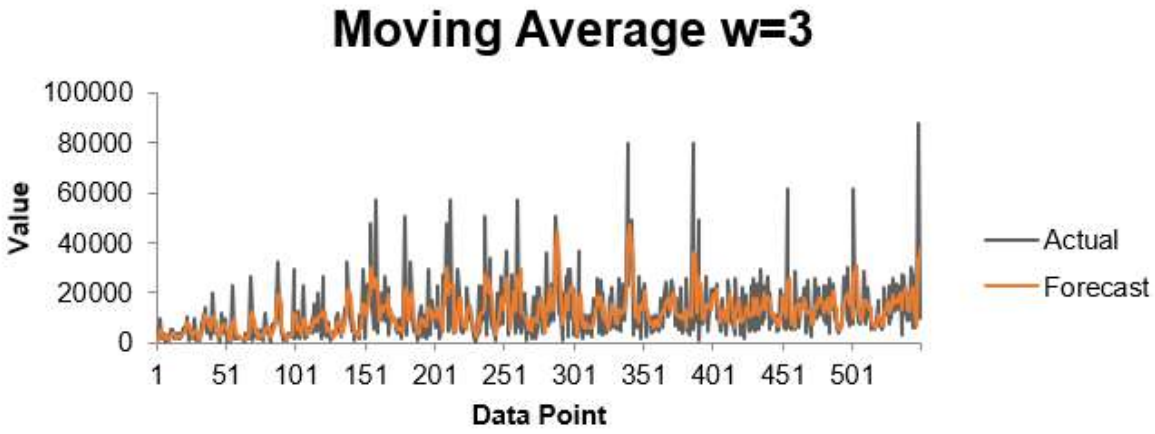


Figure 6: The result of smoothing at w = 3

The obtained data were again smoothed, but using the size of the smoothing interval $w = 5$ ($w = 3$) (Fig. 13). Smoothing of the obtained data was continued with a smoothing interval $w = 7$ ($w = 5$) (Fig. 14) and until $w = 15$ ($w = 13$) (Fig. 19). Table 4 show the obtained consecutive seven columns. After graphical construction of each smoothing interval, we obtained the following graphs (Fig. 13-19).

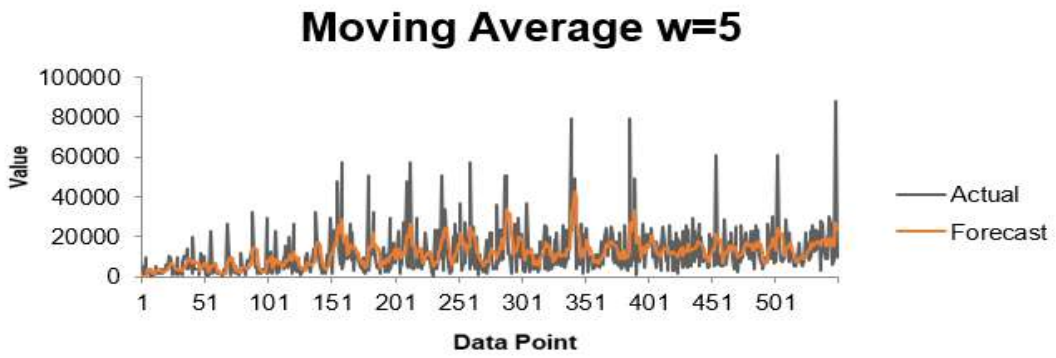


Figure 7: The result of smoothing at w = 5

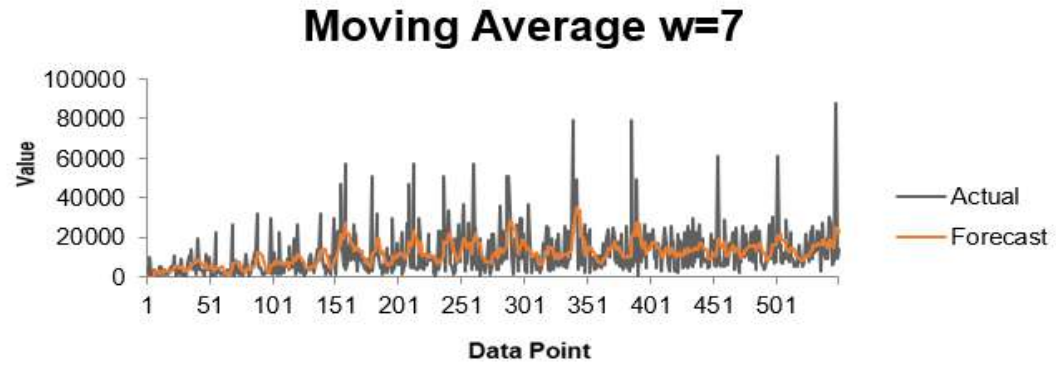


Figure 8: The result of smoothing at w = 7

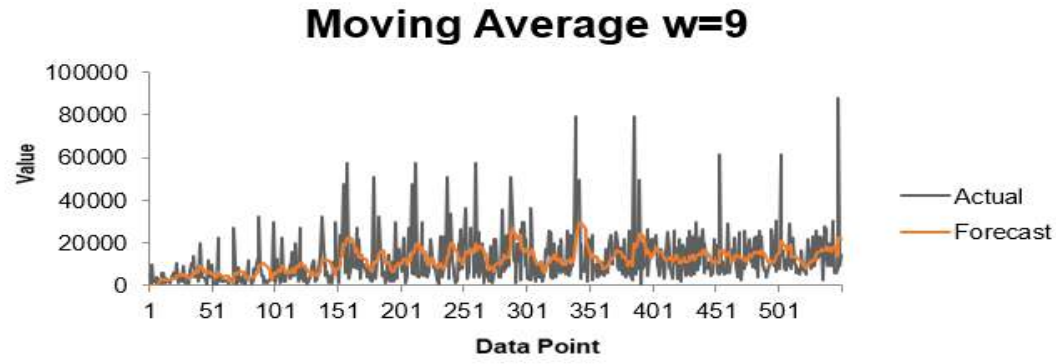


Figure 9: The result of smoothing at w = 9

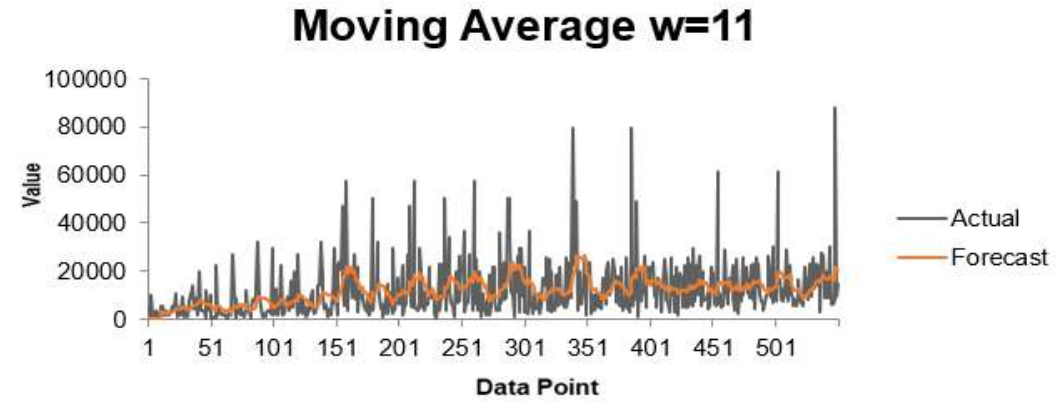


Figure 10: The result of smoothing at w = 11

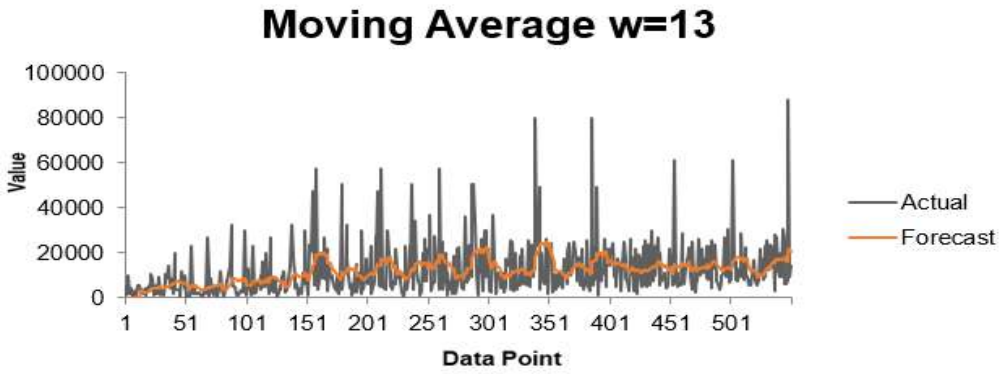


Figure 11: The result of smoothing at w = 13

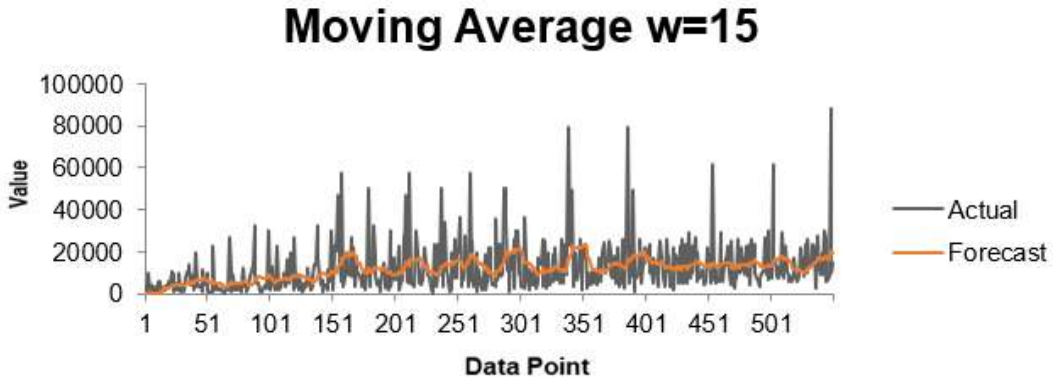


Figure 12: The result of smoothing at w = 15

Table 4 show the obtained consecutive seven columns. After graphical construction of each smoothing interval, we obtained the following graphs (Fig. 13-19).

Table 4

The result of smoothing consistently at w = 3, 5, 7, 9, 11, 13, 15

w=5 (w=3)	w=7 (w=5)	w=9 (w=7)	w=11 (w=9)	w=13 (w=11)	w=15 (w=13)	w=17 (w=15)
#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
4504,56	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
3955,08	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
3525,07	3525,07	#N/A	#N/A	#N/A	#N/A	#N/A
2956,73	3150,70	#N/A	#N/A	#N/A	#N/A	#N/A
2412,00	2893,20	3041,38	#N/A	#N/A	#N/A	#N/A
1879,00	2702,03	2702,96	#N/A	#N/A	#N/A	#N/A
1896,40	2613,14	2613,14	2811,37	#N/A	#N/A	#N/A
2333,13	2533,26	2647,64	2742,53	#N/A	#N/A	#N/A
2687,80	2435,09	2621,84	2697,29	3037,82	#N/A	#N/A
2992,33	2384,83	2631,04	2668,89	2865,30	#N/A	#N/A
3099,07	2594,26	2642,57	2642,57	2796,60	2954,15	#N/A
3088,53	2725,03	2607,92	2649,13	2756,70	2815,85	#N/A
2737,93	2752,94	2608,13	2631,25	2695,51	2729,45	2865,38
2384,80	2803,57	2582,95	2651,48	2692,69	2701,17	2755,03
2356,00	2734,83	2656,73	2687,93	2687,93	2670,96	2707,28

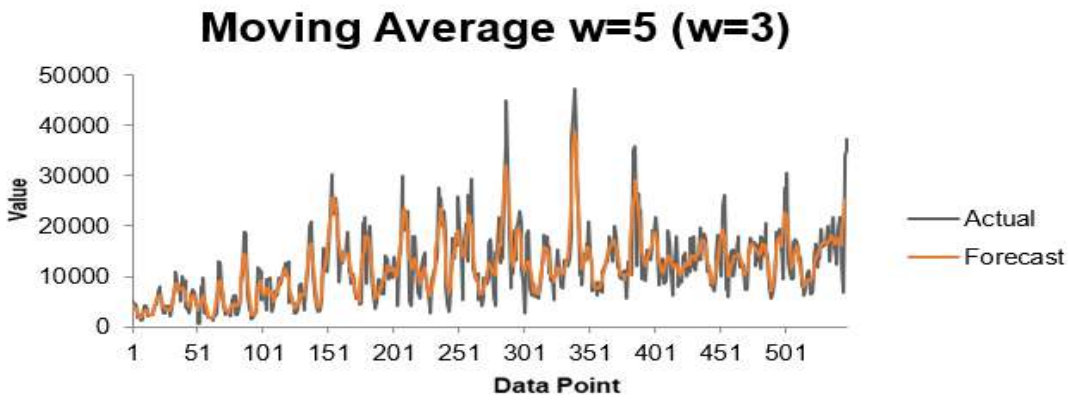


Figure 13: The result of smoothing at w = 5 (w = 3)

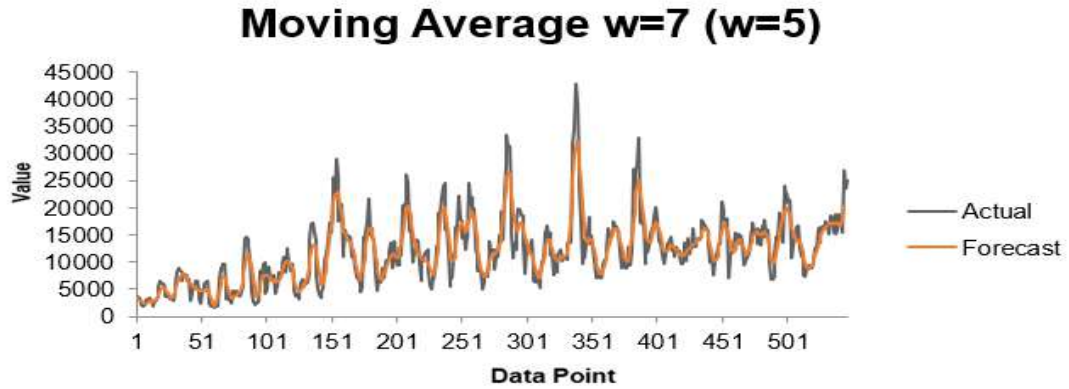


Figure 14: The result of smoothing at w = 7 (w = 5)

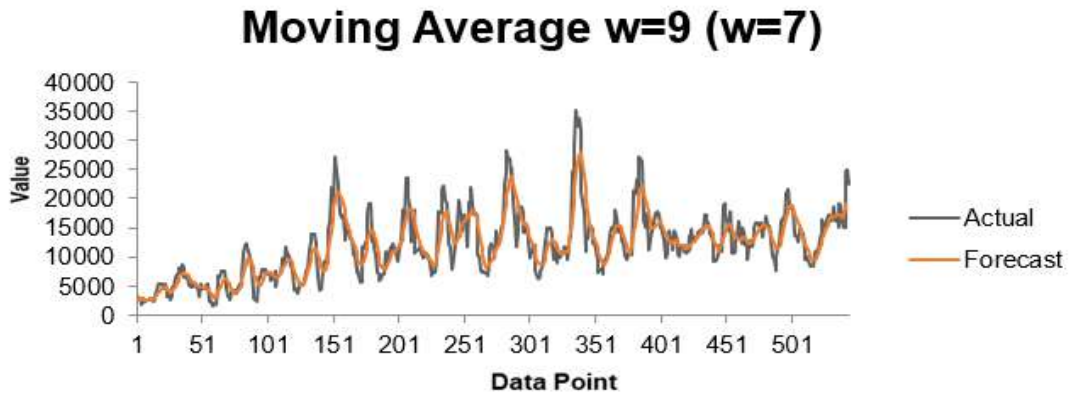


Figure 15: The result of smoothing at w = 9 (w = 7)

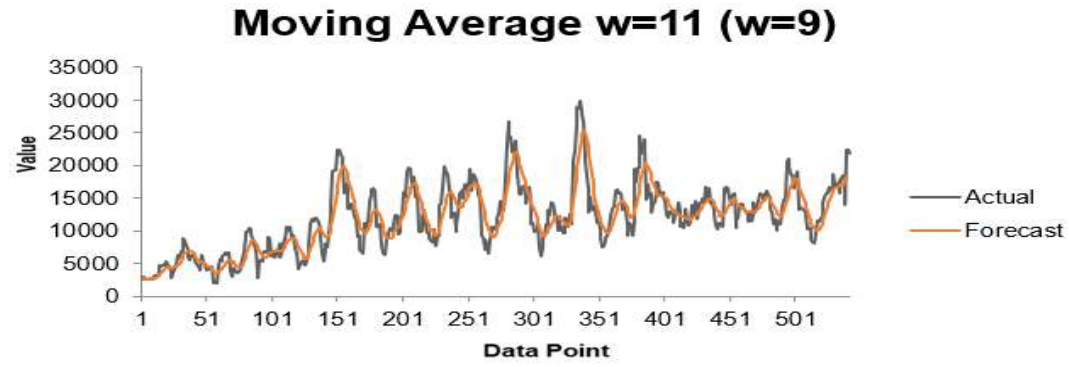


Figure 16: The result of smoothing at w = 11 (w = 9)

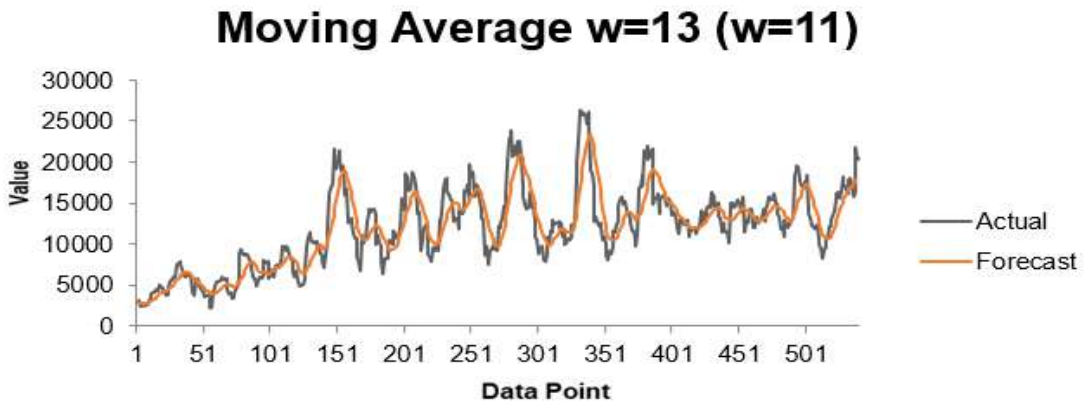


Figure 17: The result of smoothing at w = 13 (w = 11)

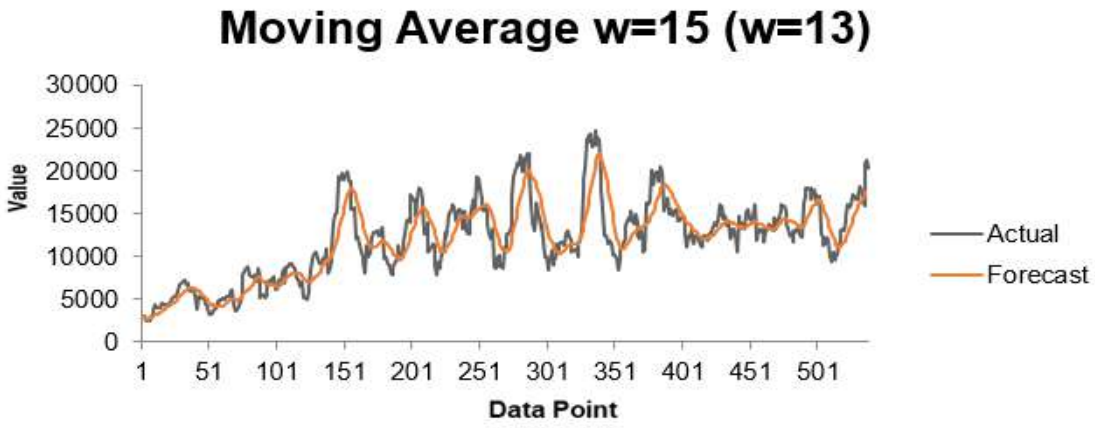


Figure 18: The result of smoothing at w = 15 (w = 13)

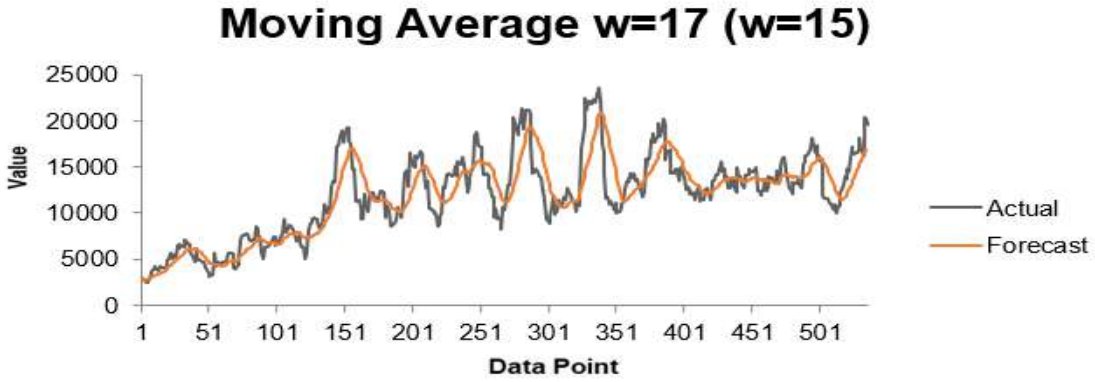


Figure 19: The result of smoothing at w = 17 (w = 15)

The number of turning points and their correlation coefficients between the original and smoothed values were found for each smoothing (Table 5).

Table 5

The number of turning points and the correlation coefficient for each w

Turn. points	w=3	w=5	w=7	w=9	w=11	w=13	w=15	w=5 (3)	w=7 (5)	w=9 (7)	w=11 (9)	w=13 (11)	w=15 (13)	w=17 (15)
Number TP	288	290	272	276	288	288	274	150	100	84	64	58	56	30
Cor. coeff.	0,646	0,556	0,490	0,429	0,400	0,385	0,365	0,353	0,266	0,234	0,221	0,213	0,206	0,204

After the results, you can see that turning points become less than individual smoothing with sequential smoothing.

Smoothing according to formulas from Pollard. Here, depending on the size of the smoothing interval, the weight for the main level will change. Smoothing was performed in the same way as in the previous paragraph. The result of smoothing according to Pollard's formulas is presented in Fig.20 – Fig. 26.

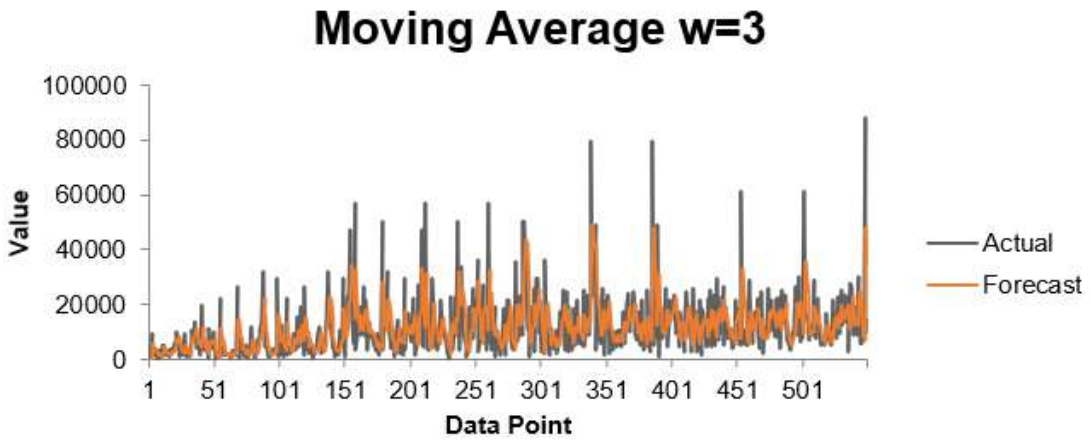


Figure 20: The result of smoothing according to Pollard formulas at w = 3

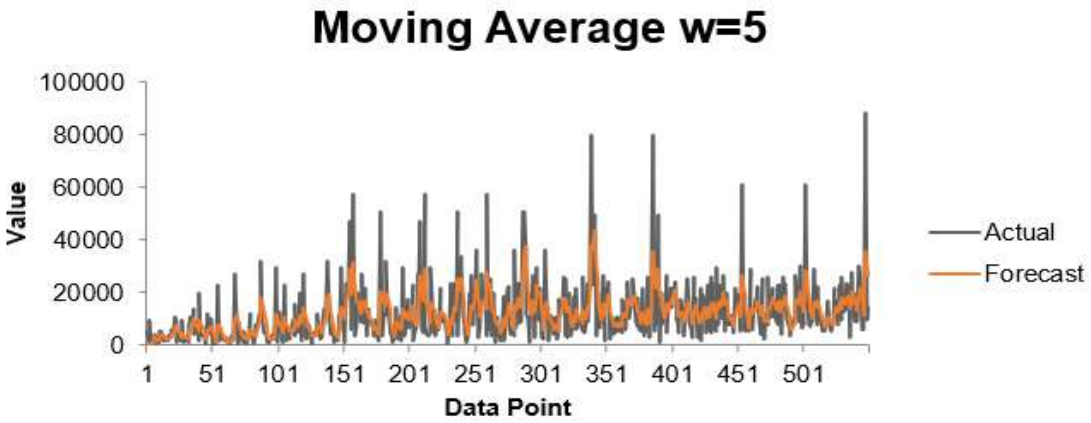


Figure 21: The result of smoothing according to Pollard formulas at w = 5

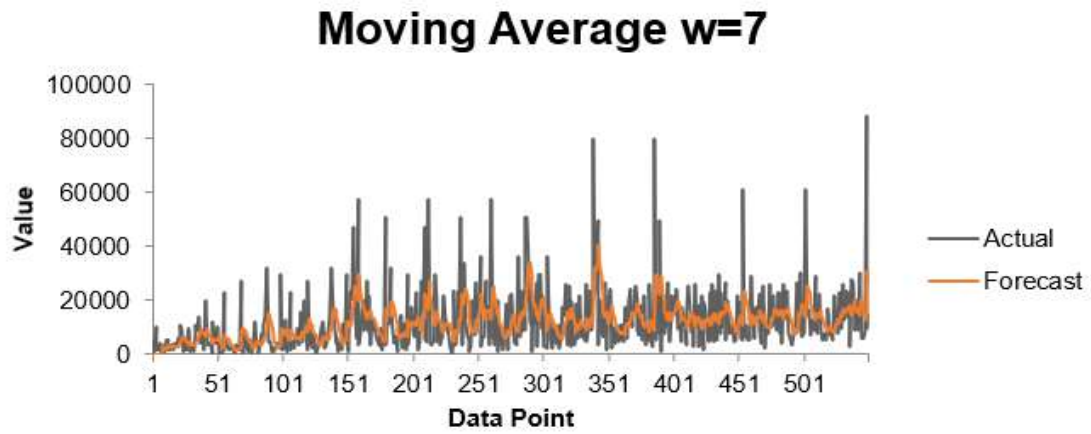


Figure 22: The result of smoothing according to Pollard formulas at w = 7

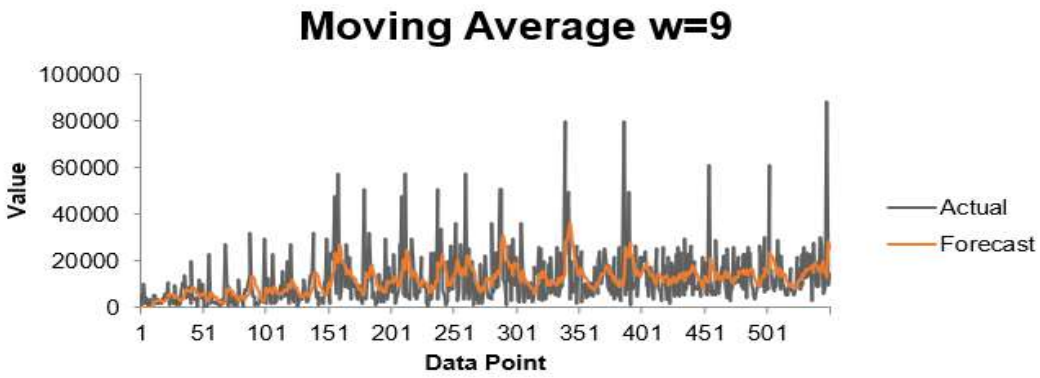


Figure 23: The result of smoothing according to Pollard formulas at w = 9

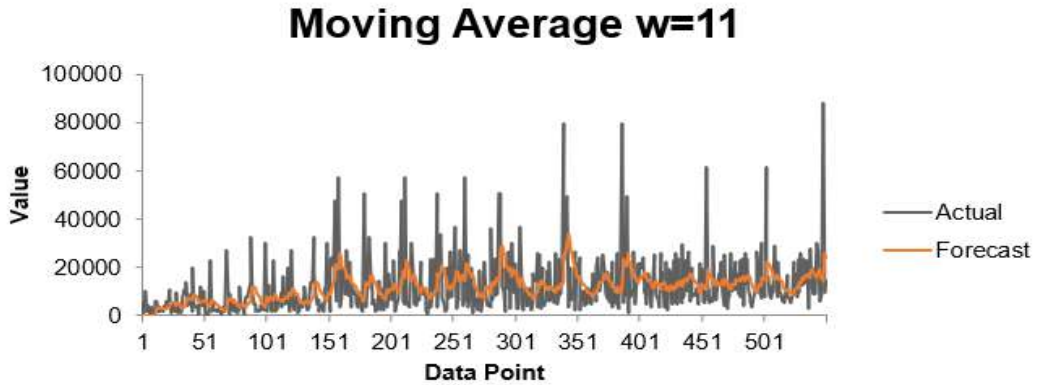


Figure 24: The result of smoothing according to Pollard formulas at w = 11

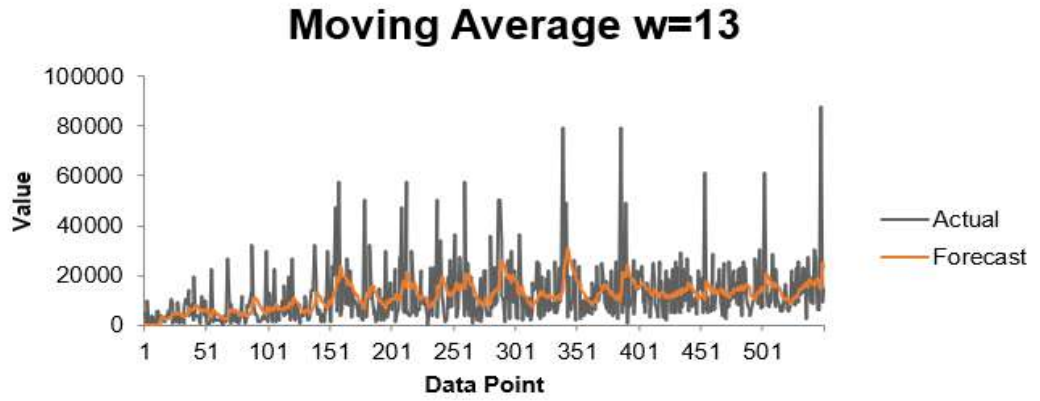


Figure 25: The result of smoothing according to Pollard formulas at w = 13

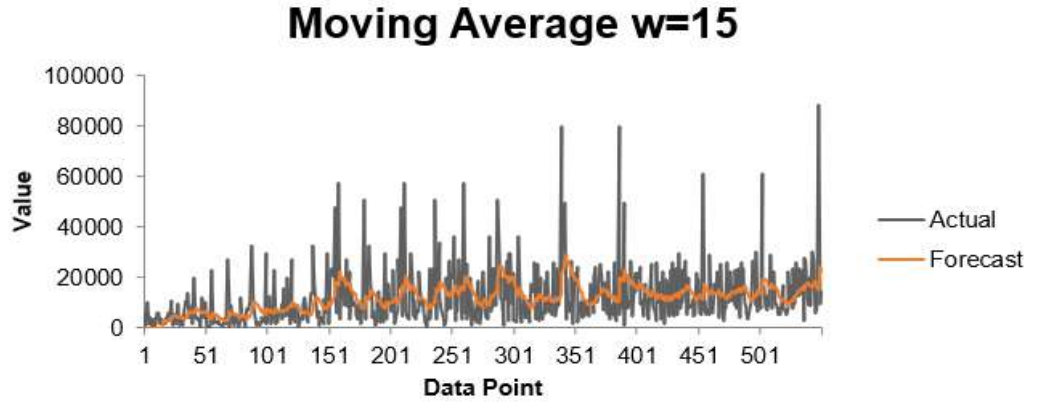


Figure 26: The result of smoothing according to Pollard formulas at w = 15

Therefore, comparing with the method of smoothing according to Kendall's formulas, we can say that smoothing according to Pollard's formulas does not give such a smooth result as the previous method, which can be seen when comparing the presented graphs. The result of smoothing according to Pollard formulas at w = 5 (w = 3) we present in Fig.27, w = 7 (w = 5) – in Fig.28, w = 9 (w = 7) – in Fig.29, w = 11 (w = 9) – in Fig.30, w = 13 (w = 11) – in Fig.31, w = 15 (w = 13) – in Fig.32, w = 17 (w = 15) – in Fig.33.

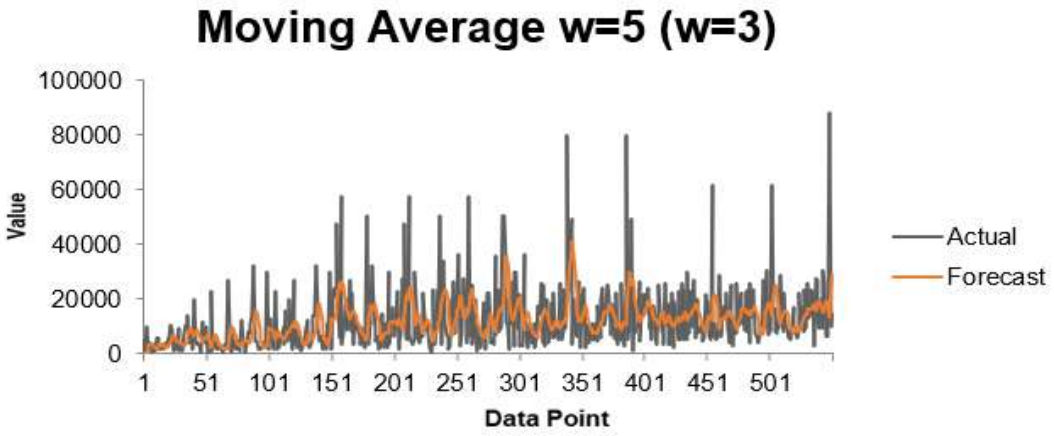


Figure 27: The result of smoothing according to Pollard formulas at w = 5 (w = 3)

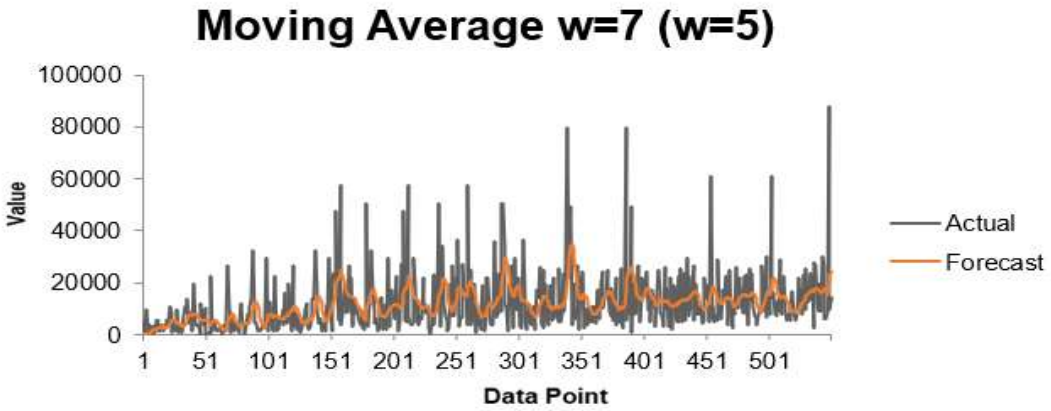


Figure 28: The result of smoothing according to Pollard formulas at w = 7 (w = 5)

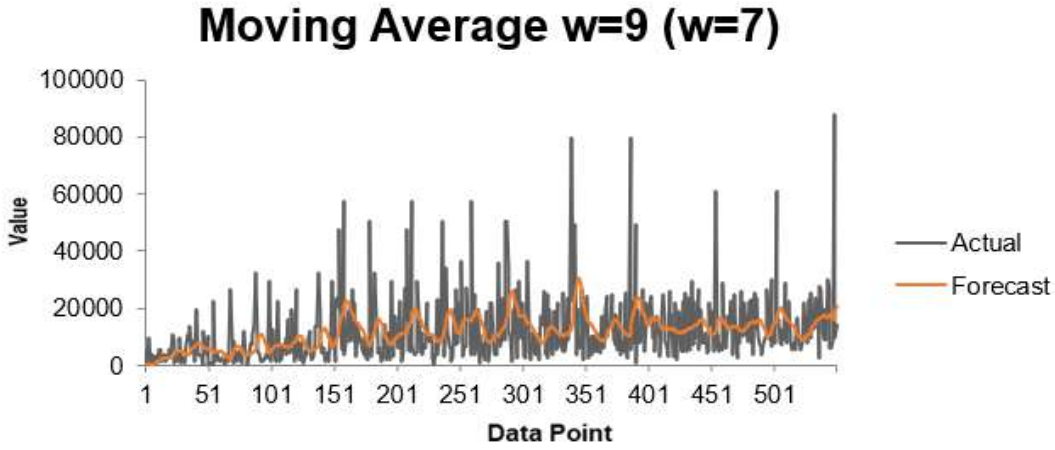


Figure 29: The result of smoothing according to Pollard formulas at w = 9 (w = 7)

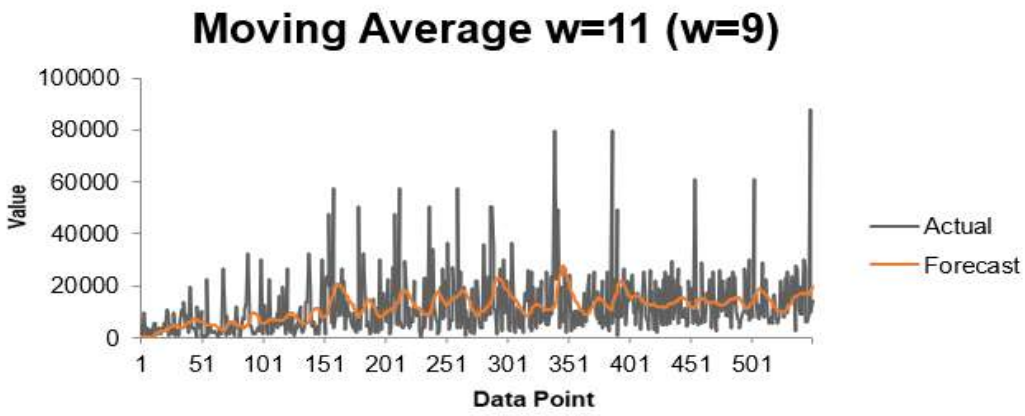


Figure 30: The result of smoothing according to Pollard formulas at w = 11 (w = 9)

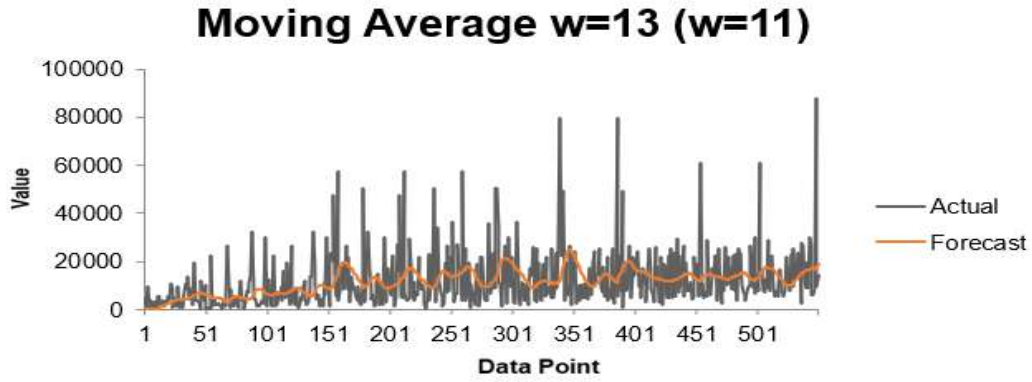


Figure 31: The result of smoothing according to Pollard formulas at w = 13 (w = 11)

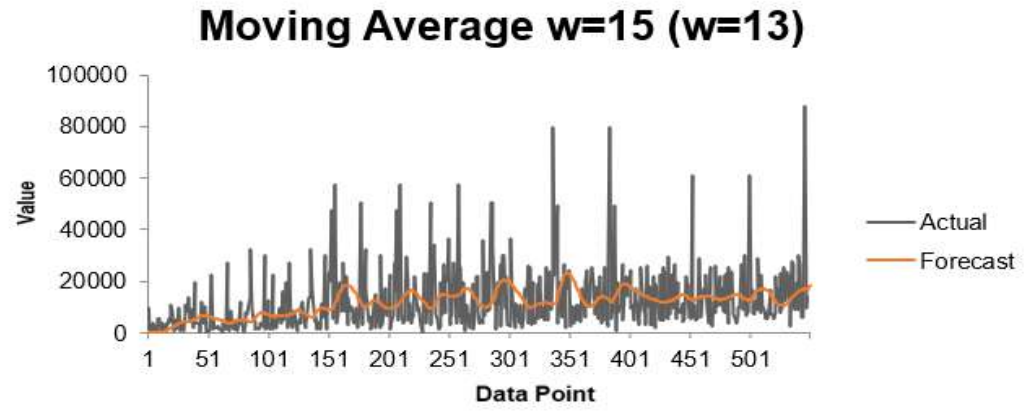


Figure 32: The result of smoothing according to Pollard formulas at w = 15 (w = 13)

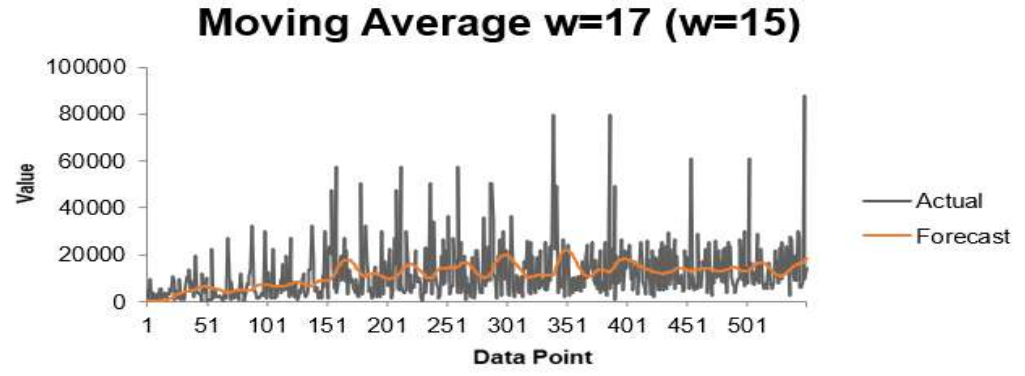


Figure 33: The result of smoothing according to Pollard formulas at w = 17 (w = 15)

Exponential smoothing. An essential feature of the use of exponential averages is the justification of the value of the smoothing parameter. The smaller it is, the more the levels in the analyzed series are smoothed. It means an increase in specific. This method of smoothing is widely used in forecasting economic time series. The main parameter of exponential smoothing is the parameter - α, which takes values from 0.1 to 0.3.

We smoothed the same series with different values of the parameter α: 0.1, 0.15, 0.2, 0.25, 0.3. In all these cases, we found for each smoothing the number of turning points and correlation coefficients between the original values and the smoothed ones. The result of exponential smoothing at $\alpha = 0.1$ we present in Fig.34, at $\alpha = 0.15$ – in Fig.35, at $\alpha = 0.2$ – in Fig.36, at $\alpha = 0.25$ – in Fig.37, at $\alpha = 0.3$ – in Fig.38.

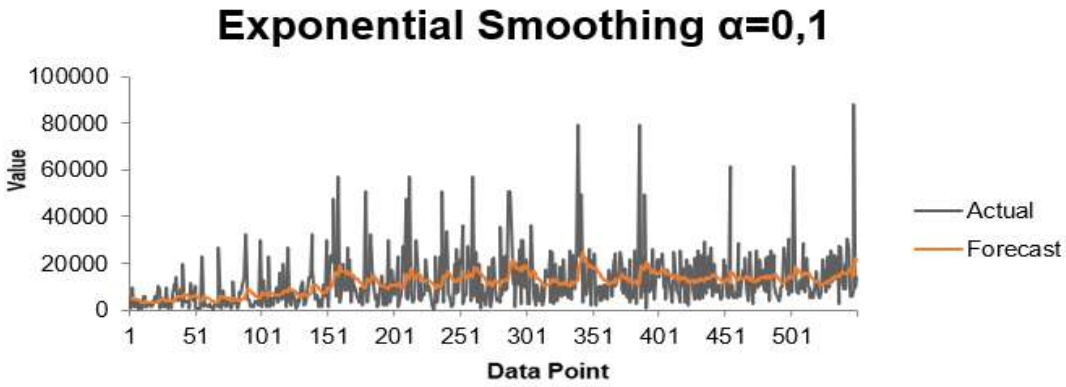


Figure 34: The result of exponential smoothing at $\alpha = 0.1$

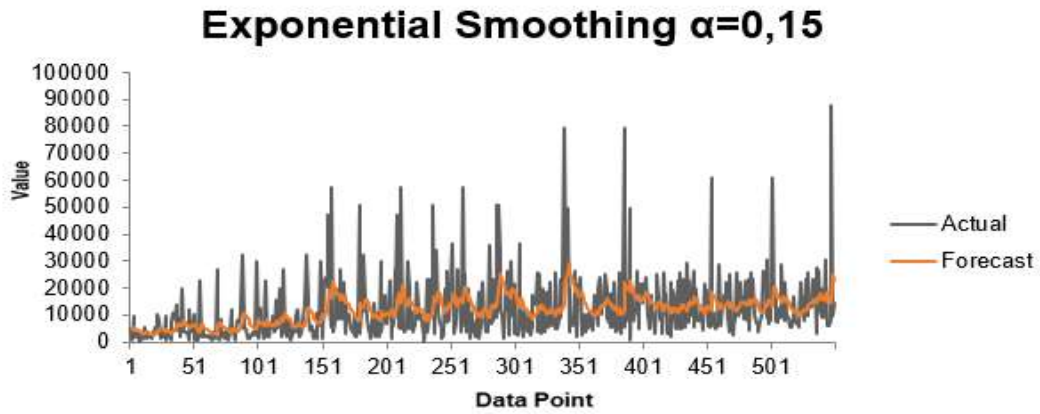


Figure 35: The result of exponential smoothing at $\alpha = 0.15$

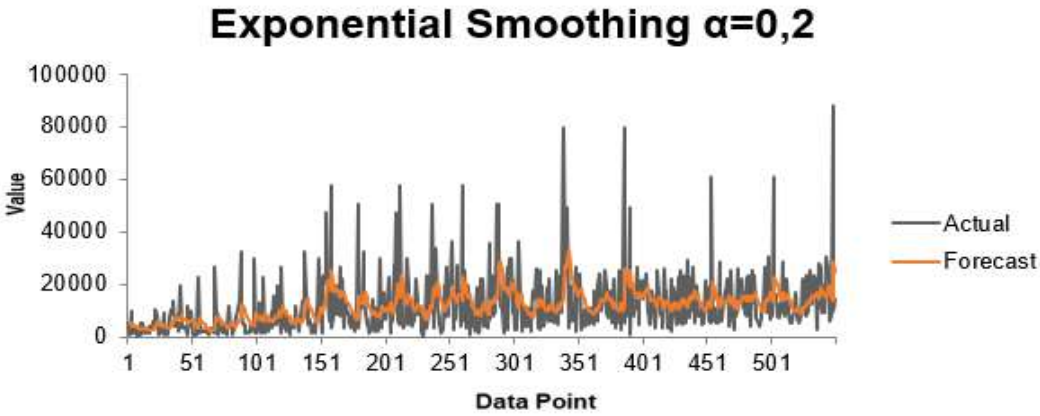


Figure 36: The result of exponential smoothing at $\alpha = 0.2$

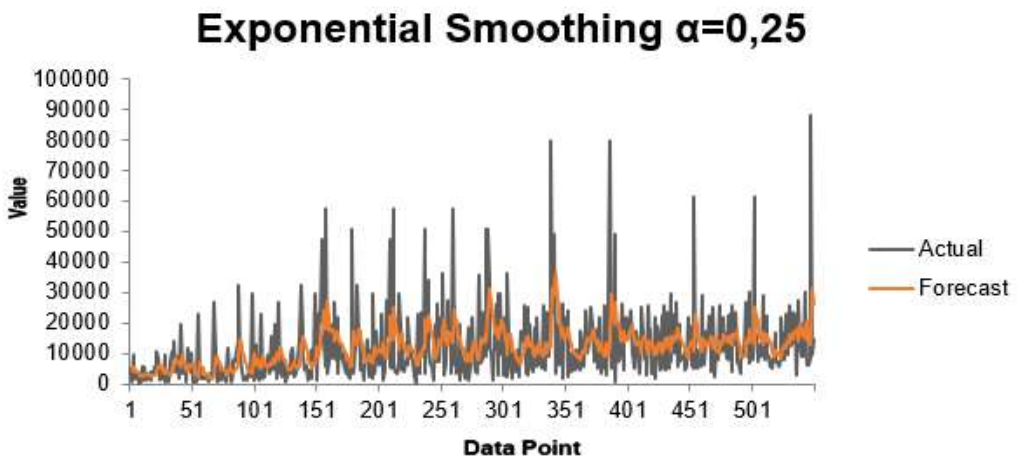


Figure 37: The result of exponential smoothing at α = 0.25

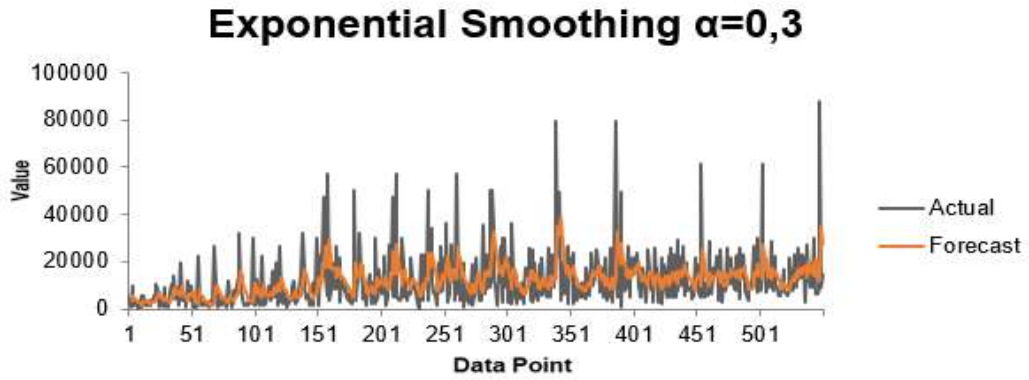


Figure 38: The result of exponential smoothing at α = 0.3

When plotting exponential smoothing graphs, we noticed that the smaller the alpha, the smoother the graph itself. The number of turning points and the correlation coefficient for each α are shown in Table 6.

When analyzing the Table 6 of the number of turning points and correlation coefficients, we can see that the greater the alpha - the greater the coefficient.

Table 6

The number of turning points and the correlation coefficient for each α

Turning points	α=0,1	α=0,15	α=0,2	α=0,25	α=0,3
Number of turning points	268	272	276	282	290
Correlation coefficient	0,245864	0,233523	0,222536	0,211596	0,20043

Median smoothing. The median refers to the distributive mean, i.e., the value of the feature that takes place in the middle of the variation series and, in contrast to the arithmetic mean, which summarizes the indicator's value, leaves the value of the indicator that corresponds to the median.

Median smoothing is as following.

- Eliminates single extreme or anomalous values of levels that are at least half the distance from the smoothing interval;
- Maintains sharp differences in trends (moving average and exponential smoothing lubricates them);
- Effectively eliminates single levels with very large or very small values that are random and stand out sharply among other levels;

- When re-applying the median smoothing at a constant window interval size, the differences in the behaviour of the levels persist as long as there are changes in the smoothed series.

The main properties of the median are as follows: the number of positive deviations from the median is equal to the number of negatives, and the sum of the absolute deviations of the sampling option relative to the median is minimal.

When working on this method, our team used the same smoothing intervals as in previous methods (w = 3, 5, 7, 9, 11, 13, 15) which is presented in the Fig.39 - Fig.45.

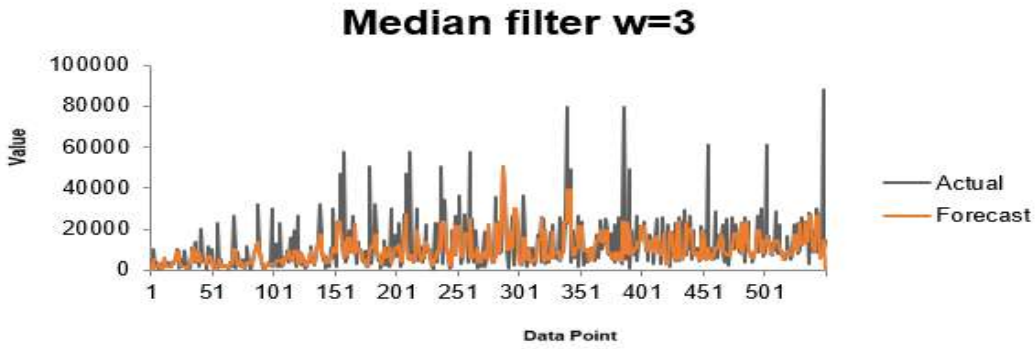


Figure 39: The result of the median filtering at w = 3

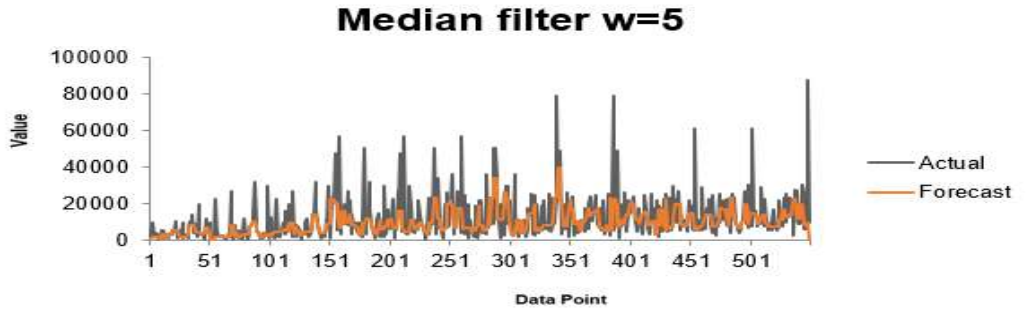


Figure 40: The result of the median filtering at w = 5

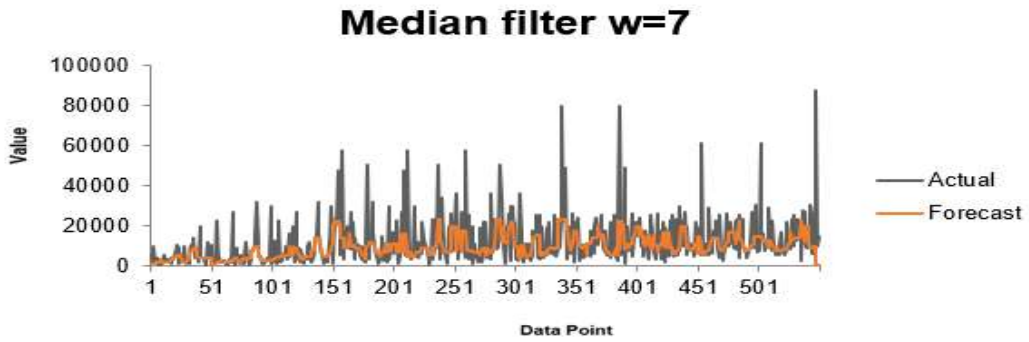


Figure 41: The result of the median filtering at w = 7

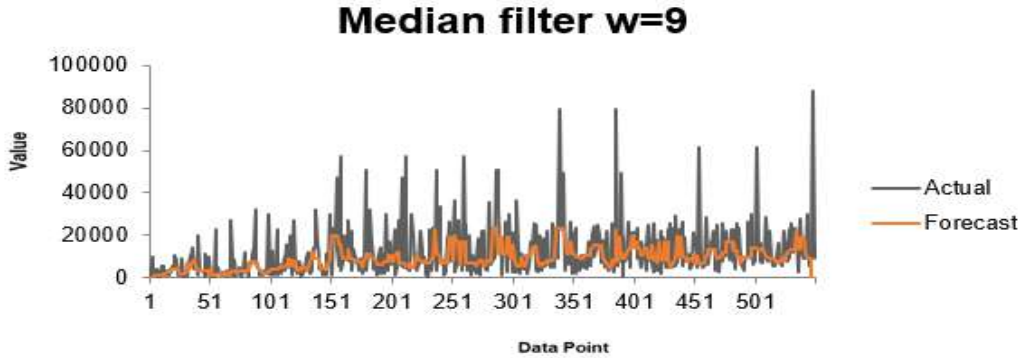


Figure 42: The result of the median filtering at w = 9

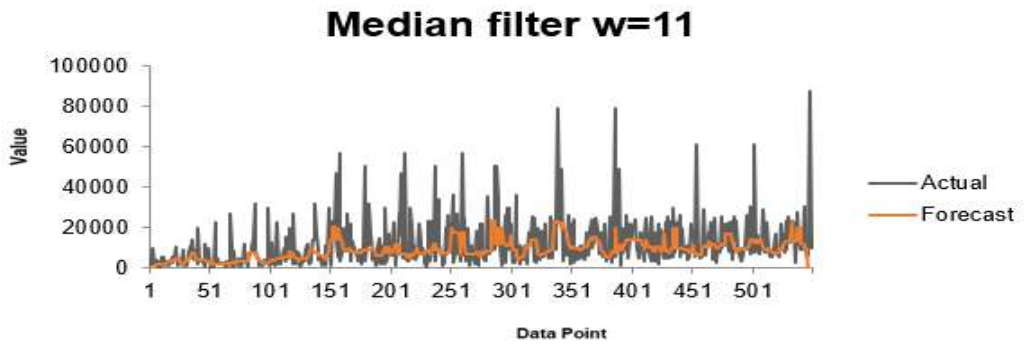


Figure 43: The result of the median filtering at w = 11

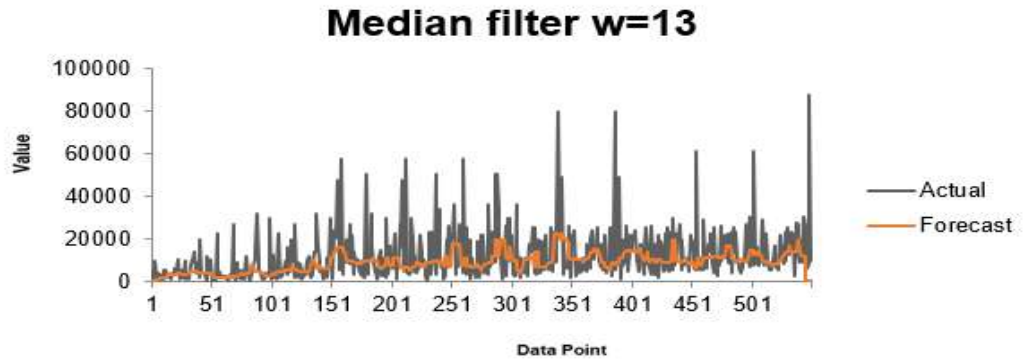


Figure 44: The result of the median filtering at w = 13

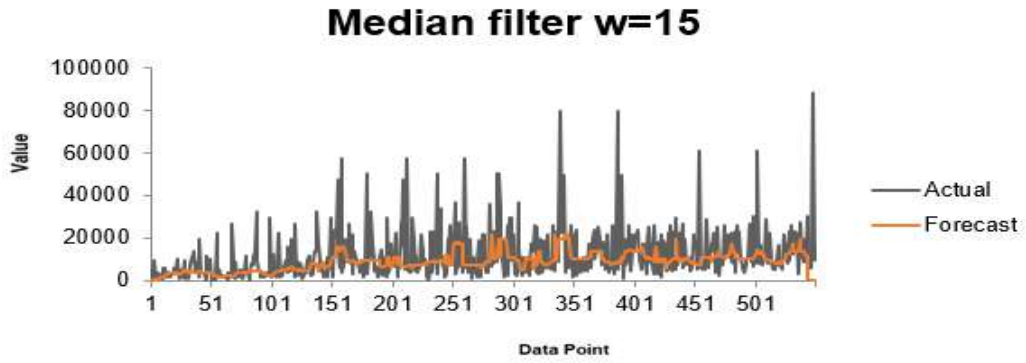


Figure 45: The result of the median filtering at w = 15

Correlation. We used the same sample of 100 elements in the correlation analysis [79-81] - ID and Reviews. First, by constructing a correlation field (Fig. 46), you can visually assess the nature of the connection.

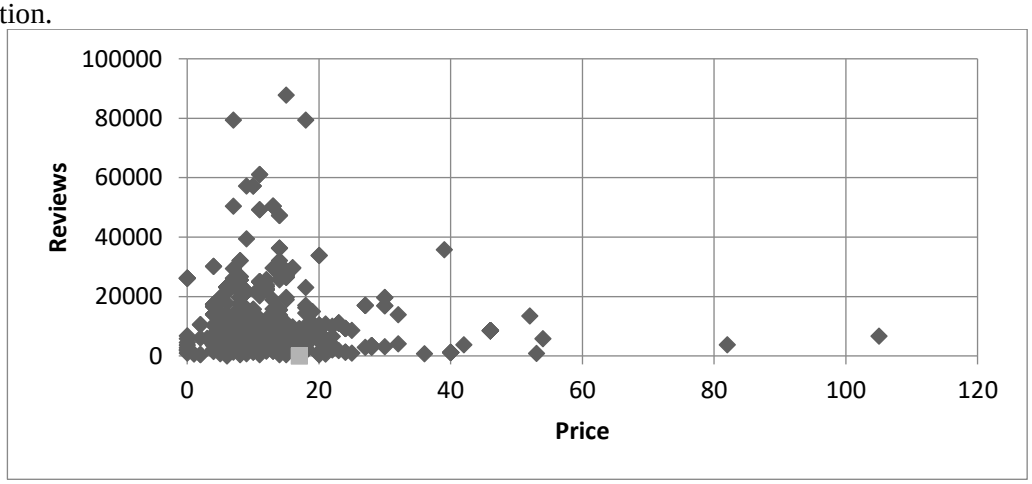
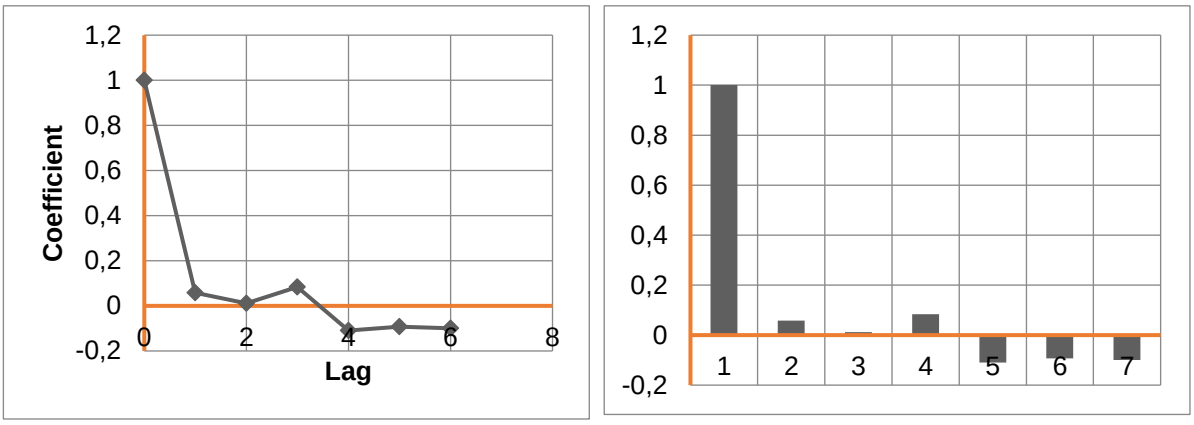


Figure 46: Visual assessment of the nature of communication

A sample correlation coefficient R is used to quantify the closeness of the connection. The sample correlation coefficient R in absolute value does not exceed one. When calculating the correlation coefficient, we obtained the following result: -0.057931 (Table 7). Graphs of autocorrelation functions were constructed (Fig. 47).

Table 7
The number of turning points and the correlation coefficient for each w

Lag	0	1	2	3	4	5	6
Correlation coefficient	1	0,05822	0,011919	0,084232	-0,10941	-0,09268	-0,09931



a) *b)*

Figure 47: The graph of autocorrelation functions (a), and choreography (b)

We divided our 550 elements into the intervals what is shown in Table 8.

Table 8
Dividing the sequence into three equal parts

Part	Part I	Part II	Part III
Interval	[1;183)	[183;366)	[366;549)
Number of sample items	182	182	182

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

We built a correlation matrix for them, which is shown in Table 9.

Table 9
Correlation matrix

Part	Part I	Part II	Part III
Part I	1		
Part II	0,030752	1	
Part III	-0,01009	-0,03414	1

Clustering. Cluster analysis is one of the methods of multidimensional statistical analysis, i.e., data analysis when each observation is presented not by one certain indicator but by a set of values of different indicators [82-101]. It includes several algorithms for forming the clusters themselves and distributing objects by clusters.

Cluster analysis, first of all, solves the problem of adding structure to the data, i.e., their group homogeneity, and provides a selection of compact, distant groups of objects, i.e., finds a "natural" breakdown of the population in the cluster of objects. We created a table "object-property" for cluster analysis, which is shown in Table 10.

Table 10
Object-property table

Authors who wrote the most bestsellers	Average rating	The average number of reviews	Average price
Bill O'Reilly	4,6333	9074,166667	10,6667
Charlaine Harris	4,4500	1633	10,2500
Dav Pilkey	4,9000	7376,833333	6,6667
E L James	4,3200	26149,2	15,6000
J.K. Rowling	4,5500	11470,5	23,8750
Jeff Kinney	4,8000	5623,5	9,5000
John Grisham	4,4000	12192,2	16,2000
Rick Riordan	4,7700	3954,1	10,2000
Stephenie Meyer	4,6571	6294	20,0000
Suzanne Collins	4,6750	24606,5	15,7500

The normalized table "object-property" shown in Table 11.

Table 11
The normalized object-property table

Authors who wrote the most bestsellers	Average rating	The average number of reviews	Average price
Bill O'Reilly	0,5402	0,3035	0,2324
Charlaine Harris	0,2241	0,0000	0,2082
Dav Pilkey	1,0000	0,2343	0,0000
E L James	0,0000	1,0000	0,5191
J.K. Rowling	0,3966	0,4013	1,0000
Jeff Kinney	0,8276	0,1628	0,1646
John Grisham	0,1379	0,4307	0,5540
Rick Riordan	0,7759	0,0947	0,2053
Stephenie Meyer	0,5813	0,1901	0,7748
Suzanne Collins	0,6121	0,9371	0,5278

They have constructed a proximity table (Table 12). For convenience, the names and surnames of the authors were replaced by a unique number from 1 to 10.

Table 12
The proximity tables

Authors who wrote the most bestsellers	1	2	3	4	5	6	7	8	9	10
1	0,000	0,439	0,520	0,927	0,787	0,327	0,530	0,316	0,556	0,703
2	0,439	0,000	0,837	1,071	0,904	0,627	0,559	0,560	0,696	1,063
3	0,520	0,837	0,000	1,362	1,180	0,249	1,043	0,334	0,882	0,961
4	0,927	1,071	1,362	0,000	0,864	1,229	0,587	1,233	1,029	0,615
5	0,787	0,904	1,180	0,864	0,000	0,970	0,516	0,932	0,360	0,746
6	0,327	0,627	0,249	1,229	0,970	0,000	0,836	0,095	0,659	0,882
7	0,530	0,559	1,043	0,587	0,516	0,836	0,000	0,801	0,551	0,694
8	0,316	0,560	0,334	1,233	0,932	0,095	0,801	0,000	0,609	0,917
9	0,556	0,696	0,882	1,029	0,360	0,659	0,551	0,609	0,000	0,787
10	0,703	1,063	0,961	0,615	0,746	0,882	0,694	0,917	0,787	0,000

The distance between objects was measured in the Euclidean metric (Table 12). It is also called the Euclidean distance and was calculated by the following formula.

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

where d_E is the Euclidean distance.

Carried out the association of clusters are presented in Table 13- Table 20.

Table 13

The merging 9 clusters

Authors	1	2	3	4	5	[6,8]	7	9	10
1	0,000	0,439	0,520	0,927	0,787	0,316	0,530	0,556	0,703
2	0,439	0,000	0,837	1,071	0,904	0,560	0,559	0,696	1,063
3	0,520	0,837	0,000	1,362	1,180	0,249	1,043	0,882	0,961
4	0,927	1,071	1,362	0,000	0,864	1,229	0,587	1,029	0,615
5	0,787	0,904	1,180	0,864	0,000	0,932	0,516	0,360	0,746
[6,8]	0,316	0,560	0,249	1,229	0,932	0,000	0,801	0,609	0,882
7	0,530	0,559	1,043	0,587	0,516	0,801	0,000	0,551	0,694
9	0,556	0,696	0,882	1,029	0,360	0,609	0,551	0,000	0,787
10	0,703	1,063	0,961	0,615	0,746	0,882	0,694	0,787	0,000

Table 14

The merging 8 clusters

Authors	1	2	[3,6,8]	4	5	7	9	10
1	0,000	0,439	0,520	0,927	0,787	0,316	0,530	0,556
2	0,439	0,000	0,837	1,071	0,904	0,560	0,559	0,696
[3,6,8]	0,520	0,837	0,000	1,362	1,180	0,249	1,043	0,882
4	0,927	1,071	1,362	0,000	0,864	1,229	0,587	1,029
5	0,787	0,904	1,180	0,864	0,000	0,932	0,516	0,360
7	0,316	0,560	0,249	1,229	0,932	0,000	0,801	0,609
9	0,530	0,559	1,043	0,587	0,516	0,801	0,000	0,551
10	0,556	0,696	0,882	1,029	0,360	0,609	0,551	0,000

Table 15

The merging 7 clusters

Authors	[1,3,6,8]	2	4	5	7	9	10
[1,3,6,8]	0,0000	0,4389	0,9269	0,7870	0,5305	0,5556	0,7027
2	0,4389	0,0000	1,0709	0,9042	0,5590	0,6962	1,0634
4	0,9269	1,0709	0,0000	0,8643	0,5868	1,0292	0,6154
5	0,7870	0,9042	0,8643	0,0000	0,5164	0,3597	0,7460
7	0,5305	0,5590	0,5868	0,5164	0,0000	0,5506	0,6942
9	0,5556	0,6962	1,0292	0,3597	0,5506	0,0000	0,7873
10	0,7027	1,0634	0,6154	0,7460	0,6942	0,7873	0,0000

Table 16
The merging 6 clusters

Authors	[1,3,6,8]	2	4	[5,9]	7	10
[1,3,6,8]	0,0000	0,4389	0,9269	0,5556	0,5305	0,7027
2	0,4389	0,0000	1,0709	0,6962	0,5590	1,0634
4	0,9269	1,0709	0,0000	0,8643	0,5868	0,6154
[5,9]	0,5556	0,6962	0,8643	0,0000	0,5164	0,7460
7	0,5305	0,5590	0,5868	0,5164	0,0000	0,6942
10	0,7027	1,0634	0,6154	0,7460	0,6942	0,0000

Table 17
The merging 5 clusters

Authors	[1,2,3,6,8]	4	[5,9]	7	10
[1,2,3,6,8]	0,0000	0,9269	0,5556	0,5305	0,7027
4	0,9269	0,0000	0,8643	0,5868	0,6154
[5,9]	0,5556	0,8643	0,0000	0,5164	0,7460
7	0,5305	0,5868	0,5164	0,0000	0,6942
10	0,7027	0,6154	0,7460	0,6942	0,0000

Table 18
The merging 4 clusters

Authors	[1,2,3,6,8]	4	[5,7,9]	10
[1,2,3,6,8]	0,0000	0,9269	0,5305	0,7027
4	0,9269	0,0000	0,5868	0,6154
[5,7,9]	0,5305	0,5868	0,0000	0,6942
10	0,7027	0,6154	0,6942	0,0000

Table 19
The merging 3 clusters

Authors	[1,2,3,5,7,6,8,9]	4	10
[1,2,3,5,7,6,8,9]	0,0000	0,5868	0,6942
4	0,5868	0,0000	0,6154
10	0,6942	0,6154	0,0000

Table 20
The merging 2 clusters

Authors	[1,2,3,4,5,7,6,8,9]	10
[1,2,3,4,5,7,6,8,9]	0,0000	0,6154
10	0,6154	0,0000

Procedure for merging clusters is presented in Table 21.

Visualization of the cluster analysis results is carried out using a dendrogram, i.e. a graphical representation of the results of sequential clustering, which is carried out in terms of the proximity matrix.

Using a dendrogram, you can graphically or geometrically represent the clustering procedure, provided that this procedure operates only with elements of the matrix of distances or similarities.

Table 22
The procedure for merging clusters

Steps	Merge	Node	Metric
	6+8	d11	0,094693412
2	3+14	d12	0,248898592
3	1+17	d13	0,316028251
4	5+9	d14	0,359741928
5	2+18	d15	0,438890676
6	7+14	d16	0,516402515
7	20+21	d17	0,530484628
8	4+41	d18	0,586804722
9	10+45	d19	0,615356846

Drawing horizontal lines in the plane of the dendrogram at a given height, in this case, allows you to select individual clusters. (Fig. 48)

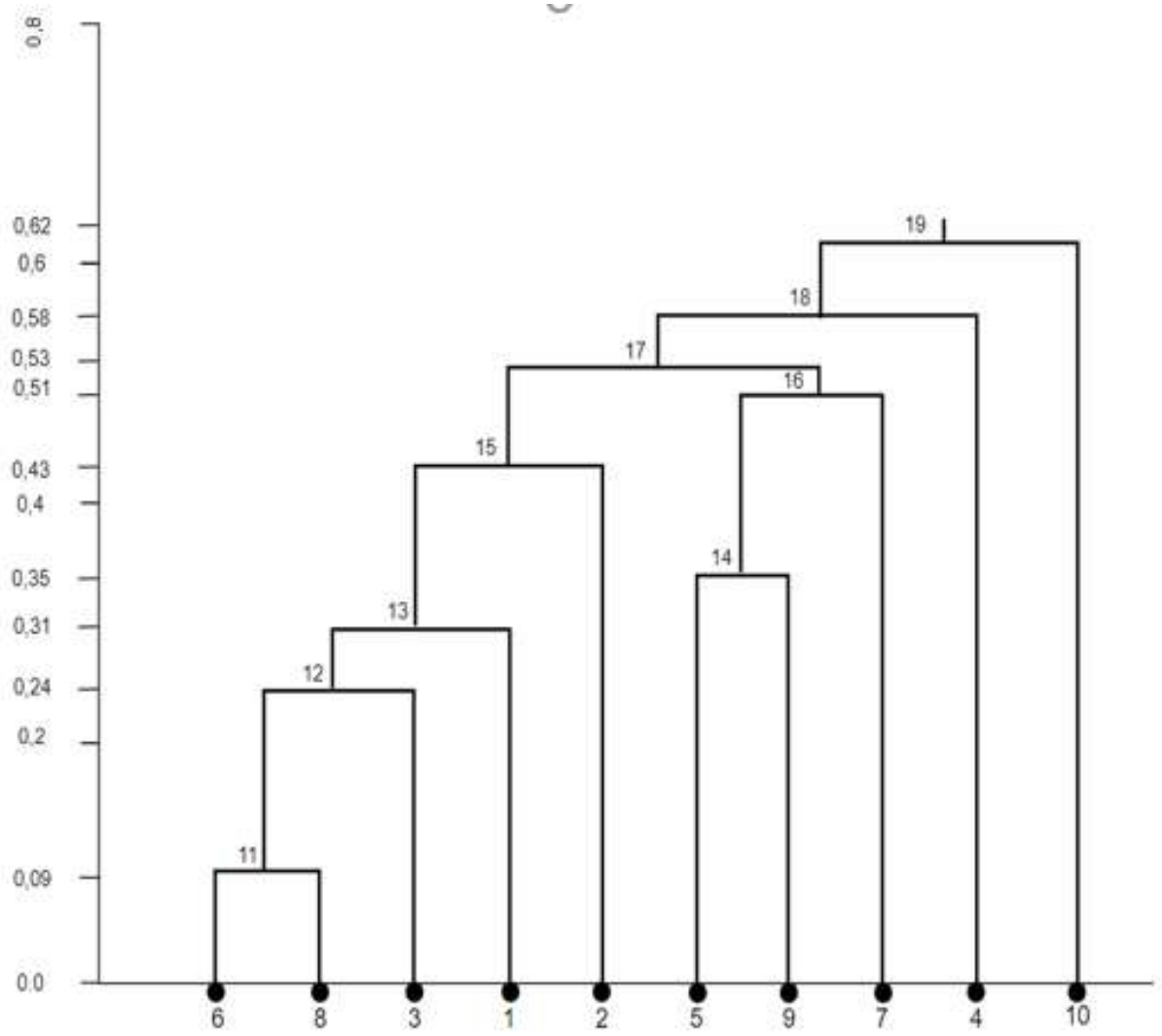


Figure 48: The dendrogram and its interpretation

Interpretation of the dendrogram is presented in Fig.48. Drawing horizontal lines in the plane of the dendrogram at a given height, in this case, allows you to select individual clusters (Fig.48). Analyzing

the constructed dendrogram, we can indicate that at the level of 0.4, we have, in principle, six fairly clearly displayed clusters, which include such objects.

- 1-st cluster - objects 6, 8, 3, 1
- 2-nd cluster - object 2
- 3-rd cluster - objects 5, 9
- 4-th cluster - object 7
- 5-th cluster - object 4
- 6-th cluster - object 10

At level 0.5 we have 5 clusters

- 1-st cluster - objects 6, 8, 3, 1, 2
- 2-nd cluster - objects 5, 9
- 3-rd cluster - object 7
- 4-th cluster - object 4
- 5-th cluster - object 10

At the level of 0.6 we have 2 clusters

- 1-st cluster - objects 6, 8, 3, 1, 2, 5, 9, 7, 4
- 2-nd cluster - object 10

7. Conclusions

A dataset of 550 elements was analyzed during the study, containing data on books: title, author, number of reviews, price, and rating.

By smoothing methods, we performed local averaging of data for another forecasting, in which non-systematic elements mutually replace each other. In this way, we get clear graphs without sharp peaks. Thus, we can conclude based on the moving average, weighted moving average, exponential smoothing, and median filtering graphs. The books that became bestsellers in 2009-2011 did not have a large response rate, mostly no more than 20,000, while in the following 2012-2019 there is an increase in the number of reviews. Their number often peaked at 40,000 reviews, with an average of 25,000 per book.

Examining the correlation between Reviews and Prices, we found a relationship that indicates that books priced above $ 60 typically have no more than 10,000 reviews. Books priced from $ 40 to $ 60 have an average of no more than 20,000 reviews. Books up to $ 20 have different reviews, but mostly this value does not exceed 40,000, and those books that have more are the exception rather than the rule.

The clustering method was applied to the 12 authors who wrote the bestsellers, and after normalizing the table and building proximity tables, we identified clusters. According to the constructed dendrogram, groups of clusters at levels 0.4, 0.5 and 0.6 can be seen. Interestingly, the ten objects are clustered only in the last iteration.

8. References

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

[2] M. Bublyk, Y. Matseliukh, Small-batteries utilization analysis based on mathematical statistics methods in challenges of circular economy, volume Vol-2870 of CEUR Workshop Proceedings, 2021, pp.1594-1603.

[3] O. Maslak, M. Maslak, N. Grishko, O. Hlazunova, P. Pererva, Y. Yakovenko, Artificial Intelligence as a Key Driver of Business Operations Transformation in the Conditions of the Digital Economy, 2021 IEEE International Conference on Modern Electrical and Energy Systems (MEES), 2021, https://doi.org/10.1109/MEES52427.2021.9598744.

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

[5] I. Jonek-Kowalska, Towards the Reduction of CO2 Emissions. Paths of Pro-Ecological Transformation of Energy Mixes in European Countries with an Above-Average Share of Coal in Energy Consumption. Resources Policy, vol. 77, 2022. doi:10.1016/j.resourpol.2022.102701.
[6] M. Bublyk , A. Kowalska-Styczen , V. Lytvyn , V. Vysotska , The Ukrainian Economy Transformation into the Circular Based he Fuzzy-Logic Cluster Analysis . Energies 2021, 14, 5951. https://doi.org/10.3390/en14185951
[7] 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 he machine learning , volume Vol-2631 of CEUR workshop proceedings , 2020, pp . 172-183.
[8] M. Bublyk , Y. Matseliukh , U. Motorniuk , M. Terebukh , Intelligent system of passenger transportation by autopiloted electric buses in Smart City , volume Vol-2604 of CEUR workshop proceedings , 2020, pp . 1280-1294.
[9] 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 , volume Vol-2631 of CEUR workshop proceedings, 2020, pp . 491-501.
[10] I. Rishnyak , O. Veres , V. Lytvyn , M. Bublyk , I. Karpov , V. Vysotska , V. Panasyuk , Implementation models application for IT project risk management , volume Vol-2805 of CEUR Workshop Proceedings, 2020, pp . 102-117.
[11] Amazon's books: [EDA/Plotly/hypothesis test], 2022. URL: https://www.kaggle.com/code/ivannatarov/amazon-s-books-eda-plotly-hypothesis-test/notebook.
[12] 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 2020 IEEE 3rd International Conference he Data Stream Mining and Processing , DSMP, 2020, pp . 432-437.
[13] 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 he Computer Sciences and Information Technologies, CSIT, 1, 2020, pp . 114-118.
[14] D. Ivanchyshyn, V. Vysotska, S. Albota, The Film Script Generation Analysis Based on the Fiction Book Text Using Machine Learning, in: Proceedings of the IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 22-25 Sept., Lviv, Ukraine. 2021, Vol. 2, pp. 68–80.
[15] O. Hladun, A. Berko, M. Bublyk, L. Chyrun, V. Schuchmann, Intelligent system for film script formation based on artbook text and Big Data analysis, in: Proceedings of the IEEE 16th International conference on computer science and information technologies on Computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 138–146.
[16] Y. Bobalo, P. Stakhiv, B. Mandziy, N. Shakhovska, R. Holoschuk, The concept of electronic textbook "Fundamentals of theory of electronic circuits", Przeglad Elektrotechniczny 88(3 A) (2012) 16-18.
[17] O. Sichevska, V. Senkivskyy, S. Babichev, O. Khamula, Information technology of forming the quality of art and technical design of books, CEUR Workshop Proceedings 2533 (2019) 45–57.
[18] N. Grabar, C. Grouin, A Year of Papers Using Biomedical Texts: Findings from the Section on Natural Language Processing of the IMIA Yearbook, Yearbook of medical informatics **28**(1) (2019) 218–222.
[19] V. Lytvyn, V. Vysotska, I. Budz, Y. Pelekh, N. Sokulska, R. Kovalchuk, L. Dzyubyk, O. Tereshchuk, M. Komar, Development of the quantitative method for automated text content authorship attribution based on the statistical analysis of N-grams distribution, Eastern-European Journal of Enterprise Technologies **6**(2-102) (2019) 28-51. doi: 10.15587/1729-4061.2019.186834
[20] B. E. Kapustiy, B. P. Rusyn, V. A. Tayanov, Peculiarities of application of statistical detection criteria for problems of pattern recognition, Journal of Automatioin and Inrormation Science **37**(2), (2005) 30-36.
[21] N. Lototska, Statistical Research of the Colour Component ЧОРНИЙ (BLACK) in R. Ivanychuk's Text Corpus, CEUR Workshop Proceedings Vol-2870 (2021) 486-497.

[22] T. Shestakevych, Modeling the Process of Analysis of Statistical Characteristics of Student Digital Text, CEUR Workshop Proceedings Vol-2870 (2021) 657-669.
[23] A. Hadzalo, Analysis of Gender-Marked Units: Statistical Approach, CEUR workshop proceedings Vol-2604 (2020) 462-471.
[24] I. Khomytska, V. Teslyuk, Statistical Models for Authorship Attribution, Advances in Intelligent Systems and Computing, 2019.
[25] I. Khomytska, V. Teslyuk, Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level, Advances in Intelligent Systems and Computing 871 (2019) 105–118. doi: 10.1007/978-3-030-01069-0_8.
[26] I. Khomytska, V. Teslyuk, The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level. In ebook : Advances in Intelligent Systems and Computing 512 (2016) 149–163.
[27] I. Kulchytskyi, Statistical Analysis of the Short Stories by Roman Ivanychuk, CEUR Workshop Proceedings Vol-2362 (2019) 312-321.
[28] Amazon Top 50 Bestselling Books 2009-2019, 2020. URL: https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019/tasks?taskId=2741.
[29] EDA: Amazon Top 50 Bestselling Books, 2022. URL: https://www.kaggle.com/code/japandata509/eda-amazon-top-50-bestselling-books/notebook.
[30] Amazon Top 50 Bestselling Books 2009-2019, 2020. URL:https://www.kaggle.com/code/aryan27/ amazon-top-50-bestselling-books-2009-2019/notebook
[31] V. Lytvyn, A. Hryhorovych, V. Hryhorovych, V. Vysotska, M. Bublyk, L. Chyrun, Medical content processing in intelligent system of district therapist, volume Vol-2753 of CEUR workshop proceedings, 2020, pp. 415–429. http://ceur-ws.org/Vol-2753/paper29.pdf
[32] M. Baran, O. Kuzmin, M. Bublyk, V. Panasyuk, K. Lishchynska, Information system for quality control of polyethylene production in a circular economy, volume Vol-2917 of CEUR Workshop Proceedings, 2021, pp. 465–502. http://ceur-ws.org/Vol-2917/paper34.pdf
[33] Standard error, 2022. URL: https://ua.nesrakonk.ru/standard-error/.
[34] Standard deviation, 2022. URL: https://studopedia.su/10_11382_standartne-vidhilennya.html.
[35] 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.
[36] Grouping of statistical data - BukLib.net Library, 2022. URL: https://buklib.net/books/35946/
[37] Graphic presentation of information, 2022. URL: https://studopedia.com.ua/1_132145_grafichne-podannya-informatsii.html.
[38] 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.
[39] Forecasting the trend of the time series by algorithmic methods, 2022. URL: http://ubooks.com.ua/books/000269/inx42.php.
[40] K.O. Soroka, Fundamentals of Systems Theory and Systems Analysis, Kharkiv, 2004.
[41] Yu.P. Surmin, Systems theory and system analysis, Kyiv, 2003.
[42] I.V. Stetsenko, Systems modeling, Cherkasy, 2010.
[43] S.S. Velykodnyi, Modeling of systems, Odessa, 2018.
[44] A. Agresti, Analysis of Ordinal Categorical Data, John Wiley & Sons, 1984.
[45] S. Babichev, V. Lytvynenko, A. Gozhyj, M. Korobchynskyi, M. Voronenko, A fuzzy model for gene expression profiles reducing based on the complex use of statistical criteria and Shannon entropy, Advances in Intelligent Systems and Computing 754 (2018) 545-554.
[46] P. Bidyuk, I. Kalinina, A. Gozhyj, Methodology of constructing statistical models for nonlinear non-stationary processes in medical diagnostic systems, CEUR Workshop Proceedings 2753 (2020) 36–45.
[47] R. Kaminskyi, N. Kunanets, A. Rzheuskyi, Mathematical support for statistical research based on informational technologies, CEUR Workshop Proceedings 2105 (2018) 449-452.
[48] I. Gorbenko, A. Kuznetsov, Y. Gorbenko, S. Vdovenko, V. Tymchenko, M. Lutsenko, Studies on Statistical Analysis And Performance Evaluation For Some Stream Ciphers. International Journal of Computing **18**(1) (2019) 82-88.

[49] Y. Butelskyy, Statistical Methods to Detect Gender Peculiarities of Communication in Vkontakte Social Network Groups, in: Proceedings of the 11th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2016, pp. 132-135. doi: 10.1109/STC-CSIT.2016.7589888.
[50] I. Sokolovskyy, N. Shakhovska, Statistical modeling of diffusion processes with a fractal structure, CEUR Workshop Proceedings 2488 (2019) 145–154.
[51] L. Chyrun, The E-Commerce Systems Modelling Based on Petri Networks, CEUR Workshop Proceedings Vol-2870 (2021) 1604-1631.
[52] O. Prokipchuk, L. Chyrun, M. Bublyk, V. Panasyuk, V. Yakimtsov, R. Kovalchuk, Intelligent system for checking the authenticity of goods based on blockchain technology, volume Vol-2917 of CEUR Workshop Proceedings, 2021, pp. 618-665.
[53] A. Berko, V. Andrunyk, L. Chyrun, M. Sorokovskyy, O. Oborska, O. Oryshchyn, M. Luchkevych, O. Brodovska, The Content Analysis Method for the Information Resources Formation in Electronic Content Commerce Systems, CEUR Workshop Proceedings 2870 (2021) 1632-1651.
[54] M. Bublyk, V. Mykhailov, Y. Matseliukh, T. Pihniak, A. Selskyi, I. Grybyk, Change management in R&D-quality costs in challenges of the global economy, volume Vol-2870 of CEUR Workshop Proceedings, 2021, pp. 1139–1151
[55] O. Bisikalo, O. Danilchuk, V. Kovtun, O. Kovtun, O. Nikitenko, V. Vysotska, Modeling of operation of information system for critical use in the conditions of influence of a complex certain negative factor, International Journal of Control, Automation and Systems (2022). https://doi.org/10.1007/s12555-021-0368-6
[56] V. Kuchkovskiy, V. Andrunyk, M. Krylyshyn, L. Chyrun, A. Vysotskyi, S. Chyrun, N. Sokulska, I. Brodovska, Application of Online Marketing Methods and SEO Technologies for Web Resources Analysis within the Region, CEUR Workshop Proceedings Vol-2870 (2021) 1652-1693.
[57] 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.
[58] 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.
[59] A. Gozhyj, I. Kalinina, V. Nechakhin, V. Gozhyj, V. Vysotska, Modeling an Intelligent Solar Power Plant Control System Using Colored Petri Nets, in: proceedings of the IEEE 11th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 22-25 Sept., Cracow, Poland, 2021, pp. 626–631.
[60] A. Gozhyj, I. Kalinina, V. Vysotska, V. Gozhyj, The method of web-resources management under conditions of uncertainty based on fuzzy logic, in: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 2018, pp. 343-346. doi: 10.1109/STC-CSIT.2018.8526761.
[61] A. Berko, Y. Matseliukh, Y. Ivaniv, L. Chyrun, V. Schuchmann, The text classification based on Big Data analysis for keyword definition using stemming, in: proceedings of IEEE 16th International conference on computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 184–188.
[62] A. Dyriv, V. Andrunyk, Y. Burov, I. Karpov, L. Chyrun, The user's psychological state identification based on Big Data analysis for person's electronic diary, in: Proceedings of IEEE 16th International conference on computer science and information technologies,Lviv, Ukraine, 22–25 September, 2021, pp. 101–112.
[63] V. Kiyko, V. Lytvyn, L. Chyrun, S. Vyshemyrska, I. Lurie, M. Hrubel, Forest cover type classification based on environment characteristics and machine learning technology, Communications in Computer and Information Science 1158 (2020) 501–524. doi: 10.1007/978-3-030-61656-4_34.
[64] Y. Tverdokhlib, V. Andrunyk, L. Chyrun, L. Chyrun, N. Antonyuk, I. Dyyak, O. Naum, D. Uhryn, V. Basto-Fernandes, Analysis and estimation of popular places in online tourism based on machine learning technology, CEUR Workshop Proceedings 2631 (2020) 457–470.

[65] D. Uhryn, V. Andrunyk, L. Chyrun, N. Antonyuk, I. Dyyak, O. Naum, Service-oriented architecture development as an integrating platform in the tourist area, CEUR Workshop Proceedings 2631 (2020) 221–236.
[66] R. Levus, A.Berko, L. Chyrun, V. Panasyuk, M. Hrubel, Intelligent System for Arbitrage Situations Searching in the Cryptocurrency Market, CEUR Workshop Proceedings Vol-2917 (2021) 407-440.
[67] V. Husak, L. Chyrun, Y. Matseliukh, A. Gozhyj, R. Nanivskyi, M. Luchko, Intelligent Real-Time Vehicle Tracking Information System, CEUR Workshop Proceedings Vol-2917 (2021) 666-698.
[68] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska, Uniform method of operative content management in web systems, CEUR Workshop Proceedings 2136 (2018) 62-77.
[69] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, Web resource changes monitoring system development, CEUR Workshop Proceedings 2386 (2019) 255-273.
[70] L. Chyrun, A. Kowalska-Styczen, Y. Burov, A. Berko, A. Vasevych, I. Pelekh, Y. Ryshkovets, Heterogeneous data with agreed content aggregation system development, volume 2386 of CEUR Workshop Proceedings, 2019, pp. 35-54.
[71] L. Chyrun, A. Gozhyj, I. Yevseyeva, D. Dosyn, V. Tyhonov, M. Zakharchuk, Web Content Monitoring System Development, CEUR Workshop Proceedings Vol-2362 (2019) 126-142.
[72] Y. Kis, L. Chyrun, T. Tsymbaliak, L. Chyrun, Development of System for Managers Relationship Management with Customers, Advances in Intelligent Systems and Computing 1020 (2020) 405-421. doi: 10.1007 / 978-3-030-26474-1_29
[73] L. Chyrun, I. Turok, I. Dyyak, Information model of the tendering system for large projects, CEUR Workshop Proceedings 2604 (2020) 1224-1236.
[74] 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, volume Vol-2631 of CEUR Workshop Proceedings, 2020, pp. 502 - 515.
[75] O. Garasym, L. Chyrun, N. Chernovol, A. Gozhyj, V. Gozhyj, I., Kalinina B., Rusyn, L. Pohreliuk, M. Korobchynskyi, Network security analysis based on consolidated threat resources, CEUR Workshop Proceedings 2604 (2020) 1004-1018.
[76] L. Chyrun, P. Kravets, O. Garasym, A. Gozhyj, I. Kalinina, Cryptographic information protection algorithm selection optimization for electronic governance IT project management by the analytical hierarchy process based on nonlinear conclusion criteria, CEUR Workshop Proceedings 2565 (2020) 205-220.
[77] I. Pelekh, A. Berko, V. Andrunyk, L. Chyrun, I. Dyyak, Design of a system for dynamic integration of weakly structured data based on mash-up technology, in: Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP, 2020, pp. 420-425. doi: 10.1109 / DSMP47368.2020.9204160.
[78] A. Berko, I. Pelekh, L. Chyrun, I. Dyyak, Information resources analysis system of dynamic integration semi-structured data in a web environment, in: Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP, 2020, pp. 414-419. doi: 10.1109 / DSMP47368.2020.9204101.
[79] 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.
[80] 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.
[81] 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, pp. 219–220.
[82] P. Kravets, Y. Burov, V. Lytvyn, V. Vysotska, Gaming method of ontology clusterization Webology **16**(1) (2019) 55-76.
[83] 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.

[84] Y. Matseliukh, V. Vysotska, M. Bublyk, T. Kopach, O. Korolenko, Network modelling of resource consumption intensities in human capital management in digital business enterprises by the critical path method, volume Vol-2851 of CEUR Workshop Proceedings, 2021, pp. 366–380.
[85] 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.
[86] 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.
[87] A. Dovbysh, V. Piatachenko, Hierarchical Clustering Approach for Information-Extreme Machine Learning of Hand Brush Prosthesis, CEUR Workshop Proceedings Vol-2870 (2021) 1706-1715.
[88] N. Boyko, S. Hetman, I. Kots, Comparison of Clustering Algorithms for Revenue and Cost Analysis, CEUR Workshop Proceedings Vol-2870 (2021) 1866-1877.
[89] 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.
[90] 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). https://doi.org/10.1007/978-3-030-26474-1_37
[91] 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.
[92] M. Bublyk, V. Vysotska, L. Chyrun, V. Panasyuk, O. Brodyak, Assessing security risks method in E-commerce system for IT portfolio management, volume Vol-2853 of CEUR Workshop Proceedings, 2021, pp. 362–379.
[93] 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.
[94] 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-238, (2019) 117-135.
[95] 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 12391 (2020) 414–424.
[96] S., Mashtalir, O., Mikhnova, M. Stolbovyi, Multidimensional Sequence Clustering with Adaptive Iterative Dynamic Time Warping. International Journal of Computing **18**(1), (2019) 53-59.
[97] 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.
[98] 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.
[99] A. Demchuk, B. Rusyn, L. Pohreliuk, A. Gozhyj, I. Kalinina, L. Chyrun, N. Antonyuk, Commercial content distribution system based on neural network and machine learning, CEUR Workshop Proceedings 2516 (2019) 40-57.
[100] 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.
[101] 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, 1, 2017, pp. 479-484.