Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles

Victoria Vysotska[1,2], Oksana Markiv[1], Sofiia Teslia[1], Yeva Romanova[1]and Inesa Pihulechko[1]

[1]*Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2]*Osnabrück University, Friedrich-Janssen-Str. 1, Osnabrück, 49076, Germany*

Abstract

The results of experimental approbation of the proposed content monitoring method used for the determination of the author style in Ukrainian scientific texts of technical profile have been studied. Authorship identification systems typically use plagiarism and rewrite metrics to determine it. There is a necessity to identify whether the work has been borrowed fully or partially. Therefore, the situation when the work has not been published yet is not taken into consideration. Quantitative content analysis of the scientific and technical texts uses the advantages of content monitoring and analysis of text based on NLP, Web-Mining and stylometry methods to identify many authors whose speech styles are similar to the studied passages. It narrows the search for further use in stylometric methods to determine the degree of the analyzed text belonging to a particular author. The method of determining the author has been decomposed on the basis of such speech coefficients analysis as lexical diversity, degree (measure) of syntactic complexity, speech coherence, indices of the text exclusivity and concentration. In parallel, the parameters of the author style, such as the text words, sentences, prepositions, conjunction quantities and the number of words with a frequency of 1, 10 or more have been analyzed.

Keywords 1

N-Grams, NLP, correlation analysis, authorship definition, Ukrainian text, distribution function density, exponential and median smoothing, linguometry, stylometric analysis

1. Introduction

Due to the increasing availability and distribution of text documents in electronic form the importance of using automatic methods to analyze the content of documents has been increased [1-3]. The tasks of text analysis include the necessity of documents classification and clustering [4-7] by various criteria, such as genre, writing format (novel, essay), emotional coloring, speech style, as well as the task of text author identification [8 -14].

With the simplification of access to various data, growth of the ability to search, copy and distribute data on networks, the task of identifying the author becomes urgent. Issues related to the determination of authorship are also important in linguistic, historical and forensic researches. The general availability of electronic devices allows to push the recognition of the author with the involvement of a large number of experts in the background, speed up and simplify this process through its automation.

The concept of author identification is defined as the process of author identification based on the set of the text general and particular features that constitute the author style [8-9].

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: victoria.a.vysotska@lpnu.ua (V. Vysotska); oksana.o.markiv@lpnu.ua (O. Markiv); sofiia.teslia.sa.2019@lpnu.ua (S. Teslia); yeva.romanova.sa.2019@lpnu.ua (Y. Romanova); inesa.pihulechko.sa.2019@lpnu.ua (I. Pihulechko)
ORCID: 0000-0001-6417-3689 (V. Vysotska); 0000-0002-1691-1357 (O. Markiv); 0000-0002-2591-2431 (S. Teslia); 0000-0003-0522-0806 (Y. Romanova); 0000-0003-2789-2902 (I. Pihulechko)

2. Related Works

Statistical methods based on the search for "author invariant" are popular in existing systems for determining the text authorship. "Author invariant" characterizes the text linguistic features (lexical, grammatical, phraseological and other ones). The invariant can be the following: the share of vowels or consonants, the frequency of certain part of speech use, the probability of transitions from one part of speech to another, "favorite" words, information entropy etc. Authors proposed a statistical method for determining the text author and genre based on the frequency distribution of letter combinations (n-grams) [10-12]. This method has shown decent results for works of Slavic research publications. Unfortunately, the accuracy of determining authorship statistical methods depends on the data specifics using the language, style and length of written texts that have been studied [13-28]. Because of this, it is difficult to conclude the accuracy of such an approach to data of a different nature. For this reason, the aim of this work is to analyze the application of such a mathematical apparatus as the distribution of letter combinations for different languages in solving the problem of establishing the texts authorship of different lengths and written in different language styles. Chosen topic, namely relative frequency n-grams, is only gaining popularity in Ukraine and is not very popular. Several literature sources to describe what n-grams are and what they are used for have been found.

N-gram is a sequence of n-elements [29]. From a semantic point of view, it can be a sequence of sounds, syllables, words or letters. In practice, N-grams are more common as a series of words, stable phrases that called collocations. A sequence of two consecutive elements is often called a bigram, a sequence of three elements is called a trigram, which have been presented in the studied dataset. At least four or more elements are denoted as N-grams, N is replaced by the number of consecutive elements. N-grams in general are used in a wide range of sciences. They can be used, for example, in the field of theoretical mathematics, biology, cartography, as well as in music. The most common uses of N-grams include the following: extracting data for the cluster of satellite images series of the Earth from space, decision which specific parts of the Earth are in the image, and searching for genetic sequences in computer compression for indexing data in search engines, using N-grams, usually indexed data related to sound. In natural language processing N-gram is used mainly for prediction based on probabilistic models. The N-gram model calculates the probability of the last word of the N-gram, if all the previous ones are known. When using this approach to modeling language, it is assumed that the appearance of each word depends only on previous words [30]. Another application of N-grams is the detection of plagiarism. If you divide the text into several small fragments represented by N-grams, they are easy to compare with each other, and thus obtain a degree of similarity of controlled documents [31]. N-grams are often used successfully to categorize text and language. In addition, they can be used to create functions that allow you to gain knowledge from textual data. Using N-grams, you can effectively find candidates to replace words with spelling mistakes. Google Research Centers have used N-gram models for a wide range of research and development. These include projects such as statistical translation from one language to another, language recognition, spelling correction, information retrieval, and more. For the purposes of these projects were used text corpora, which contain several trillion words. Google has decided to create its own educational building. The project is called Google tera corpus and it contains 1,024,908,267,229 words collected from public websites [32].

For a long time, cryptograms decryption is aided by frequency analysis the essence of which is the study of statistical patterns of symbols appearance and their compounds in original and encrypted messages [1]. In order to complicate frequency analysis ciphers have appeared in cryptography what leads to a uniform distribution of characters in the cryptogram. The principles of frequency analysis are widely used in password programs and allow to reduce the search time by several orders of magnitude [2] based on classification and clustering [4-7] of documents by various criteria, such as genre, epoch, format (novel, essay), emotional coloring, speech style, as well as the task of determining the text author [8-15].Obviously, frequency analysis requires first of all the reference frequencies of alphabet letters repetition on which the open texts are written and frequencies of N-grams repetition. For Ukrainian, English and almost all European languages the average frequency of letters, bigrams, trigrams repetition can be found in the literature [16-22].

Unfortunately, for the Ukrainian language only the frequency of letters repetition is given in the literature [23-25]. Therefore, the purpose of this work is to investigate the repetition frequency of letters and letters of the Ukrainian language on the basis of randomly selected texts in the Ukrainian language of scientific and technical orientation. The analysis of the obtained data confirms that for the Ukrainian language, as well as for other European languages, the alternation of vowels and consonants is inherent. If you study other texts, there may differ in the numbers of the given letter's frequencies, which is explained, firstly, by the length of the studied text, and, secondly, by its subject matter. For example, the generally used letter *F* can become quite common in technical texts, because it is used in such words as function, differential, diffusion, coefficient, etc. Even greater deviations from the traditional use of individual letters are observed in some works of art, especially in poems.

3. Methods and Materials

Modern systems for determining the text authorship use different approaches to the theory of mathematical statistics, pattern recognition and probability theory, cluster analysis algorithms, neural networks and others [33-45]. The systems differ in the method of author identification, the means of text analysis, the required text amount and accuracy [12]. Methods of text authorship identification based on the calculation of any text characteristics (official parts of speech, prepositions, conjunctions, particles, independent parts of speech, nouns, verbs, adjectives, word lengths, sentence lengths) also differ in comparing frequencies in the different textual content for different tasks [46-63]. The most commonly used measures of comparing texts are the following: Information entropy, Fisher information, Chi-squared test and Kullback-Leibler divergence [9-12].

When identifying the author of the text it is assumed that the text reflects the individual style of the text author, which allows to differ it from other ones. To compare the texts with each other it is necessary to compare the text with some numerical characteristic that was close to the texts of the same author and would different in the works of various authors. Such a characteristic of author in the article [9-12] uses the distribution function density (DFD) of letter combinations of three consecutive characters (3-grams). DFD is defined as the set of empirical frequencies of birth of letters or their combinations. The analysis of the text with the help of DFD does not take into account the occurrence of punctuation marks, spaces and numbers.

The task of identifying the author of the unknown text in terms of DFD is formulated as follows.

Here is a set of texts that contain works of famous authors. Let L_t be the number of works by the *a*-th author. $N_{i,t}$ – the number of symbols in the *i*-th work of the *a*-th author, $i = 1, \ldots, L_t$. All texts in this set will be presented in the form of DFD. DFD of the text, the volume of which is equal to $N_{i,t}$, is given as the set of values $p_{i,t}(j) = l_j / N_{i,t}$, l_j– the number of N-grams under the number j. The argument $j = 1, \ldots, f(n, M)$corresponds to the number of letters (n-grams) in alphabetical order, where M is the power of the alphabet of the language in which the text is written, n is the order of N-grams, i.e. the number of characters in letter combination. $f(n, M) = M^n$ is the number of N-grams in this alphabet.

Each author is identified with his weighted average DFD which is given by formula (1):

$$P_t = \frac{1}{N_t} \sum_{i=1}^{L_t} p_{i,t} N_{i,t}, N_t = \sum_{i=1}^{L_t} N_{i,t} \tag{1}$$

These DFD will play the role of copyright standards [9-12]. To compare two texts, either the text and the author standard, it is needed to specify the distance between the corresponding distribution functions. The norm in the space of summed functions is used as a distance metric. For example, the distance $w_{0,t}$ between the DFD of the unknown text p_0 and any copyright DFD will be calculated by formula (2):

$$w_{0,t} = \|p_0 - P_t\| = \sum_{j=1}^{f(n,M)} |p_0(j) - P_t(j)|, \tag{2}$$

Accordingly, the text «0» will belong to the author whose distance to the DFD will be the shortest.

When solving the problem of classification, the data set was not clearly divided into test and training sets. Weighted average DFD were built on the whole set of books by one author. The distance from the book i to "his" a -author is calculated by formula (3):

$$w_{i,t} = \frac{\|p_{i,t} - P_t\|}{1 - N_{i,t}/N_t}. \tag{3}$$

Formula (4) excludes the participation of the DFD of the document / i-article in the average DFD of "its" author [9-12]. The method of smoothing 3-gram distribution functions according to the analytical approach is impossible because the function is too complex. There is only an algorithmic approach for the implementation of which we can be focused on the main methods, such as simple or ordinary moving average, weighted moving average, exponential smoothing, median smoothing. In our case, we believe that the most optimal will be the use of the moving average method and this method is also known as the filtering method. Its application will reduce the variety of data. This fits into our analytically chosen tactics of ignoring extreme data, highs and lows. The degree of smoothing should be tied in advance to the criterion that will ensure maximum smoothing while still retaining information.

In our specific case we believe that correlation analysis of 3-gram data sequences in certain two of three selected articles can help to determine the relationship, and thus help to answer the question of how similar is the topic of articles. To do this, the function of the first studied article can be denoted by the variable *x* and the set of values of the second article (variable *y*) and perform a correlation analysis of the set of two sequences XY. The task of correlation is not only to assess connectivity, but also to reduce the target score to a numerical expression. The method of studying 3-gram sequences allows to reduce significantly the number of variables that are taken into account as important ones. Combination of the metrics-related groups forms a new cluster, which compares the metrics of closeness with others, and it is possible to end up with a fairly clear structure of the data set. Quantitative method of the potential text author identification from the set of possible ones on the basis of comparison analysis results of the reference text with the researched one is based on the technique of linguometry.

Linguometry is a branch of applied linguistics that detects, measures and analyzes the quantitative characteristics of different levels units of language or speech [33]. Using the apparatus of mathematical statistics, linguometry is involved in solving such problems of linguistics as the following criteria:

- dictionaries (including frequency and statistical) and comparisons
- automatic dictionaries, thesauri
- shorthand systems
- methods and means of automatic language detection
- methods and means of information retrieval, etc.

Each language has its own statistical parameters and knowledge of the frequency occurrence of letters and their combinations (2-gram, 3-gram, 4-gram) that allows automatically to identify it. For example, for Ukrainian texts [34-37] it was found that statistical parameters of styles can consider frequencies of vowels, consonants, spaces between words, as well as soft and sonorous groups of consonants [33]. We will show how to evaluate the speech of a particular author on a particular passage of his work [36] using a certain standard, for example, Ukrainian language letters frequencies. Consider two passages of the technical text in Ukrainian presented in a format where the letters are arranged in descending order of frequency of their appearance (frequencies are given in Table 1), the distinction between lowercase and uppercase letters has not been made. The type of letters correlation frequencies of the passages [35] and the standard [36] have been investigated. The results that confirm the conclusions have been presented, in particular, graphically.

Table 1

Frequencies of letters appearance in the standard and the studied passages

Letter	Frequency of use of Ukrainian language letters	The absolute frequency of the letters in Passage 1	The absolute frequency of letters in Passage 2	The relative frequency of letters uses in Passage 1	The relative frequency of letters use in Passage 2
ф	0.003	1	0	0.00	0.00
щ	0.004	3	1	0.01	0.00

ш	0.005	3	2	0.01	0.00
ж	0.007	3	7	0.01	0.01
й	0.007	4	6	0.01	0.01
ц	0.009	7	1	0.01	0.00
х	0.010	4	7	0.01	0.01
б	0.011	7	5	0.01	0.01
г	0.012	4	6	0.01	0.01
ю	0.012	2	2	0.00	0.00
ч	0.015	5	11	0.01	0.02
б	0.016	7	5	0.01	0.01
з	0.018	9	8	0.02	0.01
я	0.024	15	6	0.03	0.01
у	0.025	19	14	0.03	0.03
п	0.025	11	21	0.02	0.04
л	0.028	17	30	0.03	0.06
д	0.028	16	4	0.03	0.01
м	0.031	10	13	0.02	0.02
к	0.031	22	20	0.04	0.04
с	0.033	22	27	0.04	0.05
р	0.036	15	16	0.03	0.03
е	0.038	26	45	0.05	0.08
т	0.046	25	20	0.04	0.04
в	0.047	29	19	0.05	0.04
и	0.054	27	27	0.05	0.05
others	0.0605	51	34	0.09	0.06
н	0.068	33	30	0.06	0.06
а	0.074	43	31	0.08	0.06
о	0.082	37	41	0.07	0.08
« »	0.133	80	82	0.14	0.15

In the table. 1 the following data are entered for convenience: frequency of used Ukrainian language letters, absolute and relative frequencies of letters used in the studied Passage 1 (Article 1) [35] and Passage 2 (Article 2) [36]. Passage 1 contains 556 characters; Passage 2 contains 541 characters. The concept of "other" in the column of letters contains authentic letters for the Ukrainian language (ї, є, г, і), which are rarely used in most technical texts. This allows to achieve some independence in the analysis. Fig. 1 illustrates the obtained results graphically.

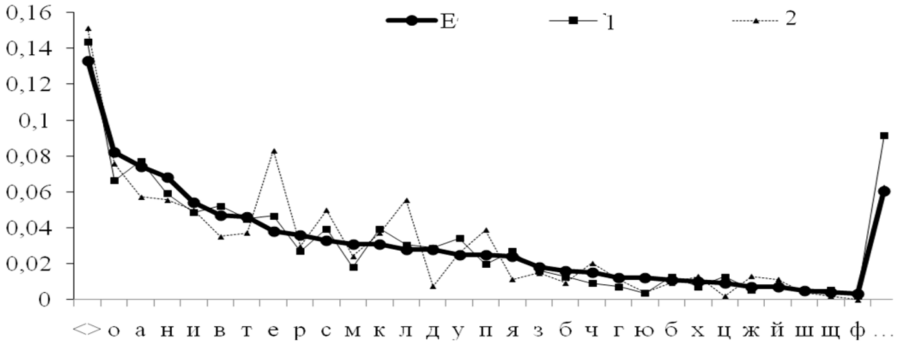


Figure1: The relative frequencies of letters in the standard and the studied passages

Graphical representation of the relative frequencies of letters in the passages gives a convincing answer to the question which of the passages was written by which author.

The distribution of 1-gram in the works is different. The optimal indicators of the texts study are the analysis of 3-grams [38-44]. We will check this in the next stages of the study. There is a sharp jump in the relative frequency of occurrence of the letter "e" for Passage 2 relative to the reference values of Standard 1 [36] (Fig. 2), so we assume that it is more likely that Standard 1 was written by the author of Passage 1 [35]. We also give the numerical values of the correlation of the frequency of letters in the passages and the standard. We find two correlation coefficients: for the standard and Passage 1 [35] and for the standard and Passage 2 [37]; factor closer to 1 will indicate that the relevant passage is more likely to belong to the standard. Calculations of the correlation coefficient for the standard and Passage1 give $R_{e\text{-}y1}$=0.962716, and the correlation coefficient for the standard and Passage 2 - $R_{e\text{-}y2}$=0.909958. Similarly, the values of relative frequencies in Standard 2 and Passages 1, 2 in Fig. 3 differ significantly, so it is likely that the author of Standard 2 [34] is not the author of Passages 1 and 2.

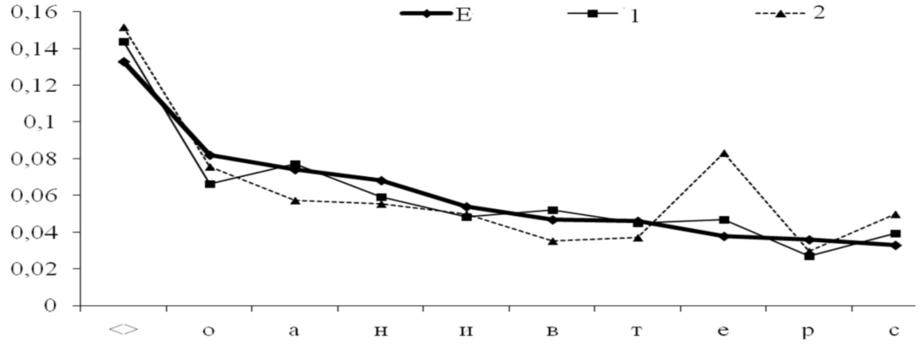


Figure2: The relative frequencies of occurrence of the ten most frequent symbols in Standard 1 and the studied Excerpts 1, 2, including omission

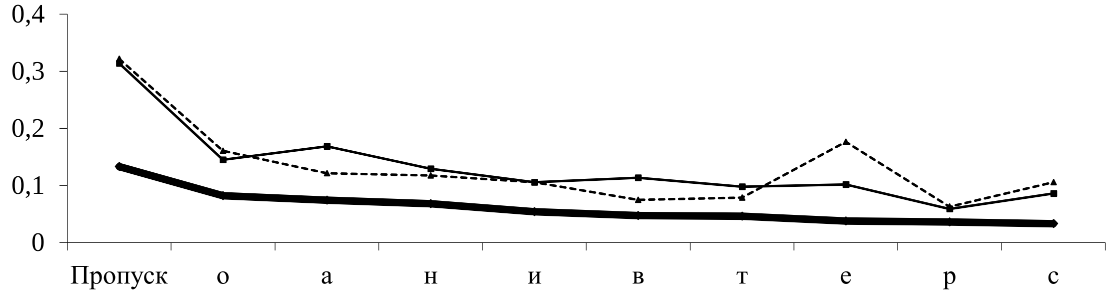


Figure3: The relative frequencies of occurrence of the ten most frequent symbols in Standard 2 and the studied Passages 1 and 2, including omission

The obtained values of the coefficients, as well as the analysis of the graphical results allow to state that the probability of belonging of Section 1 [35] to Standard 1 [36] is higher than for Section 2 [34].To achieve the research goal a system with the ability to select the language / languages of the analyzed content have been developed and implemented on the *Victana* web-resource [63]. For high-quality and effective analysis of content in determining the degree of authorship of a particular person, we propose to analyze the reference text and the study in several stages:

- Linguometric analysis of the coefficients of diversity of the author's speech (Fig. 4, Alg. 1);
- Stylometric analysis (Fig. 5);
- Analysis of stable phrases (Fig. 6);
- Linguistic and statistical analysis through N-grams (Fig. 7).

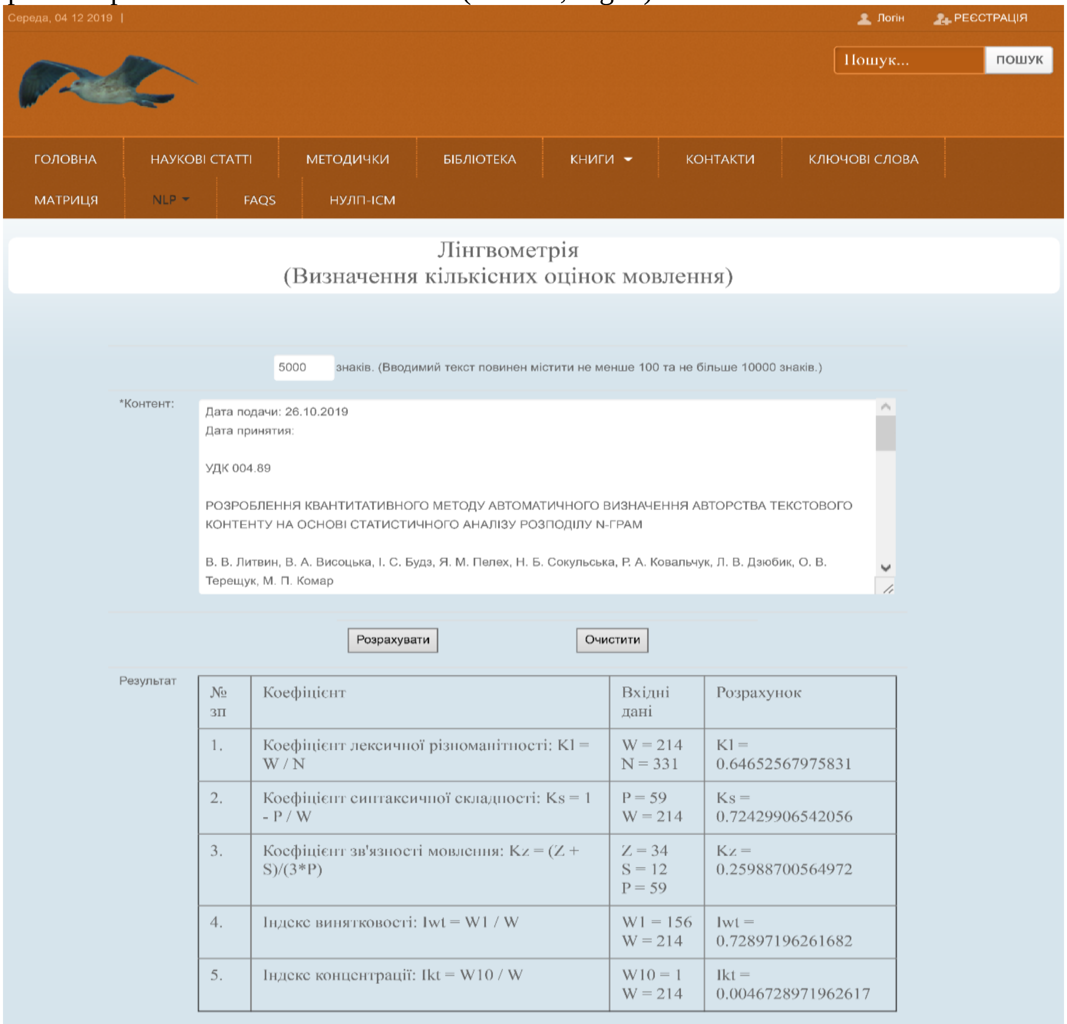
The Web-resource for linguometric analysis has the following fields (Fig. 4):

- Content -is a field where the researched text is copied from the buffer;
- Signs (the entered text must contain at least 100 and at most 10000 characters) is the maximum size of the content is a set;
- Calculation is meaning its start;
- clearance is clear the entered data.

<u>Algorithm 1. Linguometric analysis of the text to determine authorship.</u>

Step. 1. Check the length of the text - the excess is cut off.
Step. 2. Determine the number of sentences.
Step. 3. Purify the studied text (numbers, special symbols).
Step. 4. Determine the total number of words in the text N.
Step. 5. Determine the number of words W.
Step. 6. Determine the number of prepositions Z.

Step. 7. Determine the number of connectors S.
Step. 8. Calculate the coefficients of author speech.
Step. 9. Output the results to the end user (Table 2, Fig. 1).



Лінгвометрія
(Визначення кількісних оцінок мовлення)

5000 знаків. (Вводимий текст повинен містити не менше 100 та не більше 10000 знаків.)

*Контент:

Дата подачи: 26.10.2019
Дата принятия:

УДК 004.89

РОЗРОБЛЕННЯ КВАНТИТАТИВНОГО МЕТОДУ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТОВОГО КОНТЕНТУ НА ОСНОВІ СТАТИСТИЧНОГО АНАЛІЗУ РОЗПОДІЛУ N-ГРАМ

В. В. Литвин, В. А. Висоцька, І. С. Будз, Я. М. Пелех, Н. Б. Сокульська, Р. А. Ковальчук, Л. В. Дзюбик, О. В. Терещук, М. П. Комар

Результат

№ зп	Коефіцієнт	Вхідні дані	Розрахунок

1.	Коефіцієнт лексичної різноманітності: Kl = W / N	W = 214
		N = 331
			Kl = 0.64652567975831

2.	Коефіцієнт синтаксичної складності: Ks = 1 - P / W	P = 59
		W = 214
			Ks = 0.72429906542056

3.	Коефіцієнт зв'язності мовлення: Kz = (Z + S)/(3*P)	Z = 34
		S = 12
		P = 59
			Kz = 0.25988700564972

4.	Індекс винятковості: Iwt = W1 / W	W1 = 156
		W = 214
			Iwt = 0.72897196261682

5.	Індекс концентрації: Ikt = W10 / W	W10 = 1
		W = 214
			Ikt = 0.0046728971962617

Figure4: The example of linguistic analysis application result

Table 2
Example of author speech coefficients calculations

Coefficient	Incoming data	Calculation
Lexical diversity: $K_l=W/N$	W=184; N=295	K_l=0.6237
Speech connectivity: $K_z=(Z+S)/(3*P)$	Z=20; S=28; P=18	K_z=0.8889
Syntactic complexity: $K_s=1–P/W$	P=18; W=184	K_s=0.9022
Concentration index: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0.0109
Exclusivity index: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0.7663

The Web-resource for stylistic analysis has the following fields (Fig. 5):

- Select Passage 1 (2, 3) is open access to excerpts. Access to the next passage only after activating access to the previous one. Access is opened sequentially from a smaller number to a larger one.
- Reference text is the field where the Reference text is copied from the buffer.
- The text you enter must be at least 100 characters long. (Now 0) is after starting the calculation, the actual number of characters of each passage will be calculated and displayed separately.
- Passage 1 (2, 3) is the field where the corresponding excerpt text is copied from the buffer.
- Calculate is start the calculation.
- Clear is clear the entered data.

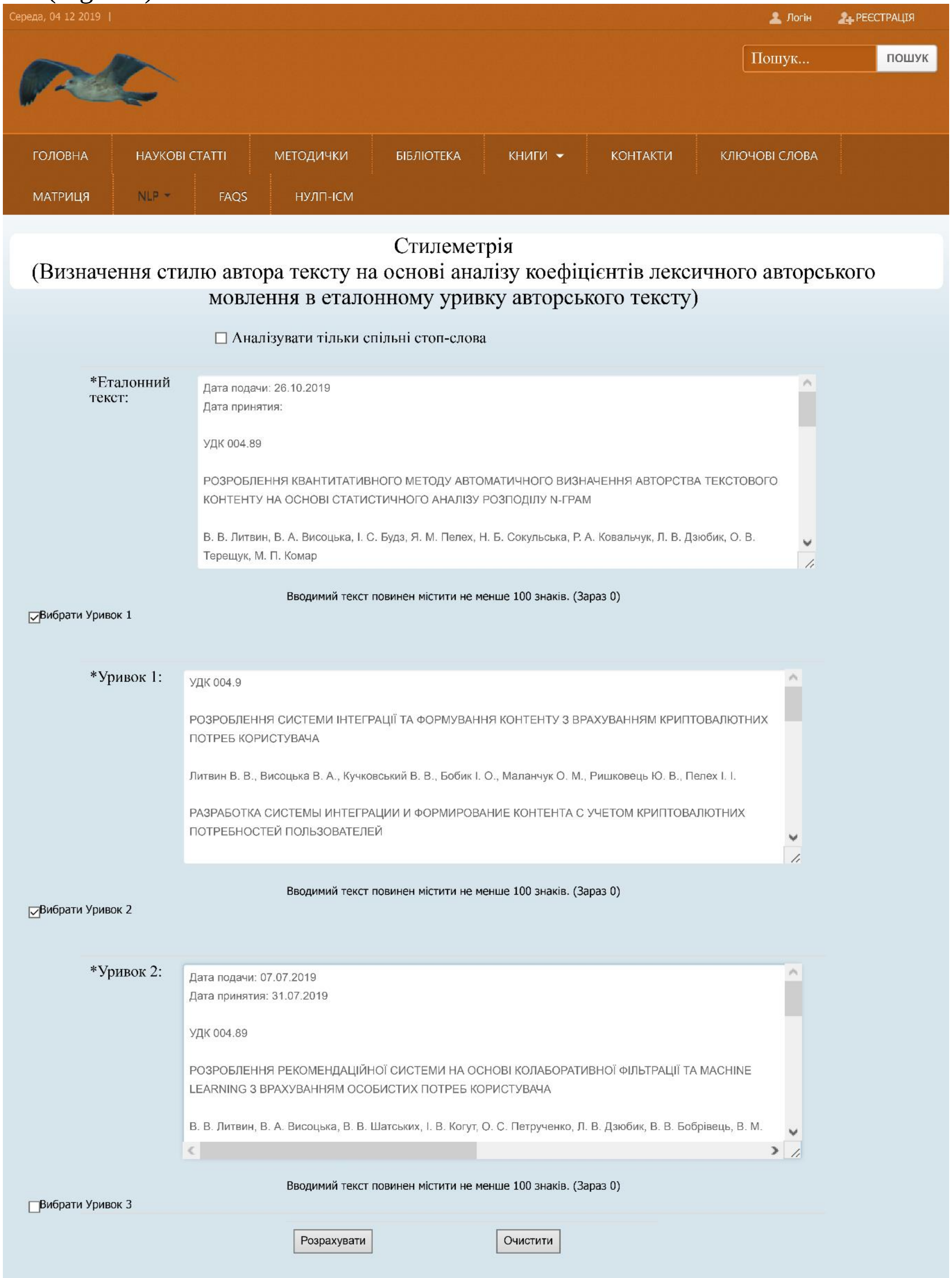
Algorithm 2. Stylometric analysis of the text to determine authorship.

Step. 1. Check the lengths of standard text and selected passages and reduce the length of the reference text to the minimum of the checked.

Step. 2. Clean the reference text from special characters, etc.

Step. 3. Determine of the words number in the text of the standard.

Step. 4. Determine the number of stop words (prepositions + conjunctions + particles) in the text of the standard (Fig. 5-6).



Стилеметрія

(Визначення стилю автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту)

□ Аналізувати тільки спільні стоп-слова

*Еталонний текст:

Дата подачи: 26.10.2019
Дата принятия:

УДК 004.89

РОЗРОБЛЕННЯ КВАНТИТАТИВНОГО МЕТОДУ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТОВОГО КОНТЕНТУ НА ОСНОВІ СТАТИСТИЧНОГО АНАЛІЗУ РОЗПОДІЛУ N-ГРАМ

В. В. Литвин, В. А. Висоцька, І. С. Будз, Я. М. Пелех, Н. Б. Сокульська, Р. А. Ковальчук, Л. В. Дзюбик, О. В. Терещук, М. П. Комар

Вводимий текст повинен містити не менше 100 знаків. (Зараз 0)

☑ Вибрати Уривок 1

*Уривок 1:

УДК 004.9

РОЗРОБЛЕННЯ СИСТЕМИ ІНТЕГРАЦІЇ ТА ФОРМУВАННЯ КОНТЕНТУ З ВРАХУВАННЯМ КРИПТОВАЛЮТНИХ ПОТРЕБ КОРИСТУВАЧА

Литвин В. В., Висоцька В. А., Кучковський В. В., Бобик І. О., Маланчук О. М., Ришковець Ю. В., Пелех І. І.

РАЗРАБОТКА СИСТЕМЫ ИНТЕГРАЦИИ И ФОРМИРОВАНИЕ КОНТЕНТА С УЧЕТОМ КРИПТОВАЛЮТНИХ ПОТРЕБНОСТЕЙ ПОЛЬЗОВАТЕЛЕЙ

Вводимий текст повинен містити не менше 100 знаків. (Зараз 0)

☑ Вибрати Уривок 2

*Уривок 2:

Дата подачи: 07.07.2019
Дата принятия: 31.07.2019

УДК 004.89

РОЗРОБЛЕННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ НА ОСНОВІ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ ТА MACHINE LEARNING З ВРАХУВАННЯМ ОСОБИСТИХ ПОТРЕБ КОРИСТУВАЧА

В. В. Литвин, В. А. Висоцька, В. В. Шатських, І. В. Когут, О. С. Петрученко, Л. В. Дзюбик, В. В. Бобрівець, В. М.

Вводимий текст повинен містити не менше 100 знаків. (Зараз 0)

□ Вибрати Уривок 3

Figure5: Example of data entry for stylometric analysis

Step. 5. The length of Passage 1 is not more than the minimum text.

Step. 6. Clear Passage 1 from special characters, etc.

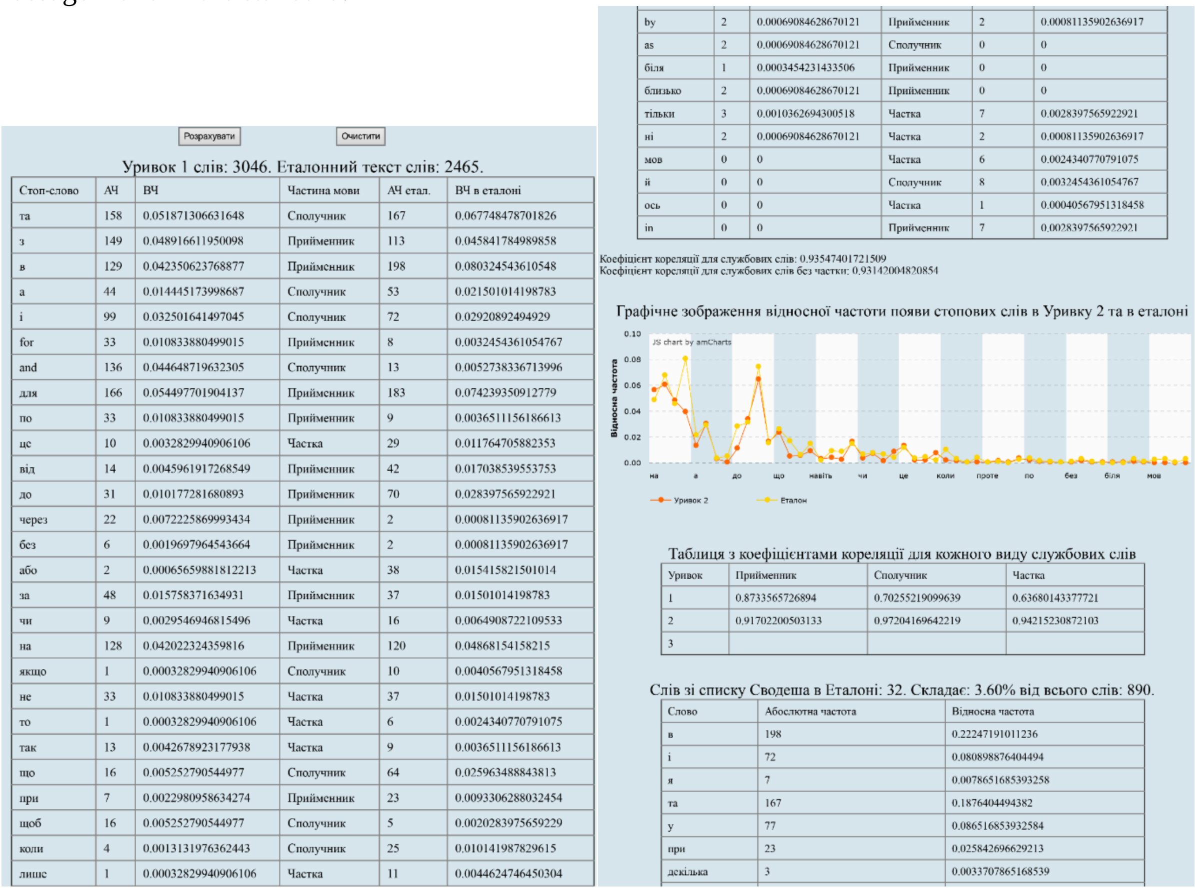
Step. 7. Determine the number of words W1 for Passage 1.

Step. 8. Determine the number of stop words (prepositions + conjunctions + particles) in the text.

Step. 9. Prepare individual arrays (excerpt and standard) to calculate the correlation coefficient (Fig. 6).

Step. 10. Call the function to calculate the correlation coefficient.

Step. 11. Form an array to form a graphical representation of the relative frequency of stop words in Passage 1 and in the standard.



Уривок 1 слів: 3046. Еталонний текст слів: 2465.

Стоп-слово	АЧ	ВЧ	Частина мови	АЧ етал.	ВЧ в еталоні
та	158	0.051871306631648	Сполучник	167	0.067748478701826
з	149	0.048916611950098	Прийменник	113	0.045841784989858
в	129	0.042350623768877	Прийменник	198	0.080324543610548
а	44	0.014445173998687	Сполучник	53	0.021501014198783
і	99	0.032501641497045	Сполучник	72	0.02920892494929
for	33	0.010833880499015	Прийменник	8	0.0032454361054767
and	136	0.044648719632305	Сполучник	13	0.0052738336713996
для	166	0.054497701904137	Прийменник	183	0.074239350912779
по	33	0.010833880499015	Прийменник	9	0.0036511156186613
це	10	0.0032829940906106	Частка	29	0.011764705882353
від	14	0.0045961917268549	Прийменник	42	0.017038539553753
до	31	0.010177281680893	Прийменник	70	0.028397565922921
через	22	0.0072225869993434	Прийменник	2	0.00081135902636917
без	6	0.0019697964543664	Прийменник	2	0.00081135902636917
або	2	0.00065659881812213	Частка	38	0.015415821501014
за	48	0.015758371634931	Прийменник	37	0.01501014198783
чи	9	0.0029546946815496	Частка	16	0.0064908722109533
на	128	0.042022324359816	Прийменник	120	0.04868154158215
якщо	1	0.00032829940906106	Сполучник	10	0.0040567951318458
не	33	0.010833880499015	Частка	37	0.01501014198783
то	1	0.00032829940906106	Частка	6	0.0024340770791075
так	13	0.0042678923177938	Частка	9	0.0036511156186613
що	16	0.005252790544977	Сполучник	64	0.025963488843813
при	7	0.0022980958634274	Прийменник	23	0.0093306288032454
щоб	16	0.005252790544977	Сполучник	5	0.0020283975659229
коли	4	0.0013131976362443	Сполучник	25	0.010141987829615
лише	1	0.00032829940906106	Частка	11	0.0044624746450304

by	2	0.00069084628670121	Прийменник	2	0.00081135902636917
as	2	0.00069084628670121	Сполучник	0	0
біля	1	0.0003454231433506	Прийменник	0	0
близько	2	0.00069084628670121	Прийменник	0	0
тільки	3	0.0010362694300518	Частка	7	0.0028397565922921
ні	2	0.00069084628670121	Частка	2	0.00081135902636917
мов	0	0	Частка	6	0.0024340770791075
й	0	0	Сполучник	8	0.0032454361054767
ось	0	0	Частка	1	0.00040567951318458
in	0	0	Прийменник	7	0.0028397565922921

Коефіцієнт кореляції для службових слів: 0.93547401721509
Коефіцієнт кореляції для службових слів без частки: 0.93142004820854

Графічне зображення відносної частоти появи стопових слів в Уривку 2 та в еталоні

Таблиця з коефіцієнтами кореляції для кожного виду службових слів

Уривок	Прийменник	Сполучник	Частка
1	0.8733565726894	0.70255219099639	0.63680143377721
2	0.91702200503133	0.97204169642219	0.94215230872103
3			

Слів зі списку Сводеша в Еталоні: 32. Складає: 3.60% від всього слів: 890.

Слово	Абослютна частота	Відносна частота
в	198	0.22247191011236
і	72	0.080898876404494
я	7	0.0078651685393258
та	167	0.1876404494382
у	77	0.086516853932584
при	23	0.025842696629213
декілька	3	0.0033707865168539

Figure6: Example of stylometric analysis application results

Step. 12. Call the function to calculate the relative frequency distribution graph (Fig. 6).

Step. 13. Call the function to calculate the correlation coefficient of Passage 2 (3) for each of the service words.

Step. 14. Form the words of the Swadesh list from the reference book, determine the number of words from the Swadesh list in the text of the Passage.

Step. 15. Form common for the Standard, Passages 1-3 and the Swadesh list.

Step. 16. The results of the study are displayed on the screen.

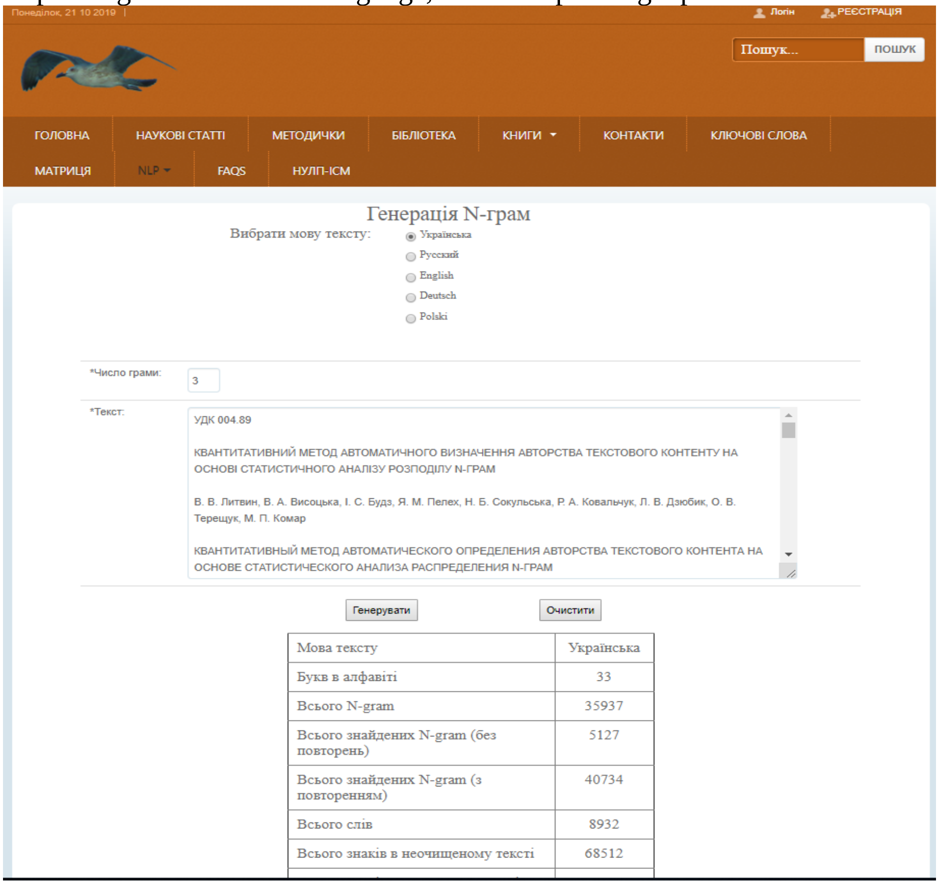
4. Experiment

When identifying the author of the text, it is assumed that the text reflects the individual style of author writing what distinguishes him from others. In order to compare texts with each other, it is necessary to compare the text with some numerical characteristics that would be close to the texts of the same author and would be significantly different for the works of different authors.

The Web-resource for the analysis of N-grams has the following fields (Fig. 7):

- Number of grams - the number of characters in grams. Default is 3 ones. Can be changed to 1, 2, 3, 4.
- Choice of the text language - the language of the text for analysis (research). The default one is "Ukrainian".
- Text - a field where the studied text is copied from the buffer.
- Restriction of text in characters.
- Generation - to start generating N-grams.
- Clearance - clear the entered data.

Algorithm 3. Linguistic and statistical analysis of N-grams of text is the following:
Step. 1. Purify the studied text (numbers, special symbols).
Step. 2. Calculate the number of words in the text.
Step. 3. All words of the text are translated in lower case.
Step. 4. Remove the spaces.
Step. 5. Depending on the selected language, the corresponding alphabet is substituted.



Генерація N-грам

Вибрати мову тексту: ● Українська ○ Русский ○ English ○ Deutsch ○ Polski

*Число грами: 3

*Текст:

УДК 004.89

КВАНТИТАТИВНИЙ МЕТОД АВТОМАТИЧНОГО ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТОВОГО КОНТЕНТУ НА ОСНОВІ СТАТИСТИЧНОГО АНАЛІЗУ РОЗПОДІЛУ N-ГРАМ

В. В. Литвин, В. А. Висоцька, І. С. Будз, Я. М. Пелех, Н. Б. Сокульська, Р. А. Ковальчук, Л. В. Дзюбик, О. В. Терещук, М. П. Комар

КВАНТИТАТИВНЫЙ МЕТОД АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВОГО КОНТЕНТА НА ОСНОВЕ СТАТИСТИЧЕСКОГО АНАЛИЗА РАСПРЕДЕЛЕНИЯ N-ГРАМ

Мова тексту	Українська
Букв в алфавіті	33
Всього N-gram	35937
Всього знайдених N-gram (без повторень)	5127
Всього знайдених N-gram (з повторенням)	40734
Всього слів	8932
Всього знаків в неочищеному тексті	68512

Figure7: Example of N-gram text analysis application

Step. 6. Depending on the set number of grams the corresponding function which calculates all possible variants of grams and saves in an array is started.

Step. 7. The function of counting the number of occurrences of words is started.

Here we calculate the relative frequency of occurrence and store it in the array: the ordinal number of the gram, the gram itself, the number of occurrences of this gram, the relative occurrence frequency of this gram.

Step. 8. The following function forms the array received in the previous function for export to the CSV file. This file is stored on the server. It can be downloaded to the computer of the user (researcher) via the link, which will be accessible after the formation of the form with the results of the study.

Step. 9. The results of the study are displayed on the screen (only those grams that are found in the text).

Step. 10. Access the export file.

Step. 11. The generalized results are deduced:

- only N-grams with repetitions were found
- only N-grams were found without repetitions
- total N-grams
- number of characters in the text that are completely cleared
- number of characters in the text with spaces
- number of words in the text
- size of the alphabet.

Three publications of scientific and technical orientation on the basis of linguistic and statistical analysis of 3-grams have been compared. Articles 1 and 2 have been written by one team, Article 3 has been written by another author (Table 3). The language of the text is Ukrainian (letters in the alphabet - 33, all possible N-grams – 35937)

Table 3
Values of parameters for the analyzed Articles 1–3

Parameters	Article 1	Article 2	Article 3
Total characters in plain text	29967	32570	37062
Total characters in the raw text	39792	39663	47084
Total words	5475	5358	6060
Total N-grams found (with repetition)	29494	29862	36383
Total N-grams found (no iterations)	4354	4377	3890
Total N-gram	35937	35937	35937

When comparing articles only those 3-grams that are found in the text at the same time in three articles at least once have been taken into account. Therefore, for this particular example, all 3-grams are 2147. That is, for Article 1 78.4814% 3-grams have been analyzed, for Article 2 - 72.6332% and for Article 3 - 84.1271%. Accordingly, the difference in consumption of the relevant 3-grams between Articles 1 and 2 is R_{12}=56.5254 %, between Articles 2 and 3 - R_{23}=69.4271 %, between Articles 1 and 3 - R_{13}=62.9839 %. These indicators themselves show that the characteristics of Articles 1 and 2 are more similar (R_{23}>R_{12}on 12.9017 %, R_{23}> R_{13}on 6.4432 %, R_{13}> R_{12}on 6.4585 %, that is R_{23}>R_{13}>R_{12}) than the characteristics under Articles 1-3 and 2-3. The smaller the R_{ij}, the greater the degree that the articles are written by the same author. Then in the case of Articles 1 and 2 it is more likely to be written by one author / team than Articles 2-3 and Articles 1-3 respectively. But we will analyze the use of individual clusters of 3-grams in the relevant articles and compare the results.

Fig. 8 presents the results of the analysis of use in Articles 1-3 of 3-grams, starting with the letter *a* (appearance in Articles 1-3 in the range of 6.1125-6.7087%). Most often the curve lines for Articles 1-2 (4.2322%) and Articles 1-3 (4.197%) coincide or approach each other (average discrepancy is 0.02713% and 0.0269%, respectively). But not always there is a coincidence with Article 2-3 (4.6322%) and there are significant differences (the average difference is 0.02969%). If you analyze only such 3-grams it turns out that all three articles are written more likely by one author. This is due to the fact that this letter is one of the most commonly used for the formation of Ukrainian words.

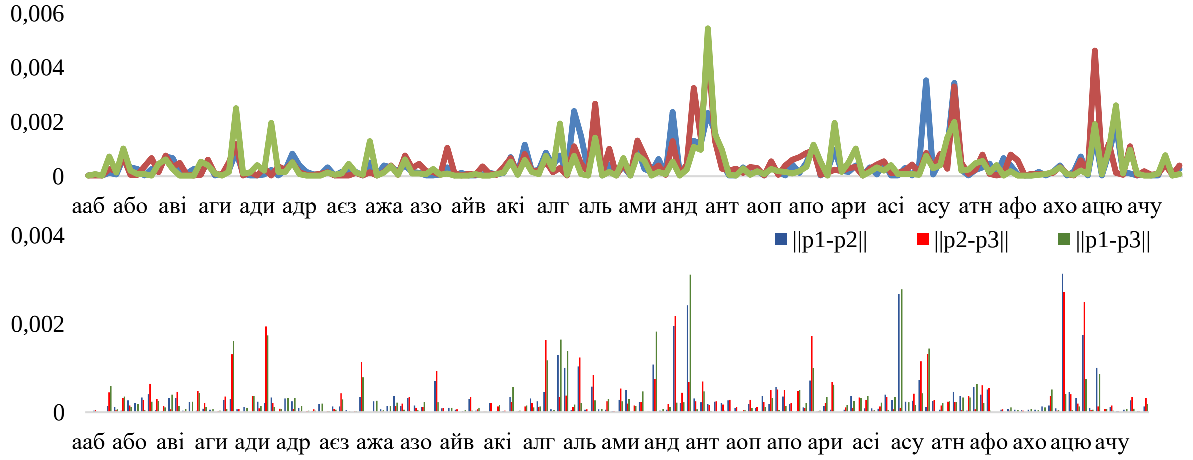


Figure8: The use of 3-grams, starting with the letter *a* (Article 1 – blue, Article 2 – red, Article 3 – green)

Fig. 9 presents the analysis results of use in Articles 1-3 of 3-grams, starting with the letter *б* (letter *b* in English) (appearance in Articles 1-3 in the range of 0.48884-0.77738%). Most often the curve lines for Articles 1-2 (0.594%) as opposed to Articles 1-3 (0.7072%) and Articles 2-3 (1.1208%) coincide or approach. But the trajectory of the curve of Article 1 and Article 3 often coincides (most likely articles are written by one author, the average discrepancy is 0.01809%, while for Articles 1-2 - 0.0261% and

Articles 2-3 - 0.02866%. If analyze only such 3-grams (which are less common), it turns out that all Articles 1-2 are written more likely by one author, and Article 3 - by another one. This is due to the fact that this letter would be rare in the formation of Ukrainian words. And some authors use such words more often because of habit and / or because of the subject matter of their publications (this requires further research).

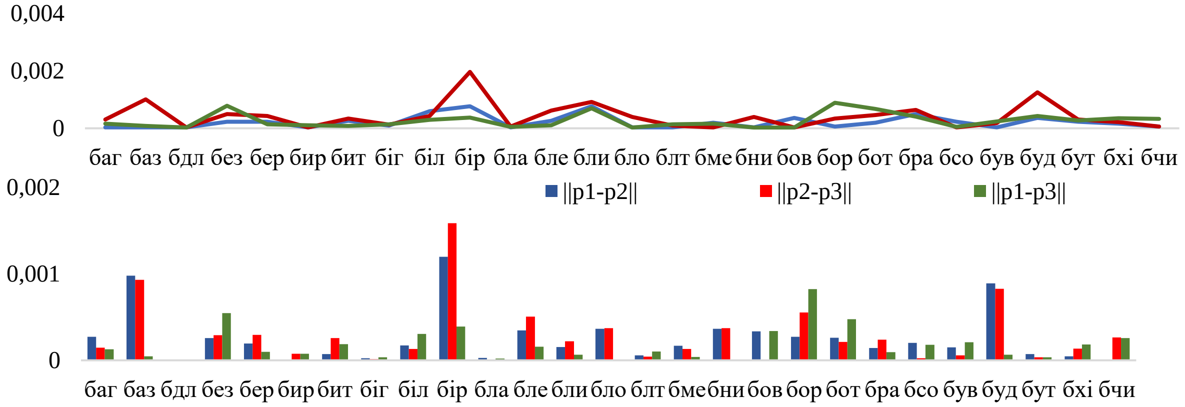


Figure9: The use of 3 grams, starting with the letter б (Article 1 – blue, Article 2 – red, Article 3 – green)

According to Table 4 and Fig. 10-12, a part of the letters in the Ukrainian language are most often used, others - much less often. For the most frequently used letters, the frequency of occurrence of 3-grams with such initial letters will have almost the same distribution (top values in the graph of Fig. 12), and not for other letters.

Table 4

Distribution of frequencies of 1-gram in Articles 1–3

1 gram	N1	N2	N3	P1	P2	P3
о	2824	0.094240	2472	0.075898	3870	0.103601
н	2471	0.082460	2370	0.072766	2888	0.077312
а	2255	0.075252	2698	0.082837	2491	0.066685
т	2102	0.070146	1956	0.060055	2141	0.057315
і	1789	0.059701	1967	0.060393	2250	0.060233
и	1732	0.057799	1852	0.056862	2036	0.054504
в	1654	0.055196	1590	0.048818	1915	0.051265
с	1549	0.051692	1327	0.040743	1384	0.037050
е	1404	0.046853	1453	0.044612	2090	0.055950
р	1335	0.044550	1722	0.052871	1893	0.050676
к	1279	0.042682	1110	0.034080	1453	0.038897
л	1116	0.037242	927	0.028462	906	0.024254
у	987	0.032937	960	0.029475	1195	0.031990
д	859	0.028666	939	0.028830	1319	0.035310
м	808	0.026964	976	0.029966	1399	0.037451
п	647	0.021591	825	0.025330	1138	0.030464
я	647	0.021591	681	0.020909	864	0.023129
з	623	0.020790	644	0.019773	946	0.025325
ь	498	0.016619	418	0.012834	613	0.016410
ч	459	0.015317	289	0.008873	574	0.015366
г	408	0.013615	373	0.011452	651	0.017427
х	355	0.011847	384	0.011790	482	0.012903
б	284	0.009477	569	0.017470	428	0.011458
ж	246	0.008209	210	0.006448	176	0.004712
й	239	0.007976	260	0.007983	265	0.007094
ц	224	0.007475	334	0.010255	299	0.008004
є	188	0.006274	165	0.005066	347	0.009289
ф	179	0.005973	209	0.006417	137	0.003668

ї	174	0.005807	217	0.006663	270	0.007228
ю	156	0.005206	277	0.008505	289	0.007737
ш	117	0.003904	169	0.005189	281	0.007522
щ	95	0.003170	52	0.001597	128	0.003427

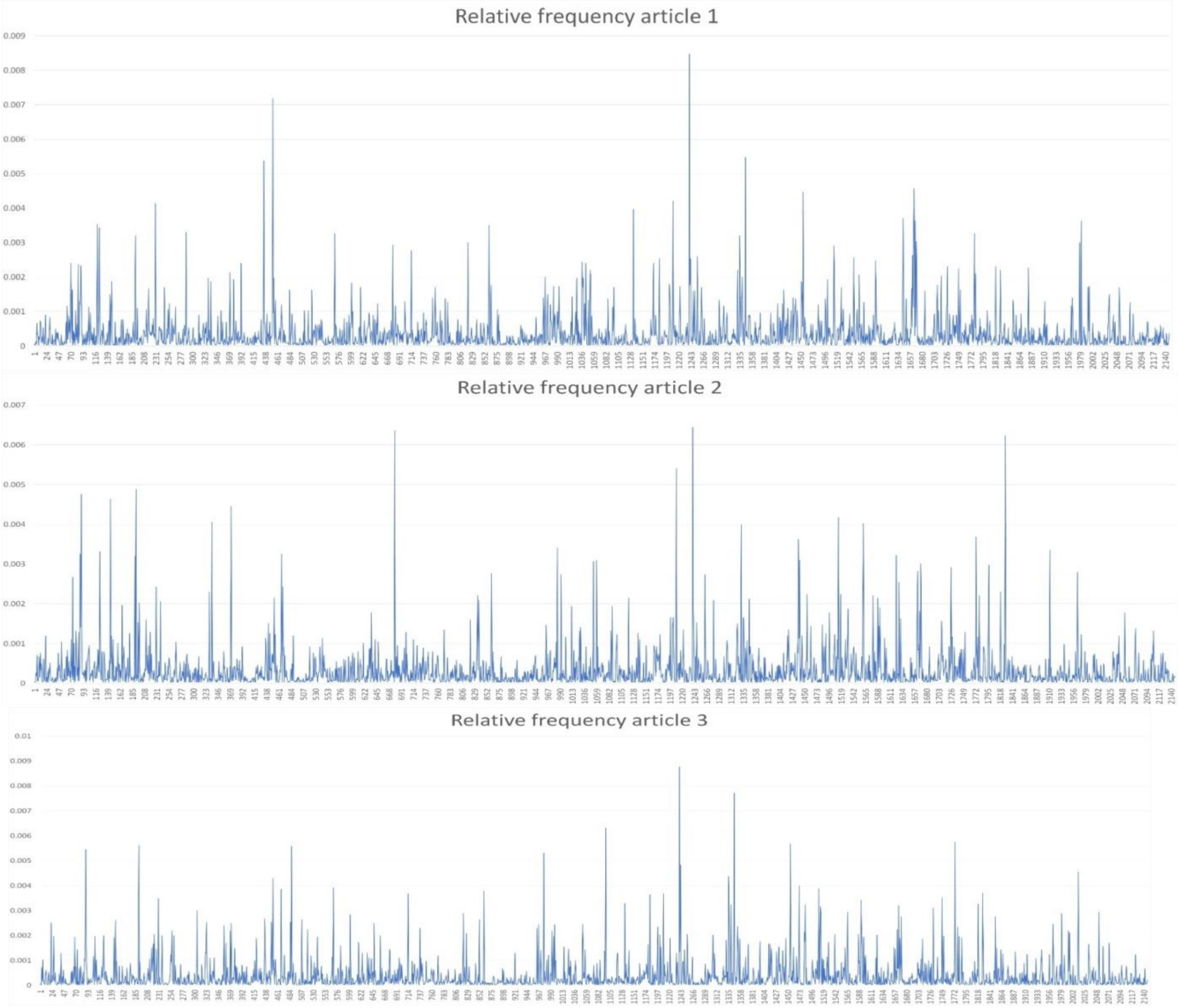


Figure10: Relative frequency for Article 1-3

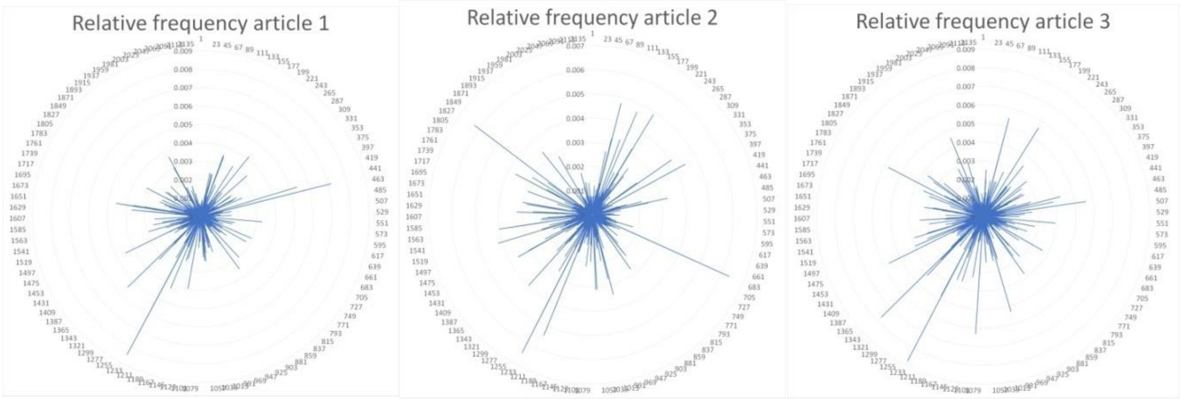


Figure 1: Relative frequency for Article 1-3

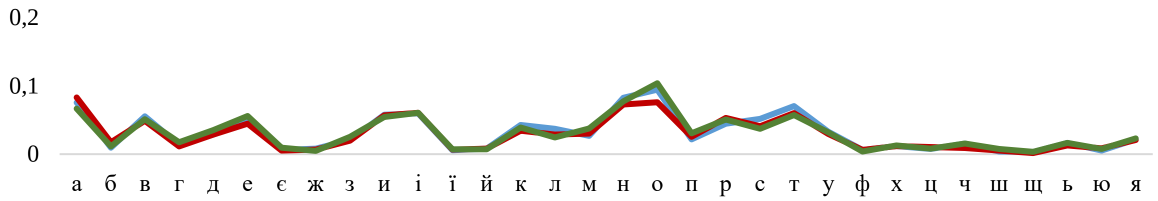


Figure12: Frequency distribution of 1-gram in Articles 1-3 (Article 1 – blue, Article 2 – red, Article 3 – green)

Therefore, it is advisable to study only the trigrams for the initial letters, which are less common in the texts of a particular language to determine the degree of belonging of the text to the author (for example, Fig. 12). Thus, for 3-grams of the letter are (the appearance in Articles 1-3 in the range of 0.2517-0.707%) most often the lines of curves for Articles 1-2 (0.2508%) in contrast to Articles 1-3 (0.6077 %) and Articles 2-3 (0.5443%) that coincide or approach each other. But the trajectory of the curve of Article 1 and Article 2 often coincides (most likely articles written by one author - the average discrepancy is 0.0114%, while for Articles 2-3 - 0.02478% and Articles 1-3 - 0.02762% this value is higher twice as much).

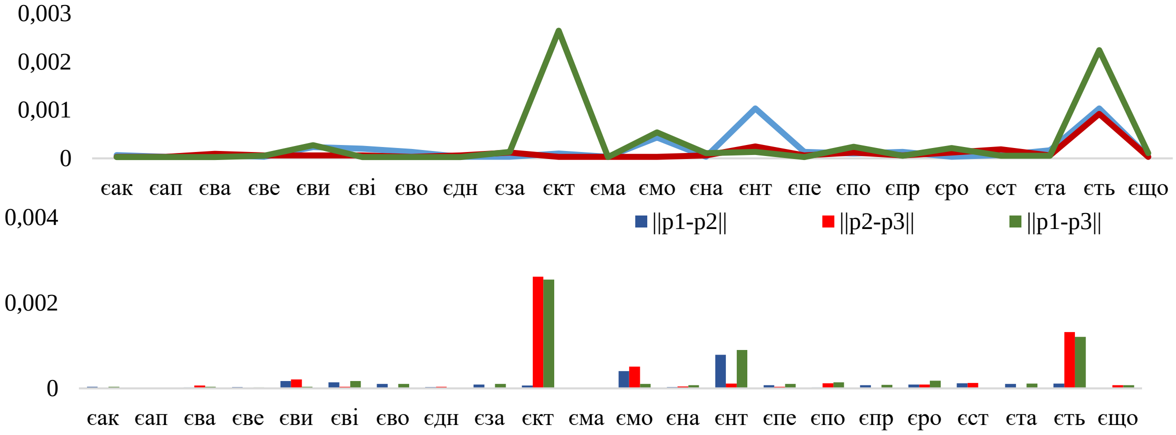


Figure12: The use of 3-grams, starting with the letter є (Article 1 – blue, Article 2 – red, Article 3 – green)

Table 4 shows frequencies of letters appearance in the standard and the studied passages. Fig. 14 shows histograms of the relative frequency of n-grams in 1-3 articles. Low frequency (noise) values are the most common and form the main volume of the data. We can ignore them (Fig. 15).

Table 5

Frequencies of letters appearance in the standard and the studied passages

Indexes	Article 1	Article 2	Article 3
Average	0.000366529	0.000339199	0.000392978
Standard error	1.28793E-05	1.24565E-05	1.53165E-05
Median	0.000167	0.000154	0.000162
Fashion	0.000033	0.000031	0.000027
Standard deviation	0.000596773	0.00057718	0.000709699
Sampling variance	3.56138E-07	3.33136E-07	5.03673E-07
Kurtosis	37.42530062	32.63050249	29.5089837
Asymmetry	4.881688545	4.62453506	4.54877741
Interval	0.008443	0.006417	0.008742
Minimum	0.000033	0.000031	0.000027
Maximum	0.008476	0.006448	0.008769
Sum	0.786938	0.72826	0.843723
Amount	2147	2147	2147
Reliability level (95.0%)	2.52573E-05	2.4428E-05	3.00366E-05

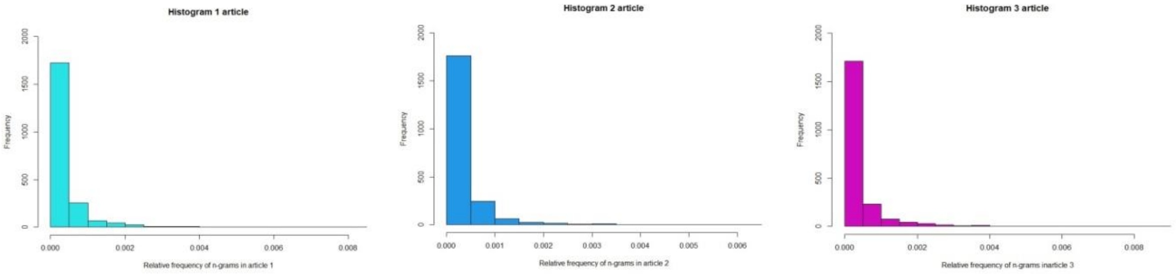


Figure14: Histogram of the relative frequency of N-grams in Articles 1-3

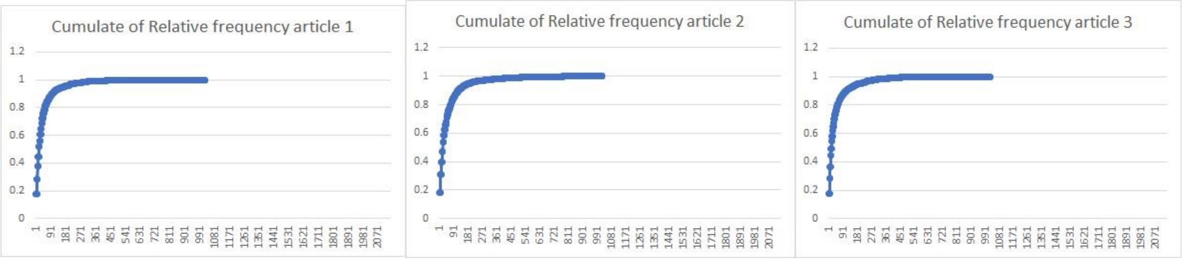


Figure15: Cumulates of common N-gram frequencies in Articles 1-3

All graphs of the distribution of the frequency of 3-grams in articles show a significantly noticeable gradation of 3-grams on underused (noise-like) and widely used peak values. This allows to see the specific examples of the three articles, the fact that to reduce the amount of information analyzed it is desirable to proceed to the analysis of the distribution function from a certain threshold value of frequency and at the same time cover the main information content is visible. To compare the distribution function in the context of the three studied articles, it is necessary to compare clearly expressed average values. After analyzing the most commonly used 3-grams, we conclude that they are caused by the stylistics or grammar of the Ukrainian language and are not relevant to determine the specific topic of articles. The most used 3-grams in Article:

- 1: *ння* [nnya] 0.008476, *енн* [enn] 0.007175, *ого* [oho] 0.005473.
- 2: *ння* [nnya] 0.006448, *ист* [yst] 0.006356, *ува* [uva] 0.006233.
- 3: *ння* [nnya] 0.008769, *ого* [oho] 0.007717, *мет* [met] 0.006314.

5. Results

In the algorithmic approach, the appearance of the trend is obtained due to various algorithms that practically implement smoothing procedures. 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 the following simple or ordinary moving average (Fig. 16), weighted moving average, exponential smoothing - median smoothing. In this part of the calculation work, the relative frequency of consumption of 3-grams in three texts has been smoothed by the method of moving average, exponential smoothing and median smoothing.

The moving average method is one of the oldest known methods of smoothing the time series. It is based on the transition from the initial values of the series to their average values in the time interval, the length of which is selected in advance. The selected time interval slides along the row. Moving averages can smooth out both random and periodic fluctuations, identify existing trends in the process and therefore serve as an important tool in filtering time series components. The moving average method estimates the average level over a period of time. The longer the time interval to which the average belongs, the smoother the level will be, but the less accurately the trend of the original time series will be described. In all figures, the gray graph is the graph of the initial Relative frequency, and the red graph is the graph of the smoothed Relative frequency data.

At small values of the size of the interval w, the efficiency in terms of smoothing effect is not very high, as can be seen in the following Figures 16-18 for Article 1 (smooth the data using the size of the smoothing interval w = 3, 5, 7, 9,11, 13, 15).

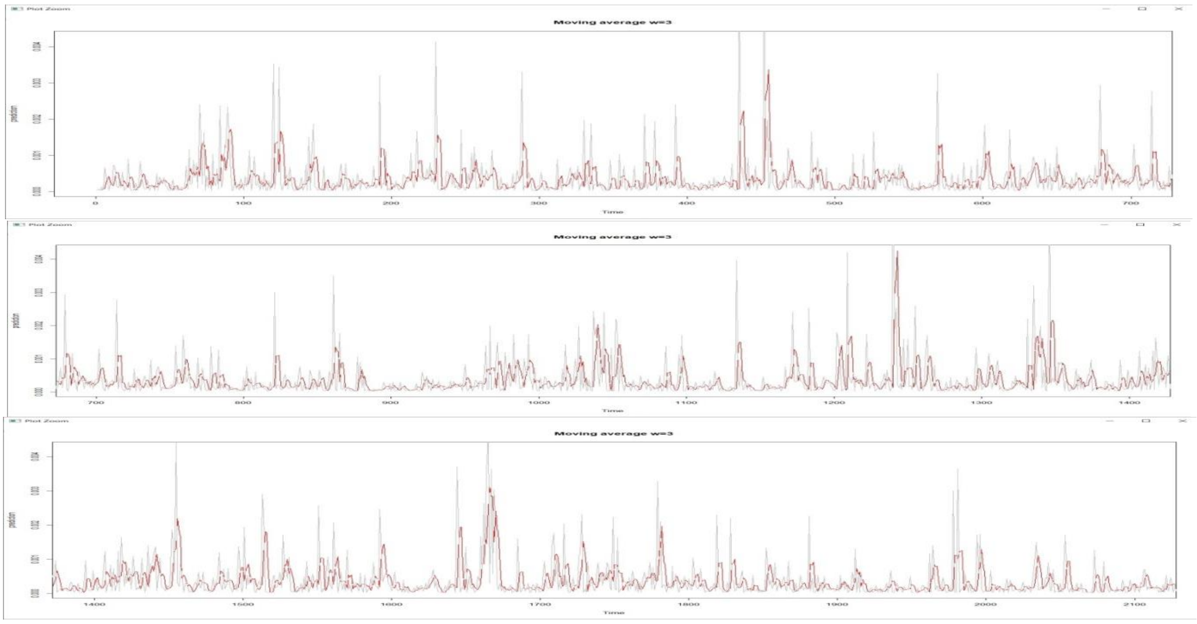


Figure 2: Moving Average of Article 1 for w=3 for the interval 0-700, 700-1400 and 1400-2100

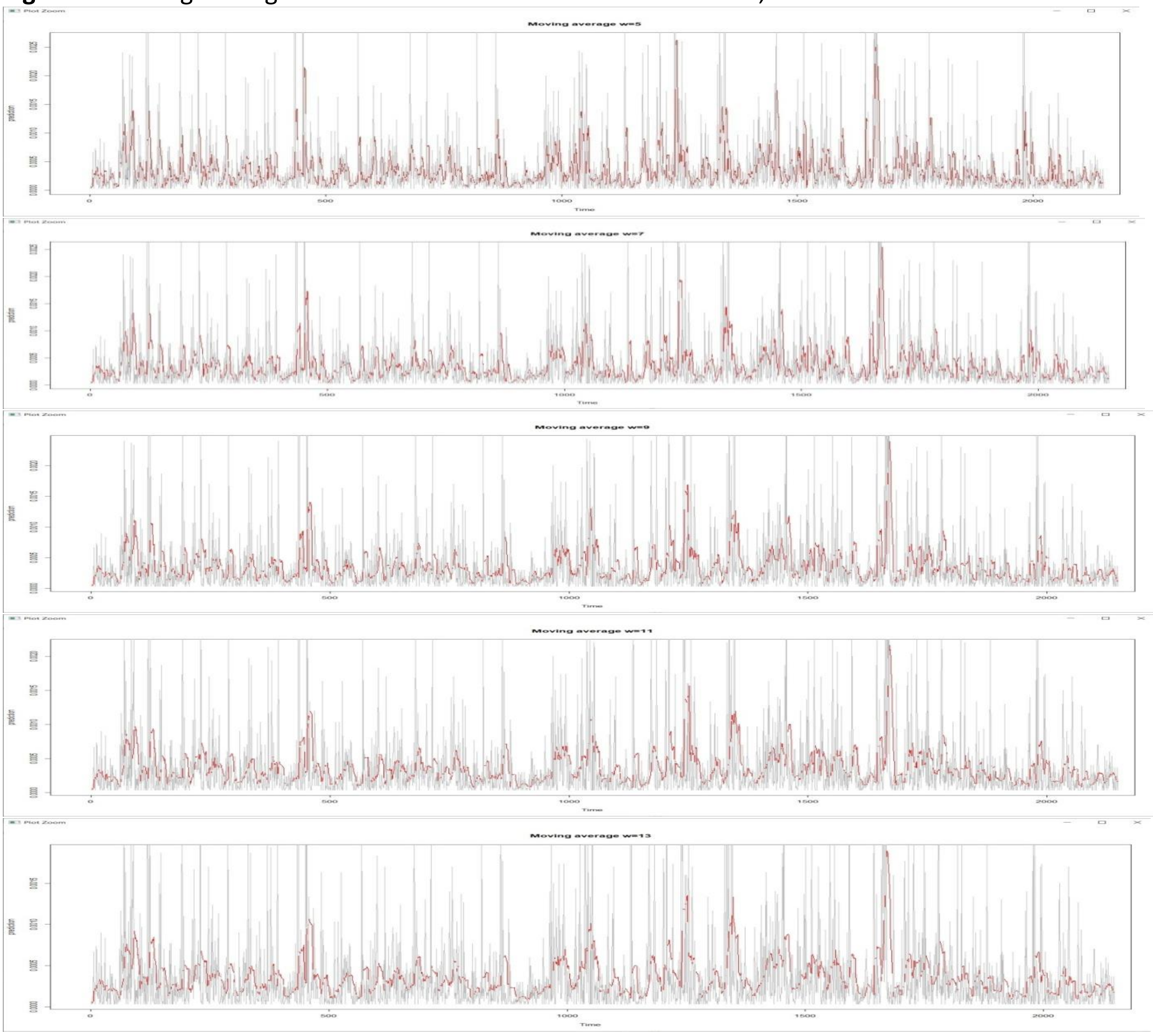


Figure 3: Moving Average of Article 1 for w = 5, 7, 9,11, 13

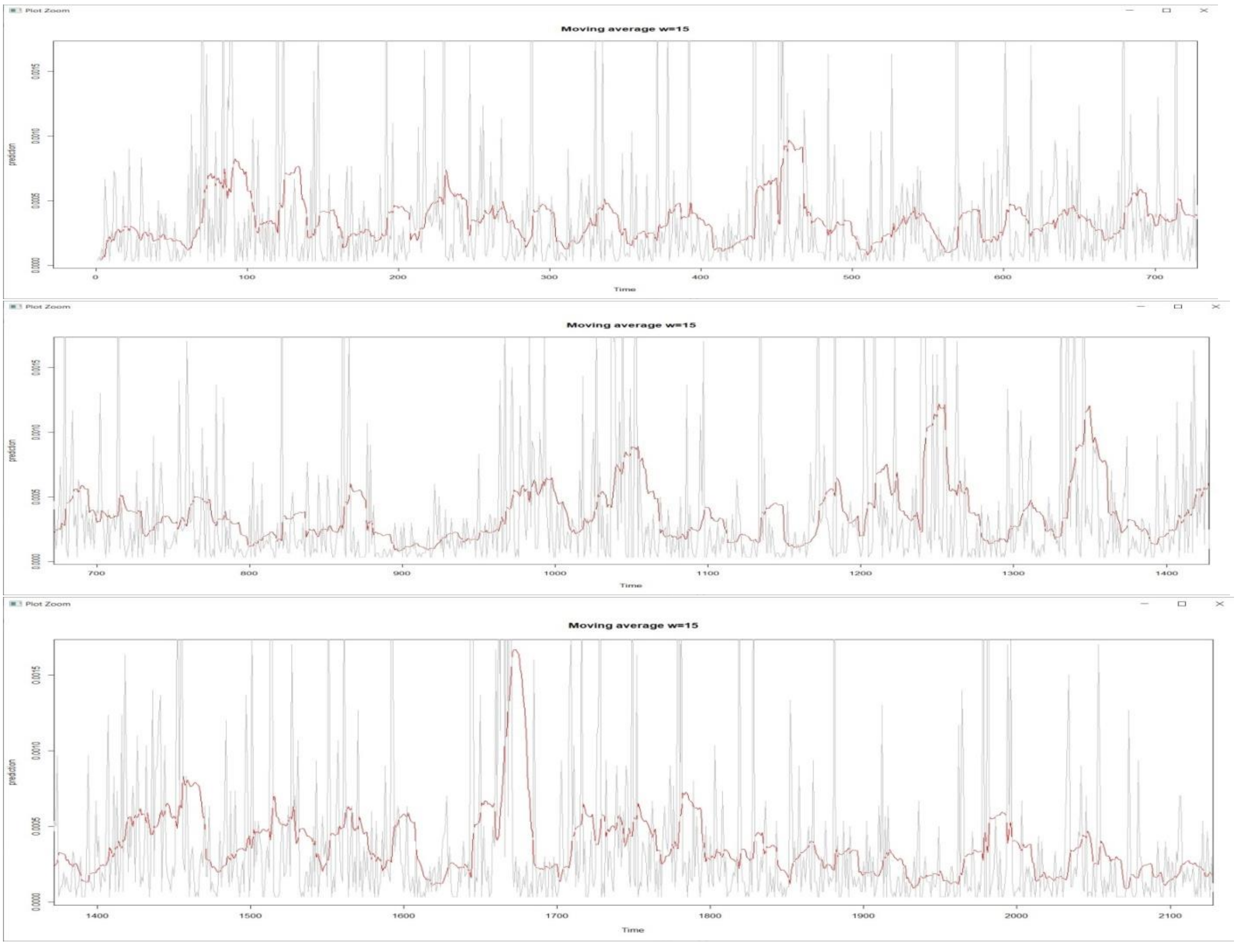


Figure 4: Moving Average of Article 1 for w=15 for the interval 0-700, 700-1400 and 1400-2100

It is needed to smooth the data using the size of the smoothing interval w = 3 (Fig. 19), then smooth the obtained data again but using the size of the smoothing interval w = 5. Then continue smoothing the obtained data with the smoothing interval w = 7 and so on to w = 15 (Fig. 20-21).

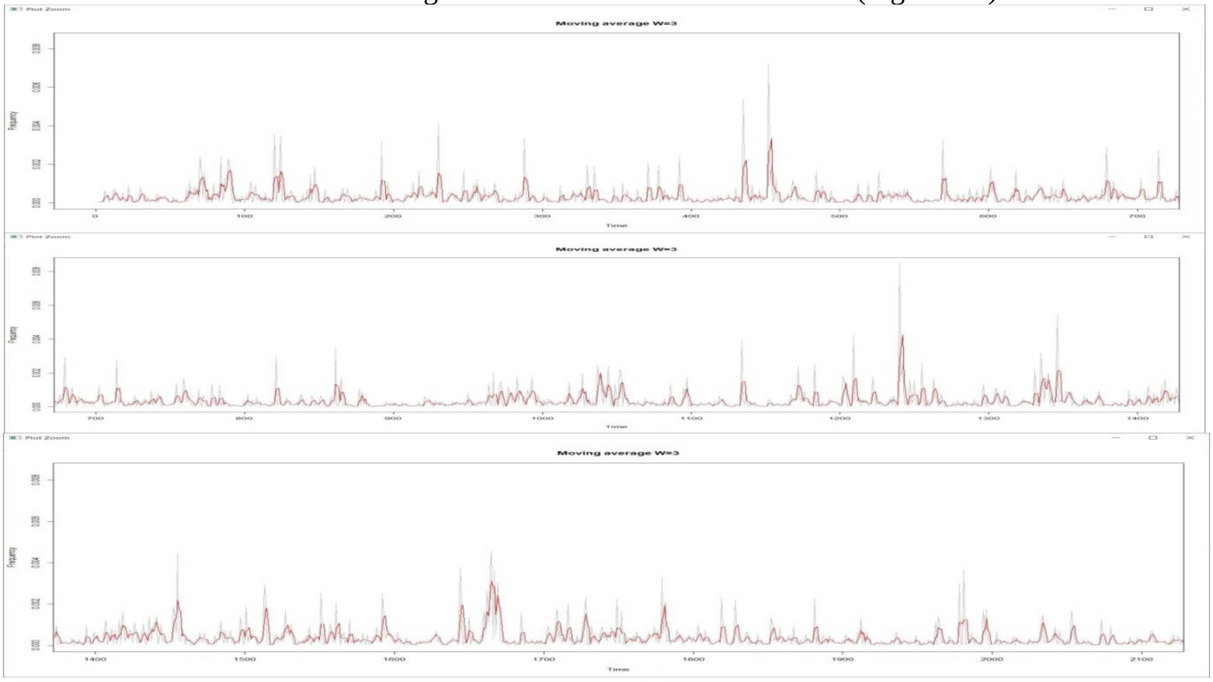


Figure 5: Moving Average of article 1 for w=3 for the interval 0-700, 700-1400 and 1400-2100

At small values of the size of the interval w, the efficiency in terms of the smoothing effect decreases, which can be seen in the following figures. Also, the smoothing method, using pre-smoothed rows, smoothest the data very effectively.

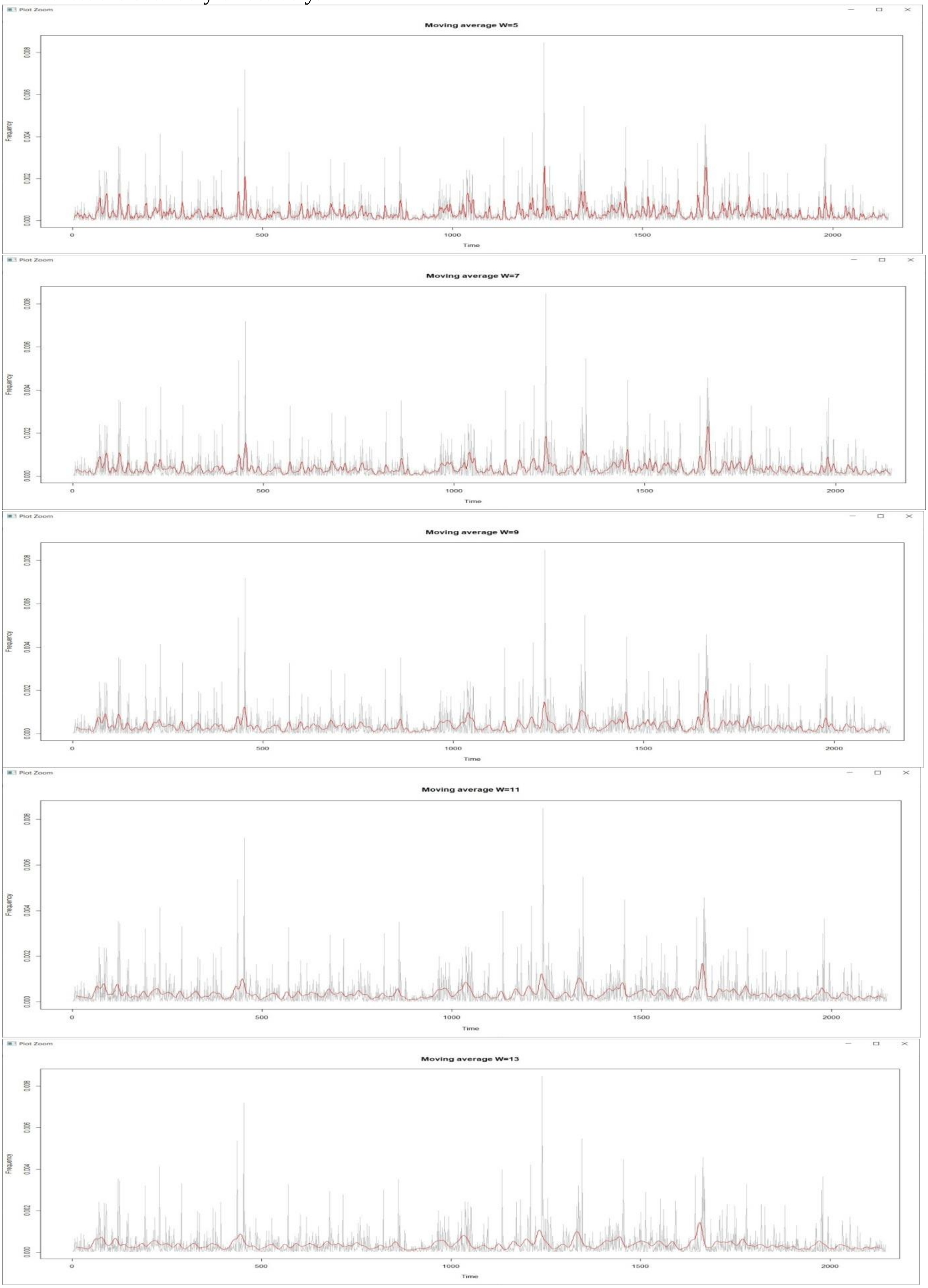


Figure 6: Moving Average of Article 1 for w = 5, 7, 9,11, 13

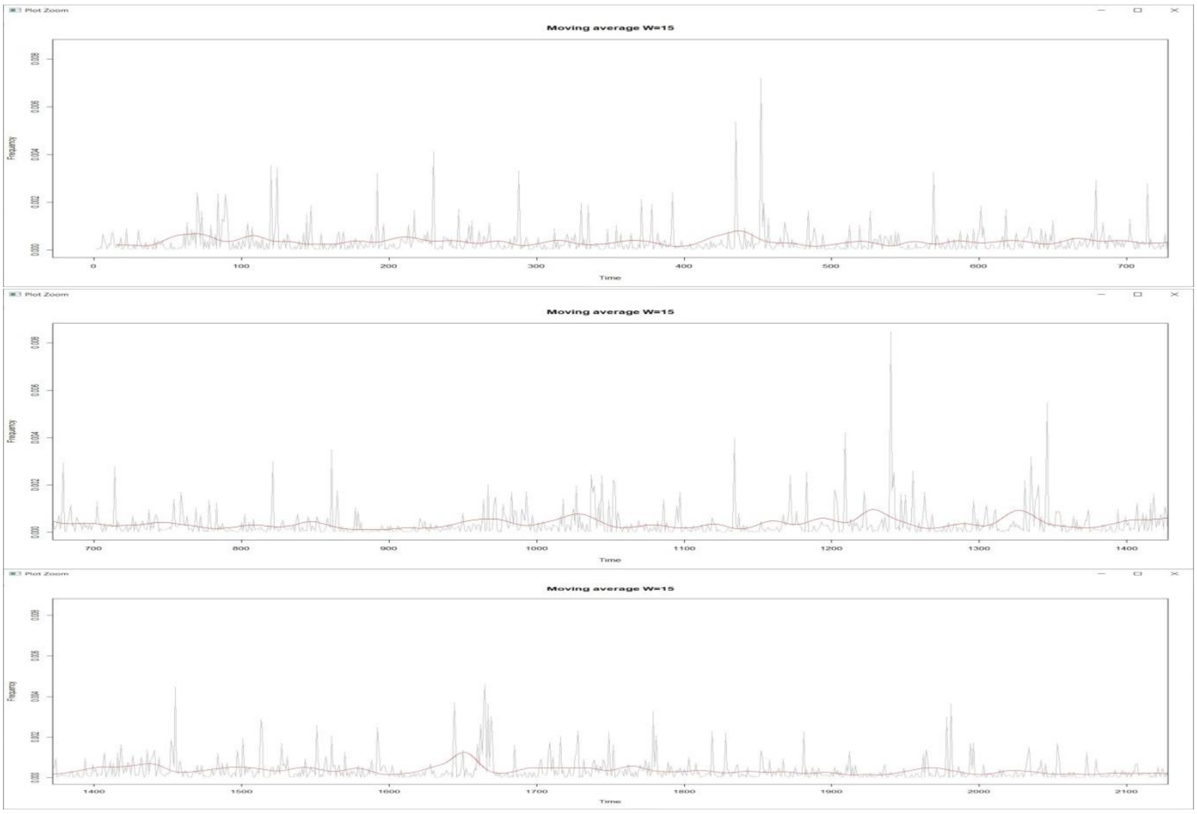


Figure 7: Moving Average of Article 1 for w=3 for the interval 0-700, 700-1400 and 1400-2100

We smooth the data using the dimensions of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 for Article 2, the moving average showed a trend in the interval better than for Article 1 (Fig. 22-24).

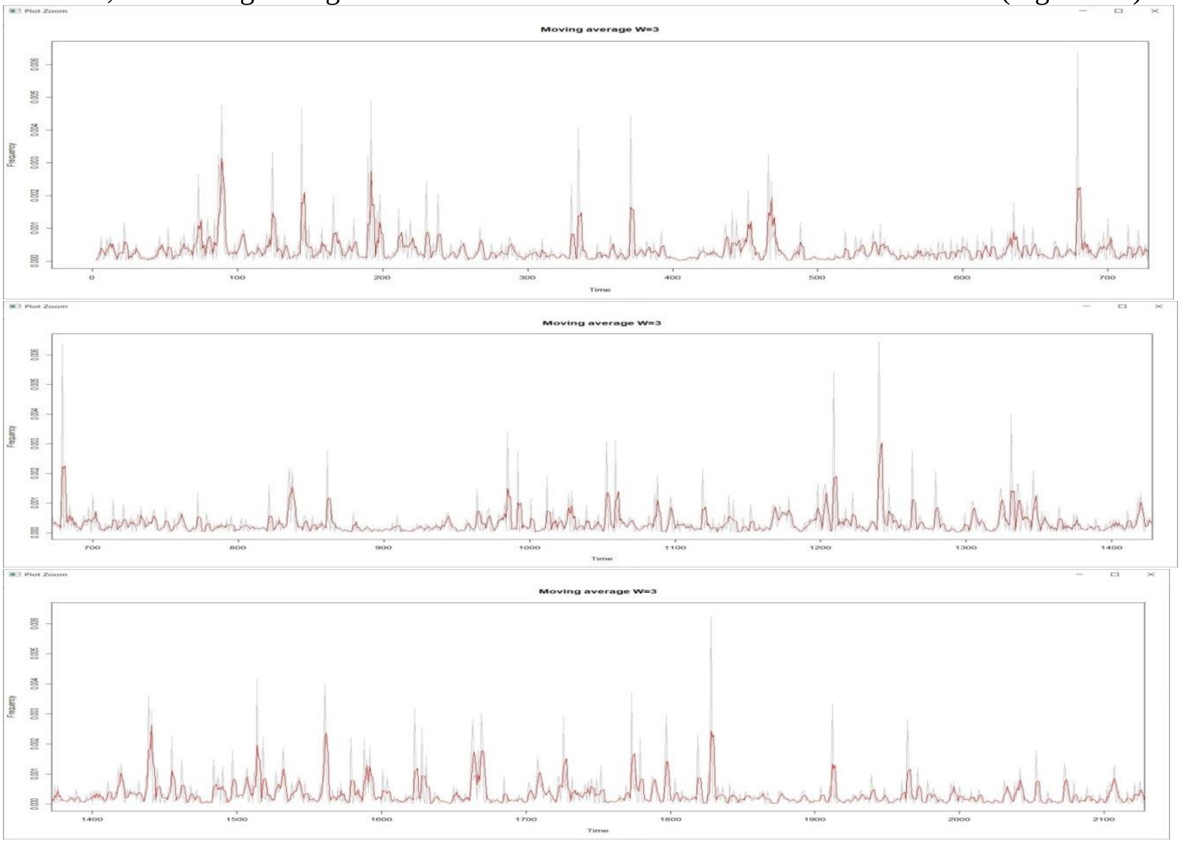


Figure 8: Moving Average of Article 2 for w=3 for the interval 0-700, 700-1400 and 1400-2100

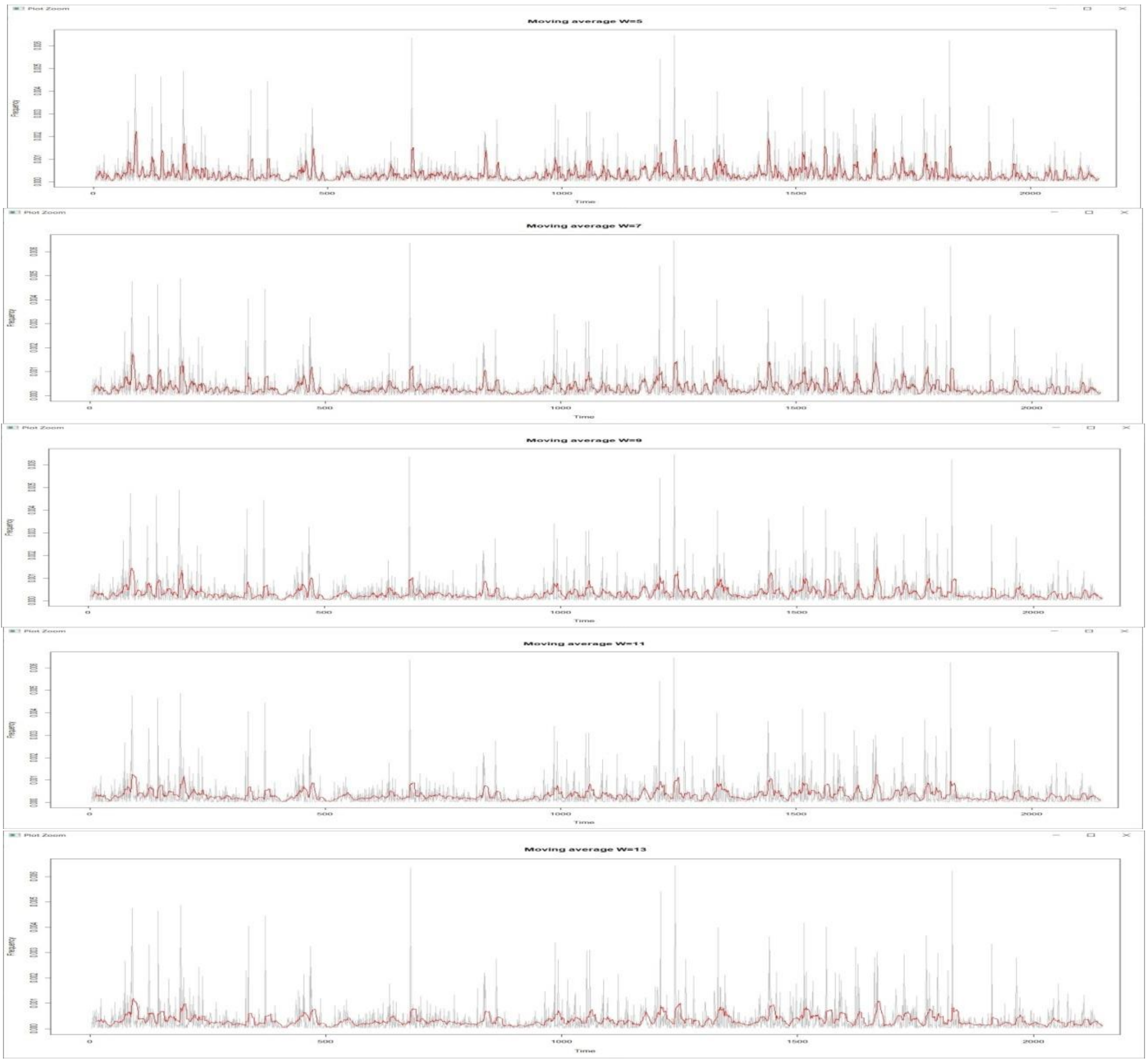


Figure 9: Moving Average of Article 2 for w = 5, 7, 9,11, 13

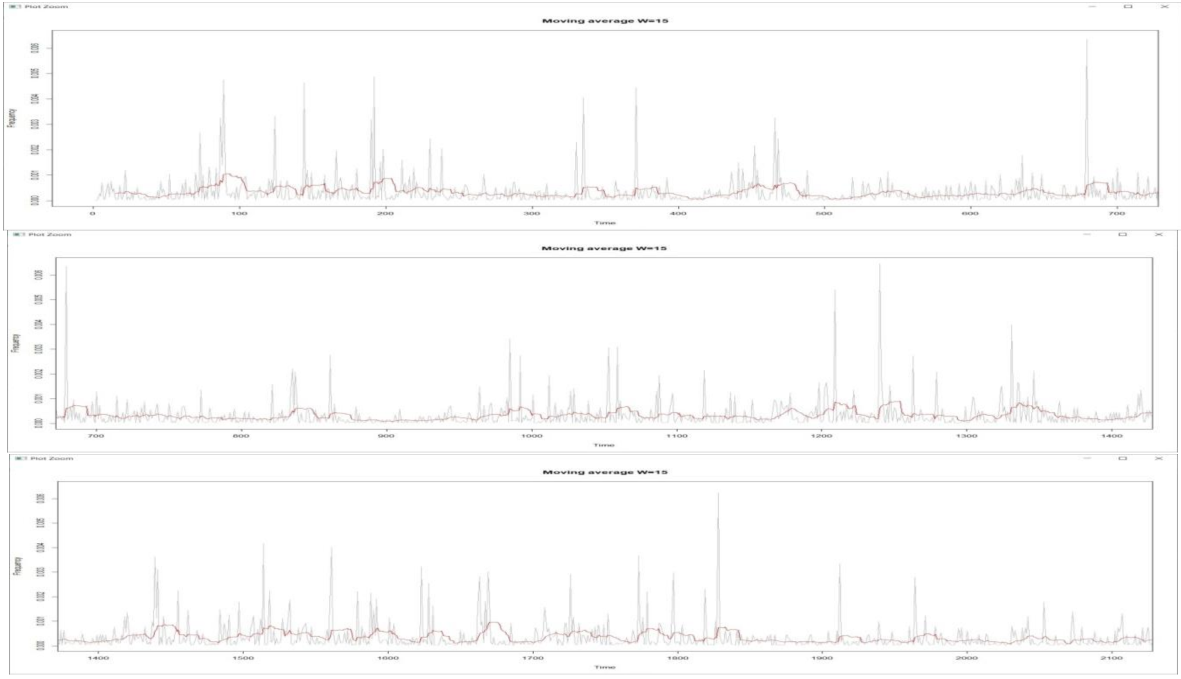


Figure 10: Moving Average of Article 2 for w=3 for the interval 0-700, 700-1400 and 1400-2100

It is needed to smooth the data using the size of the smoothing interval w = 3 for Article 2 (Fig. 25-27), then smooth the obtained data again but using the size of the smoothing interval w = 5. Then continue smoothing the obtained data with the smoothing interval w = 7 and so on to w = 15.

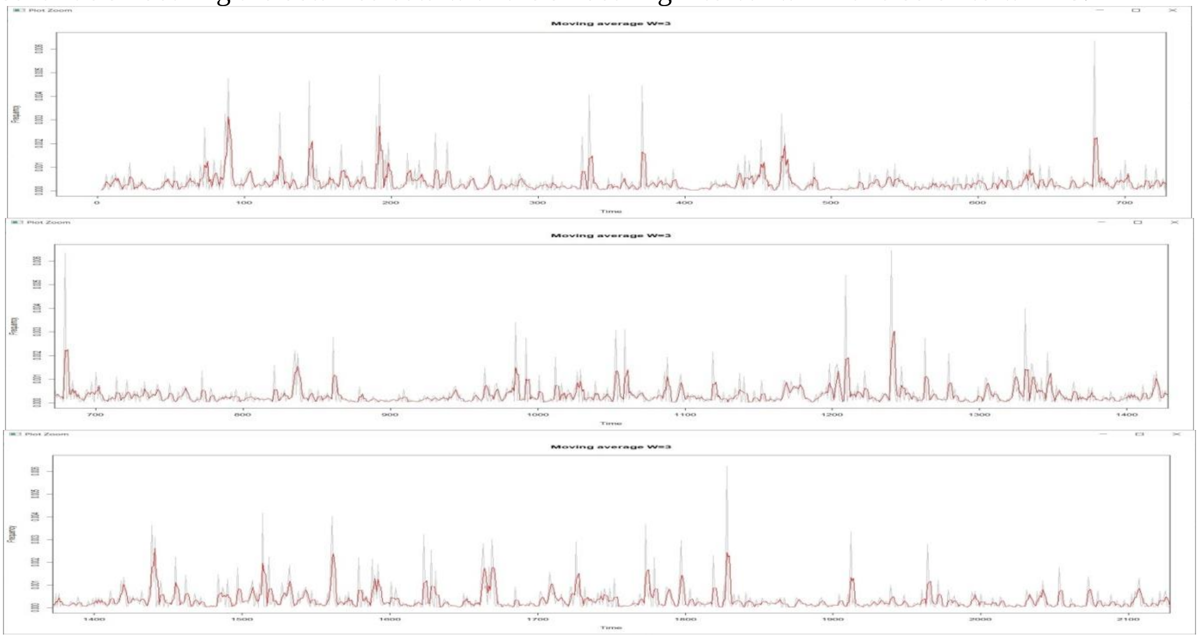


Figure 11: Moving Average of Article 2 for w=3 for the interval 0-700, 700-1400 and 1400-2100

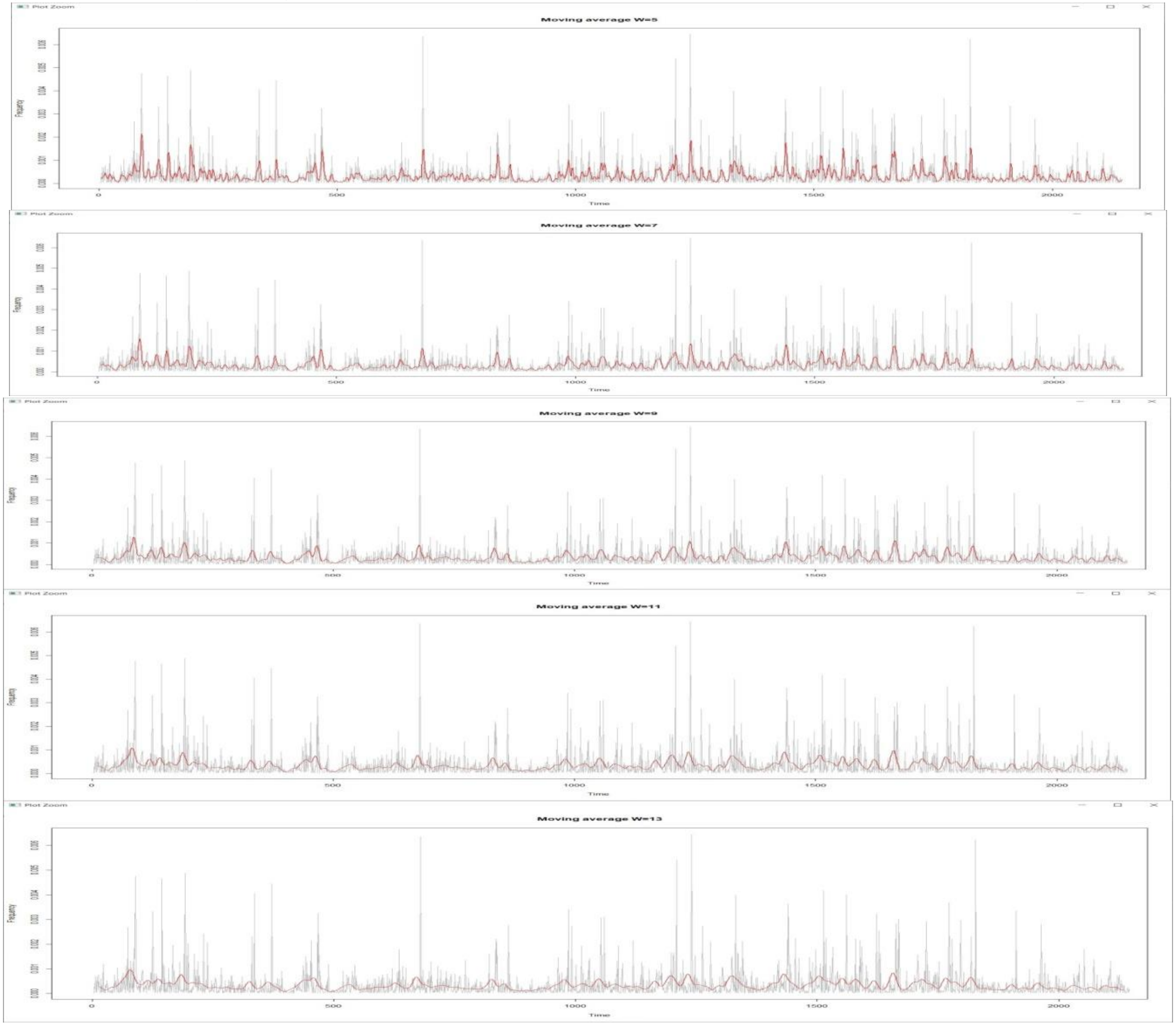


Figure 12: Moving Average of Article 2 for w = 5, 7, 9,11, 13

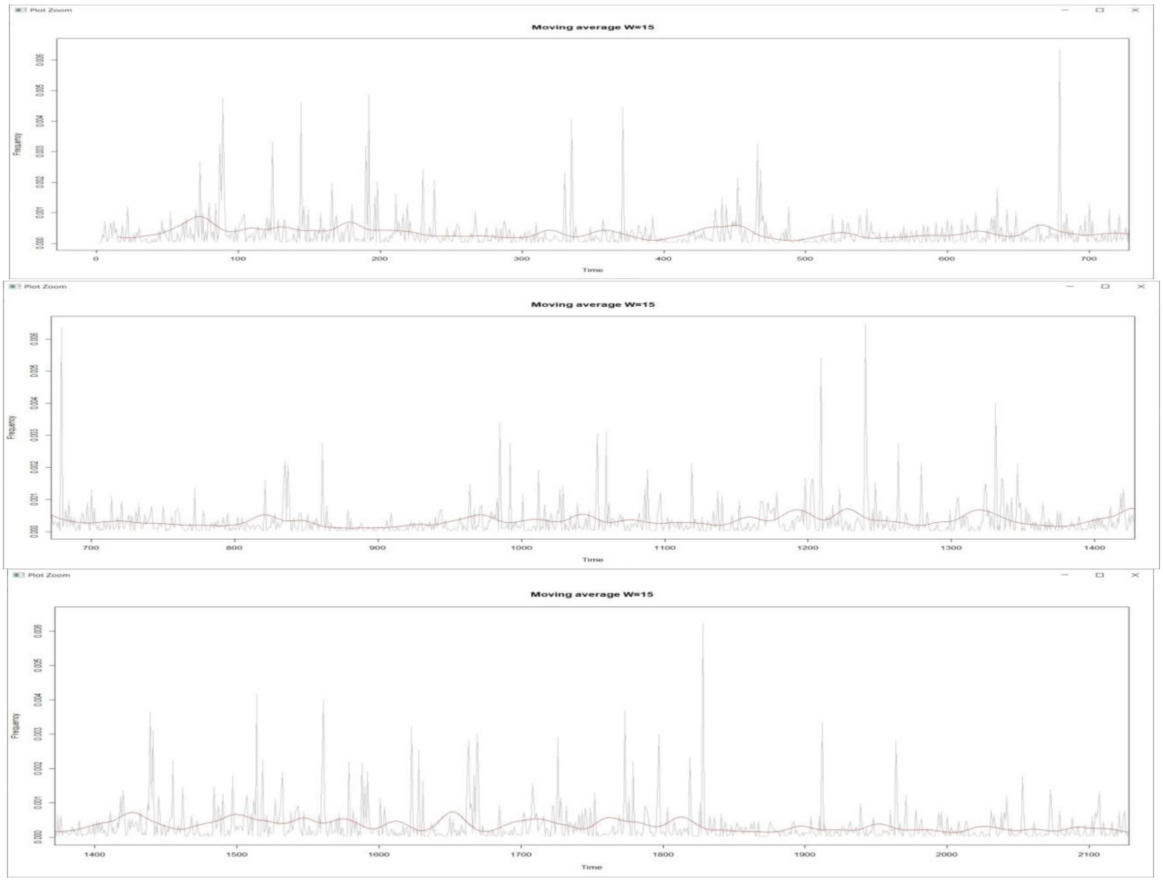


Figure 13: The Moving Average of Article 2 for w=3 for the interval 0-700, 700-1400 and 1400-2100

Smooth the data using the size of the smoothing interval w = 3, 5, 7, 9,11, 13, 15 for Article 3, the moving average smoothed the data with about the same efficiency as in Article 1 (Fig. 28-10).

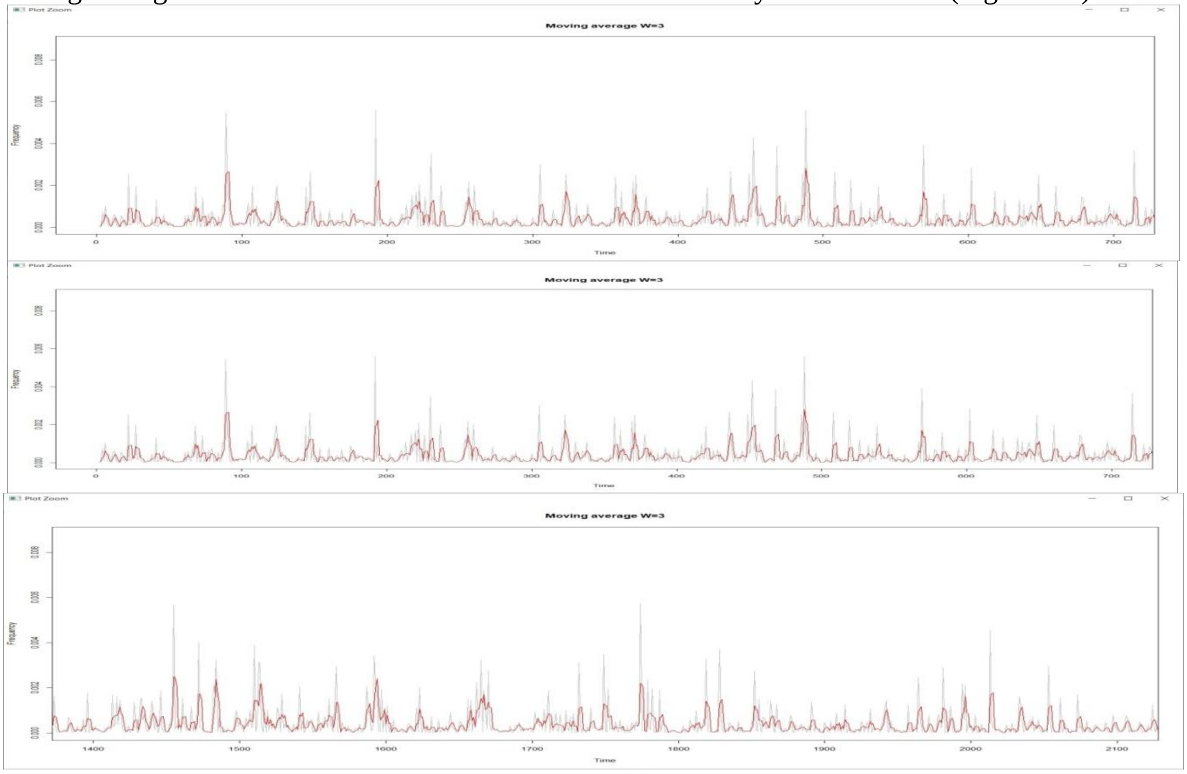


Figure 14: The Moving Average of Article 3 for w=3 for the interval 0-700, 700-1400 and 1400-2100

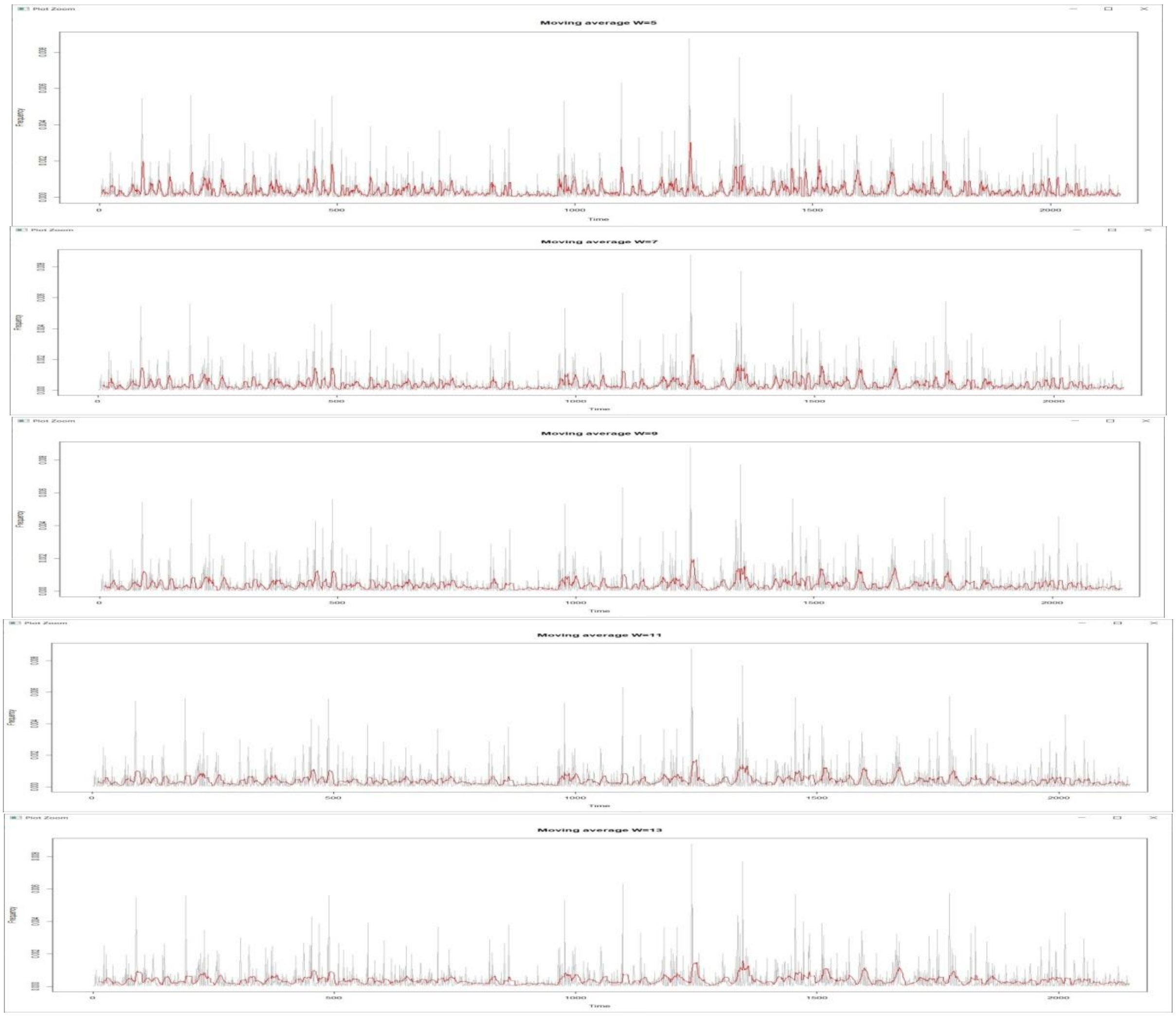


Figure 15: The Moving Average of Article 3 for w = 5, 7, 9,11, 13

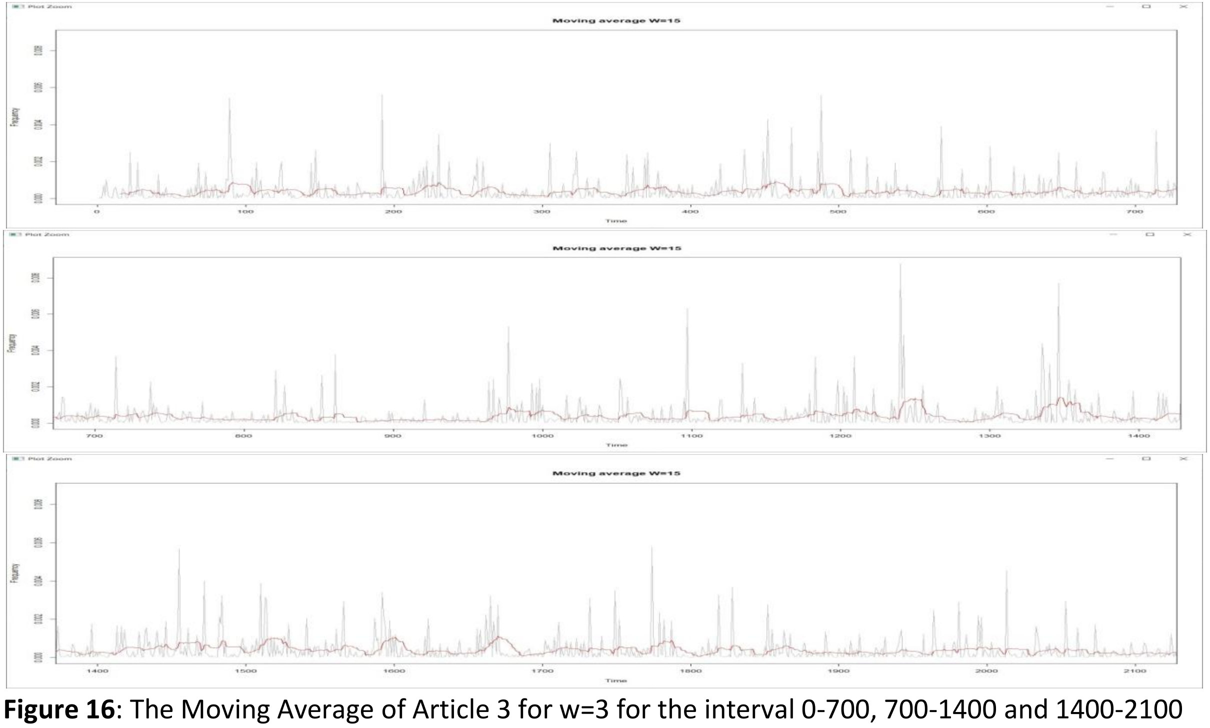


Figure 16: The Moving Average of Article 3 for w=3 for the interval 0-700, 700-1400 and 1400-2100

It is needed to smooth the data using the size of the smoothing interval w = 3 (Fig. 31), then smooth the obtained smoothed data again but using the size of the smoothing interval w = 5. Then continue smoothing the obtained data with the smoothing interval w = 7 and so on to w = 15 (Fig. 32-33).

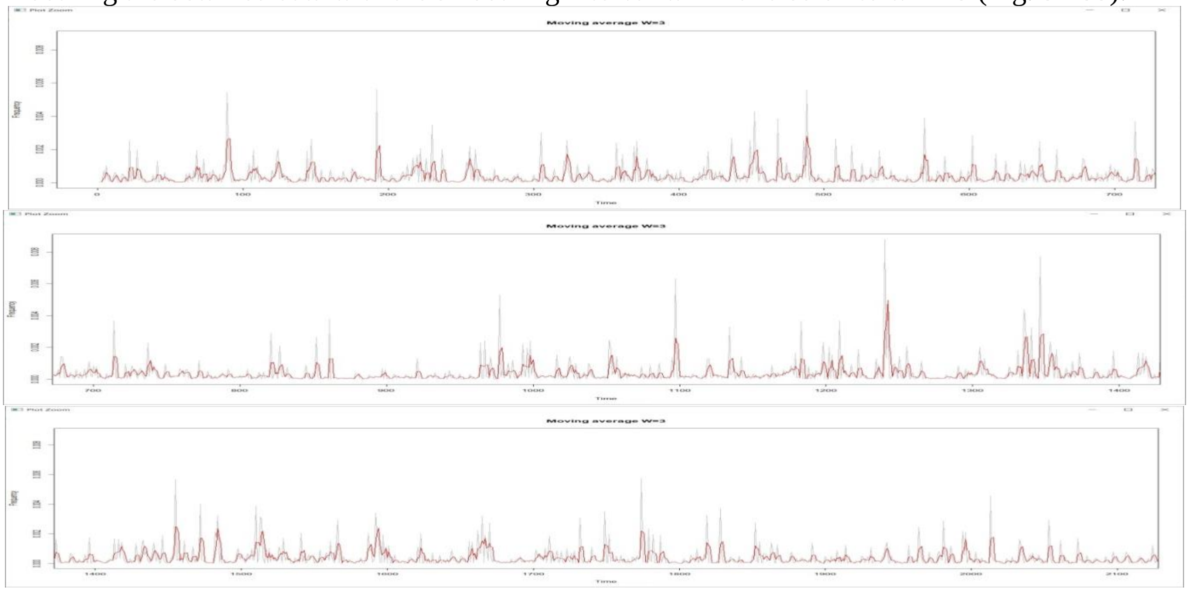


Figure 17: The Moving Average of Article 3 for w=3 for the interval 0-700, 700-1400 and 1400-2100

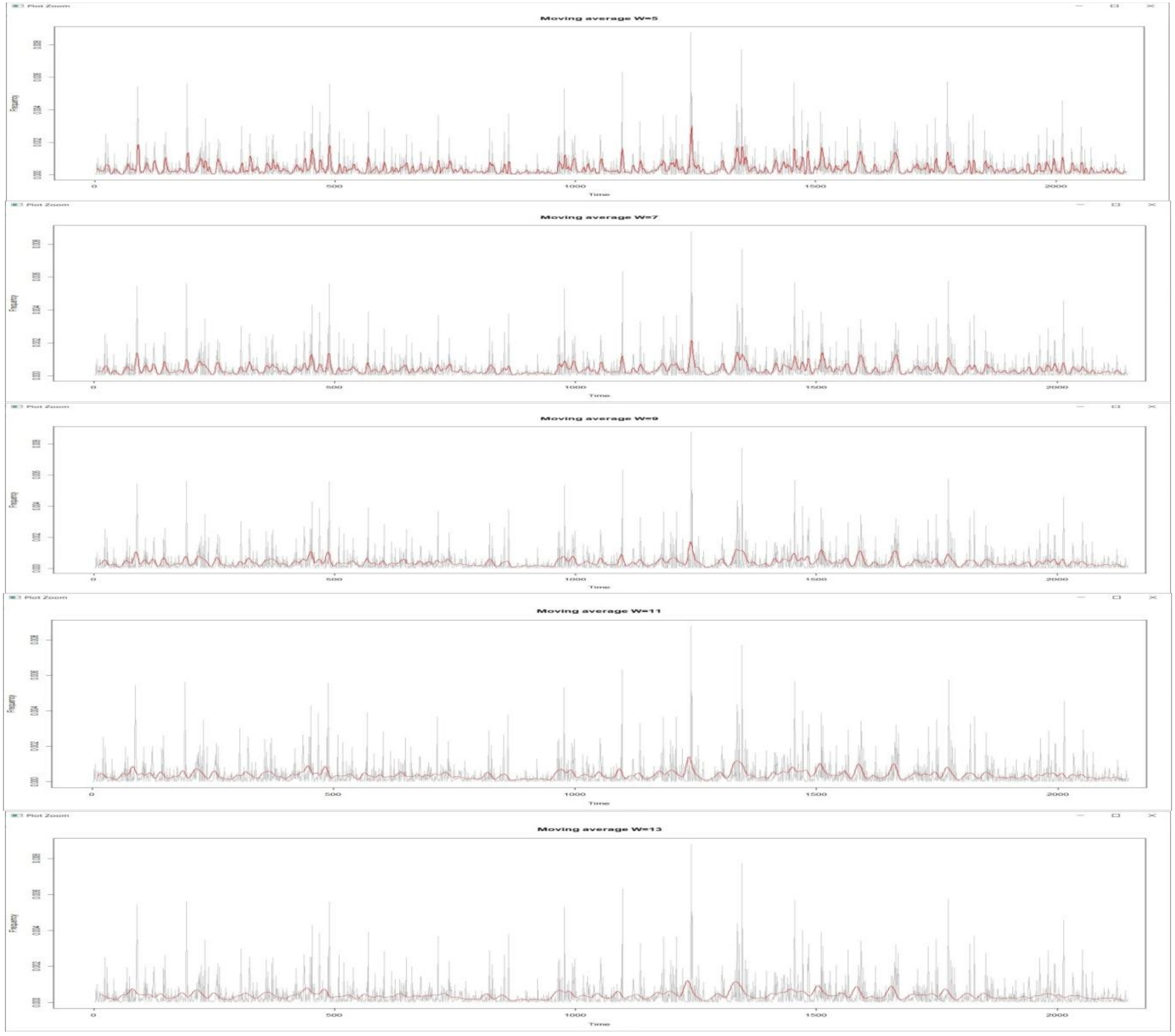


Figure 18: The Moving Average of Article 3 for w = 5, 7, 9,11, 13

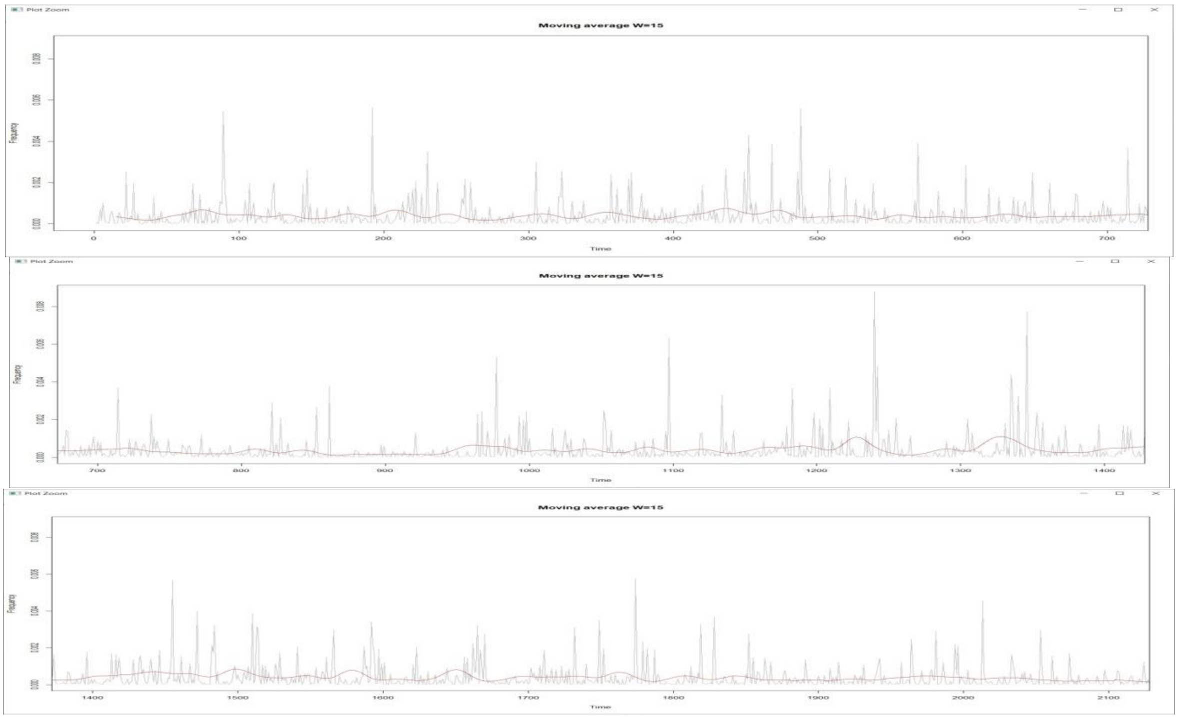


Figure 19: The Moving Average of Article 3 for w=3 for the interval 0-700, 700-1400 and 1400-2100

Exponential smoothing. The main parameter of exponential smoothing is a parameter that takes values in the range of 0.1 0.3. It is necessary to perform smoothing of the same series with the values of the parameter = 0.1, 0.15, 0.2, 0.25, 0.3 and in all these cases to find for each smoothing the correlation coefficients between the original values and smoothed ones. An essential feature of the use of exponential averages is the substantiation of the value of the smoothing parameter α. The smaller it is, the more the levels in the analyzed series are smoothed. This means an increase in the specific α.

For data on the frequency of 3-grams use exponential smoothing for all three texts did not give a "delay". Exponential smoothing has softened these data a little and it is harder to see the general trend. Also, the correlation coefficients of the data are very low (Fig. 34-42).

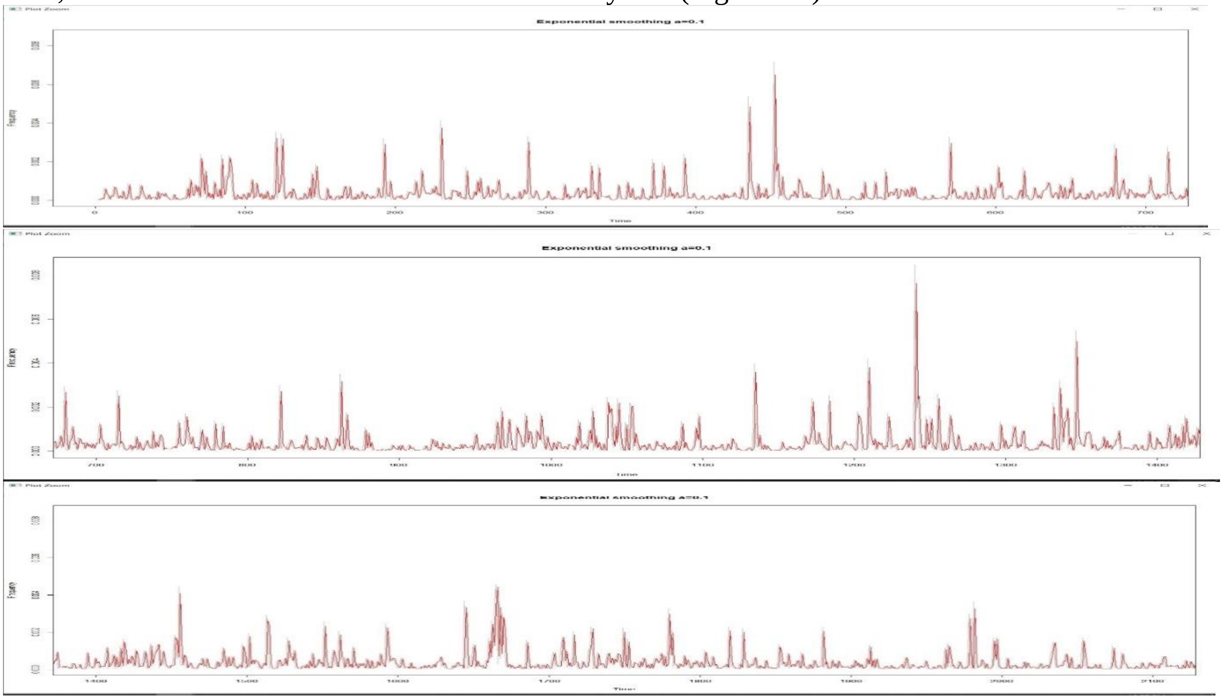


Figure 20: Exponential smoothing a=0.1 of Article 1 for the interval 0-700, 700-1400, 1400-2148

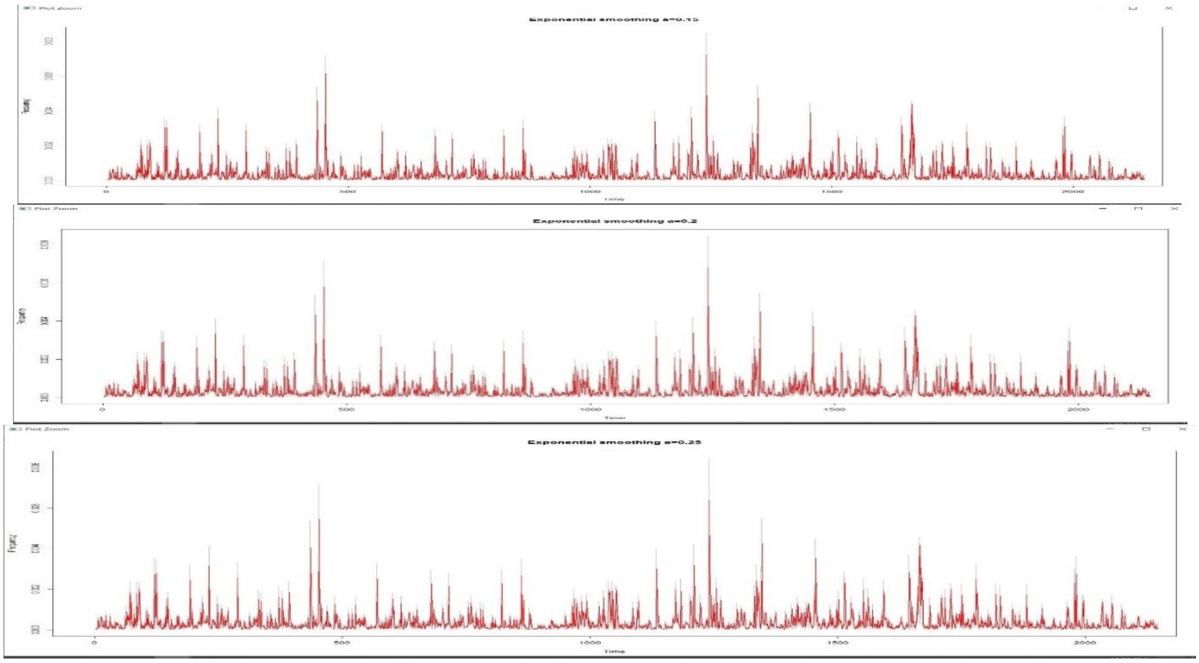


Figure 21. Exponential smoothing of Article 1 for a=0.15, 0.2, 0.25

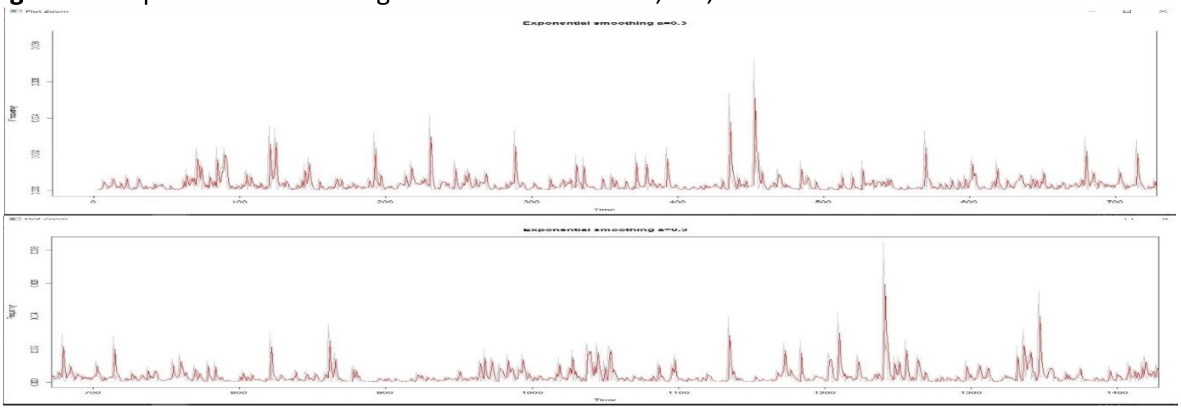


Figure 22: Moving Average of Article 1 for a=0.3 for the interval 0-700 and 700-1400

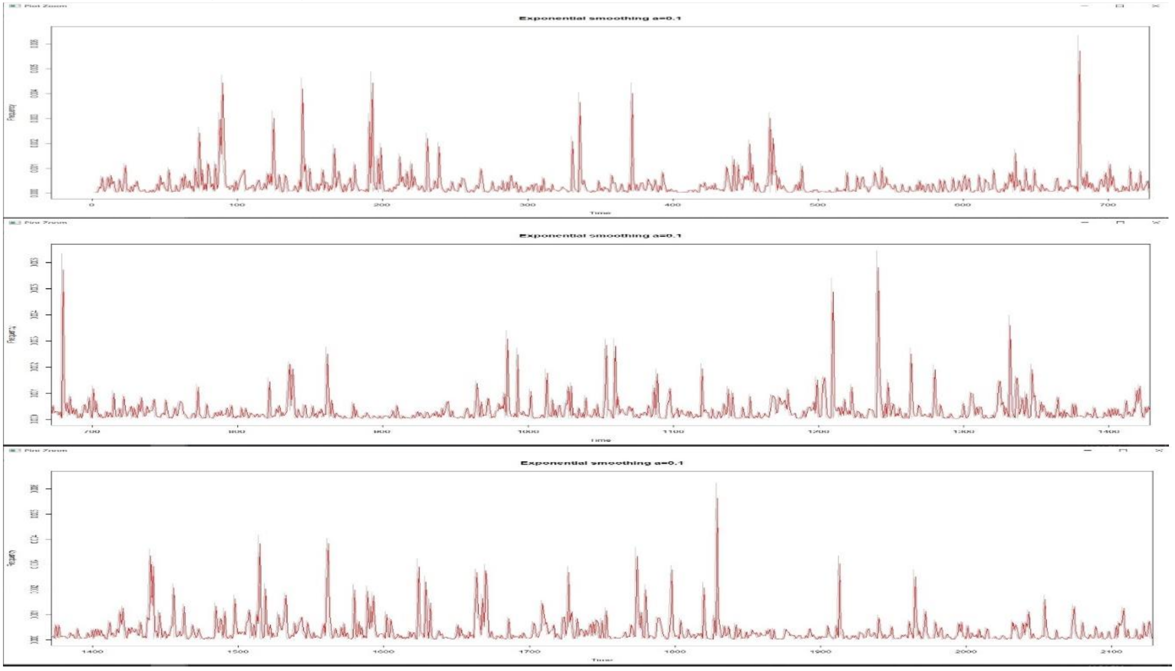


Figure 23: Exponential smoothing a=0.1 of Article 2 for the interval 0-700, 700-1400, 1400-2148

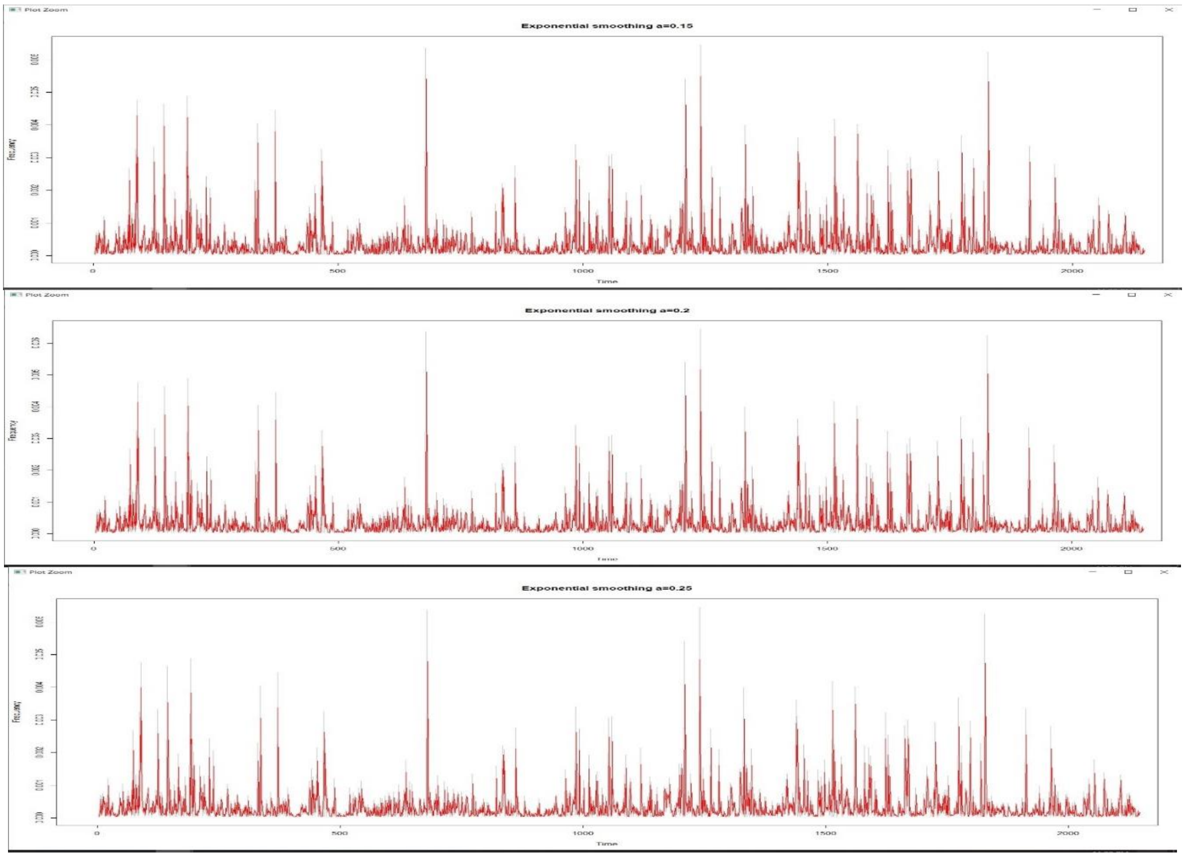


Figure 24 Exponential smoothing of Article 2 for a=0.15, 0.2, 0.25

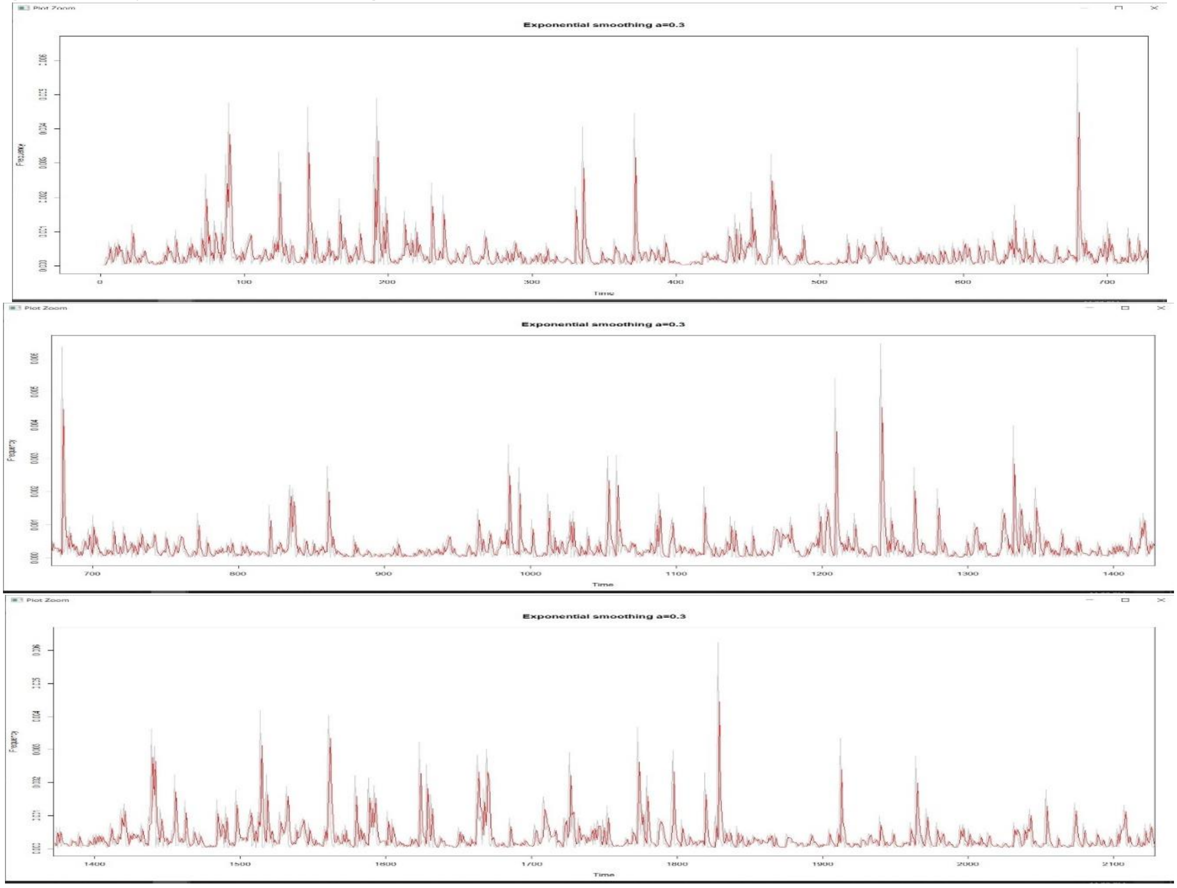


Figure 25: Exponential smoothing a=0.3 of Article 2 for the interval 0-700, 700-1400, 1400-2148

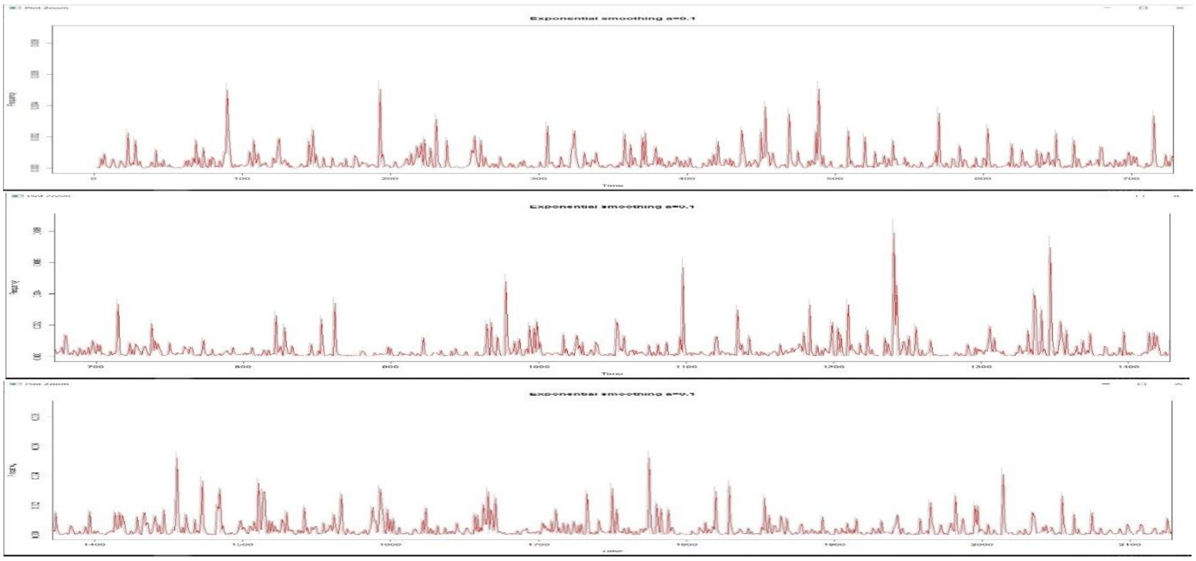


Figure 26: Exponential smoothing a=0.1of Article 3 for the interval 0-700, 700-1400, 1400-2148

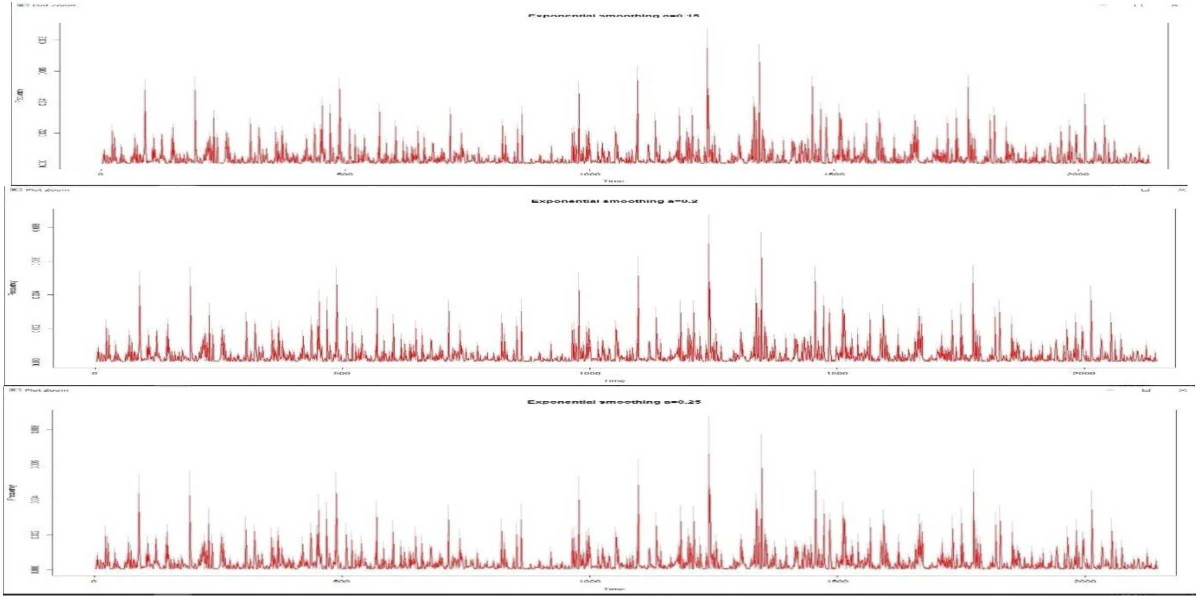


Figure 27 Exponential smoothing of Article 3 for a=0.15, 0.2, 0.25

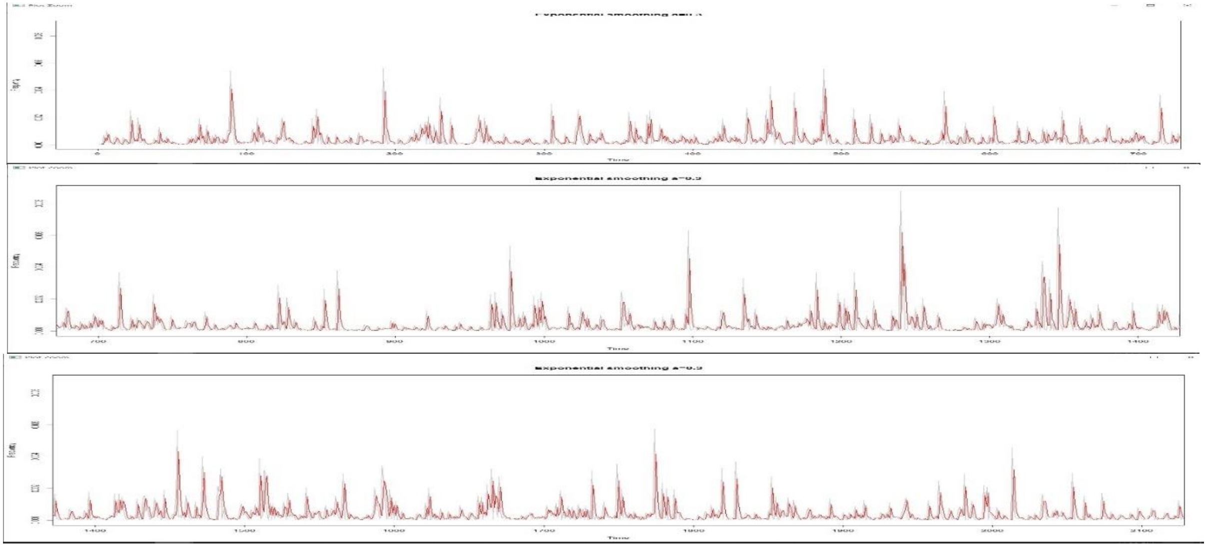


Figure 28: Exponential smoothing a=0.3 of Article 3 for the interval 0-700, 700-1400, 1400-2148

Median smoothing. In this case, use the same dimensions of the smoothing interval and the operation as in paragraph 1.Characteristic feature of median smoothing is that it leaves monotonic parts of the data sequence and sharp differences unchanged, and for nonmonotonic areas within the size of the sliding smoothing interval leaves only a centered value equal to their median, i.e. effectively eliminates those levels that violate monotonicity. It is needed completely to eliminate single extreme or anomalous values of levels that are at least half the distance from the smoothing interval, maintain 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. These characteristics of the median smoothing were confirmed during the median smoothing for relative frequency in Article 1. lower than the moving average. We smooth the data using the dimensions of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 (Fig. 43-45). Graphics are arranged in the appropriate order: Median w = 3 (for intervals 0-700, 700-1400, 1400-2148), Median w = 5, 7, 9, 11, 13, Median w = 15 (for intervals 0-700, 700-1400, 1400-2148).

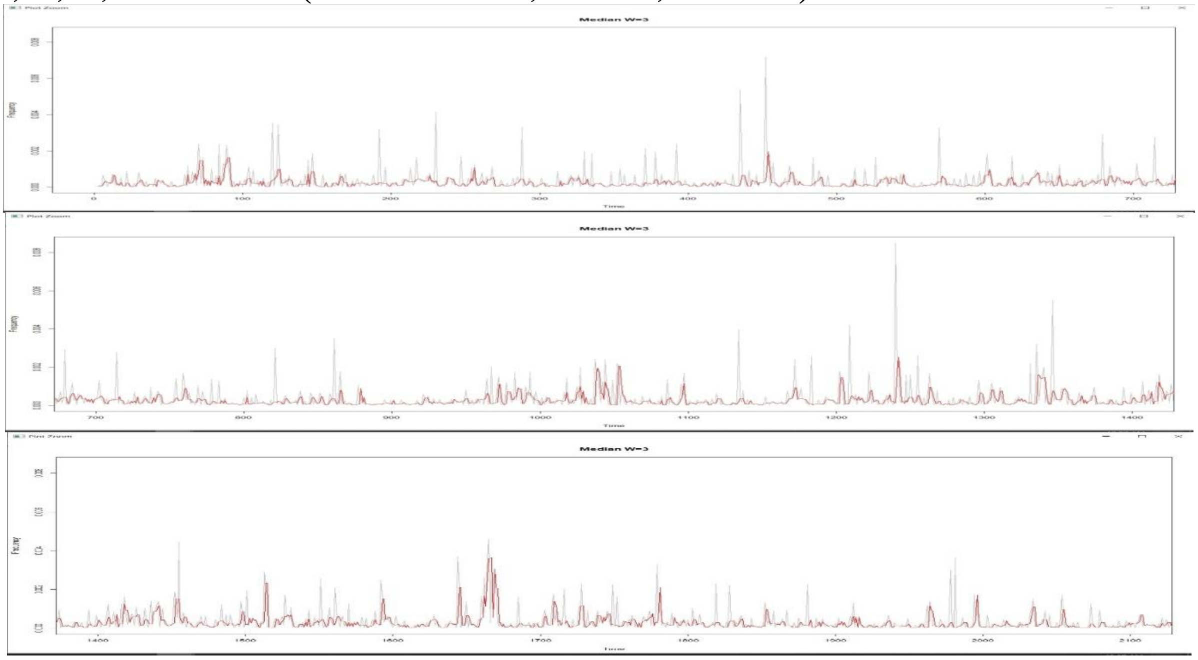


Figure 29: Median smoothing w = 3 of Article 1 for the interval 0-700, 700-1400, 1400-2148

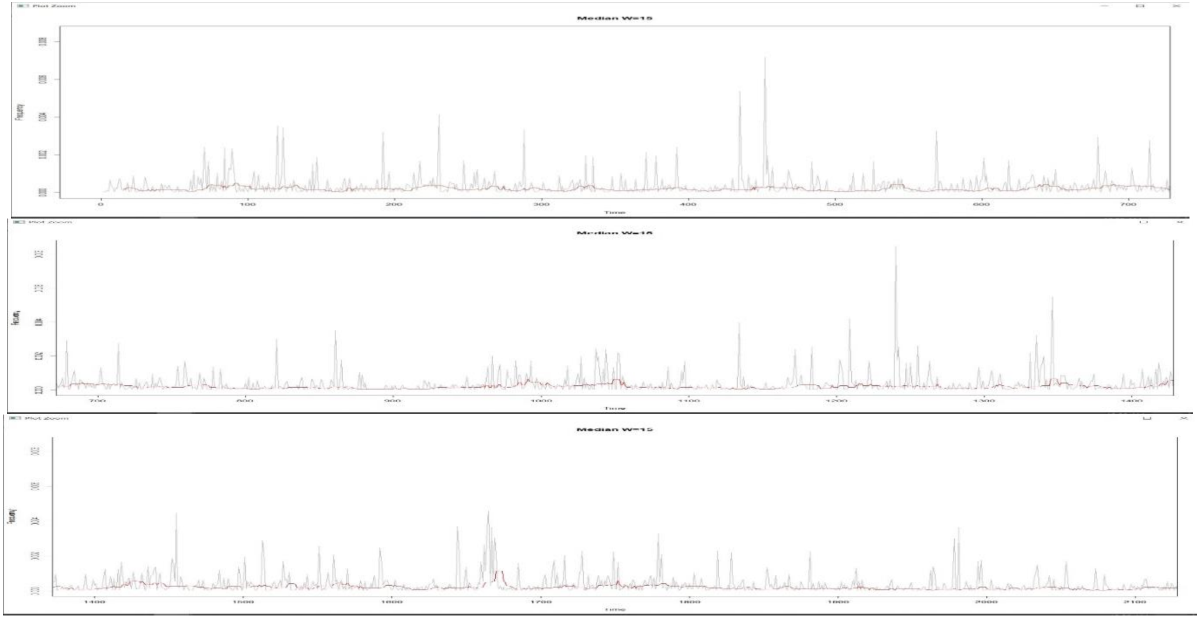


Figure 30: Median smoothing w = 15 of Article 1 for the interval 0-700, 700-1400, 1400-2148

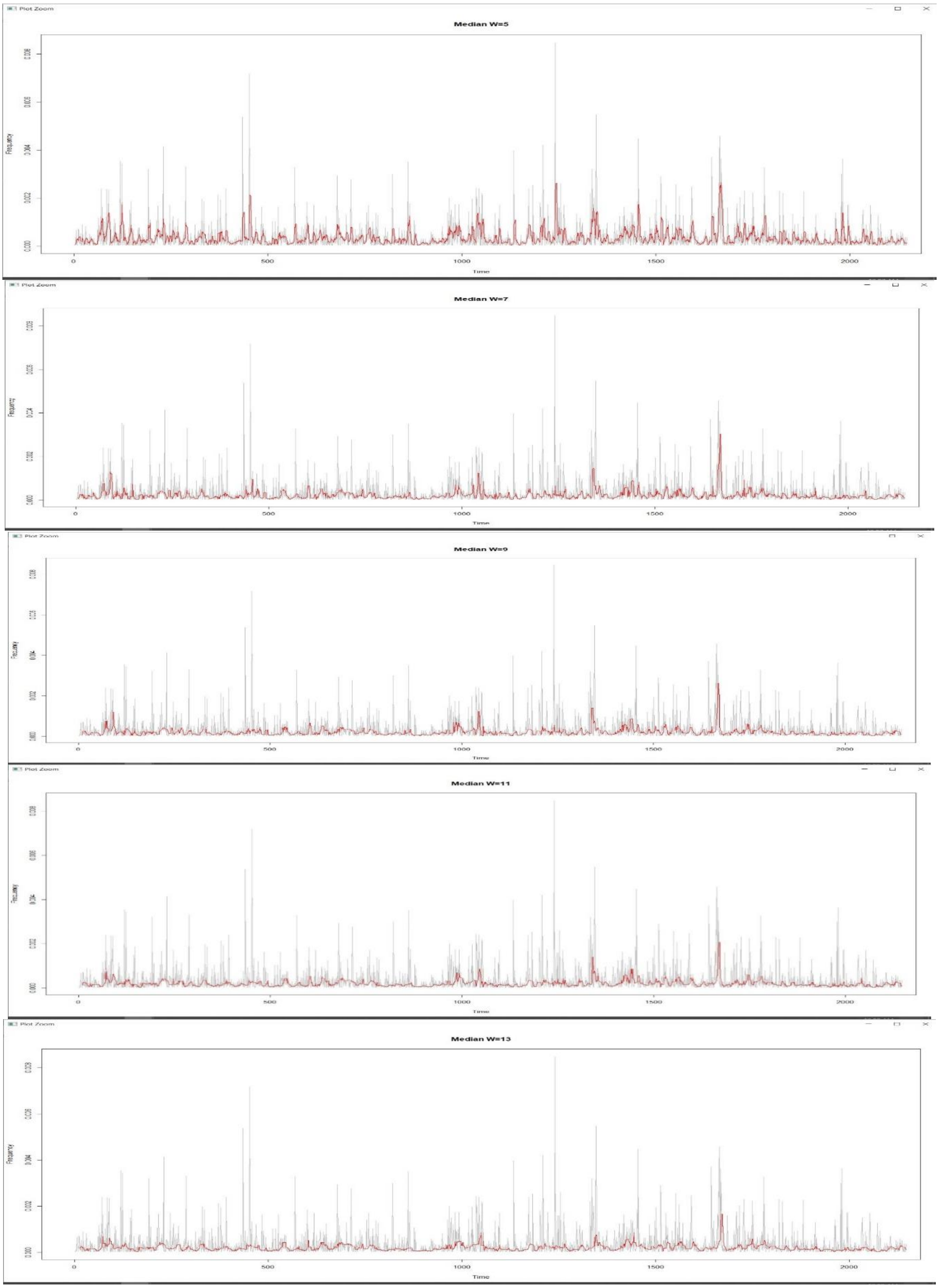


Figure 31 Median smoothing of Article 1 for w = 3, 5, 7, 9, 11, 13

It is needed to smooth the data using the size of the smoothing interval w = 3, then smooth the obtained smoothed data again but using the size of the smoothing interval w = 5. We continue smoothing the obtained data with the smoothing interval w = 7 and so on to w = 15 (Fig. 46-48).

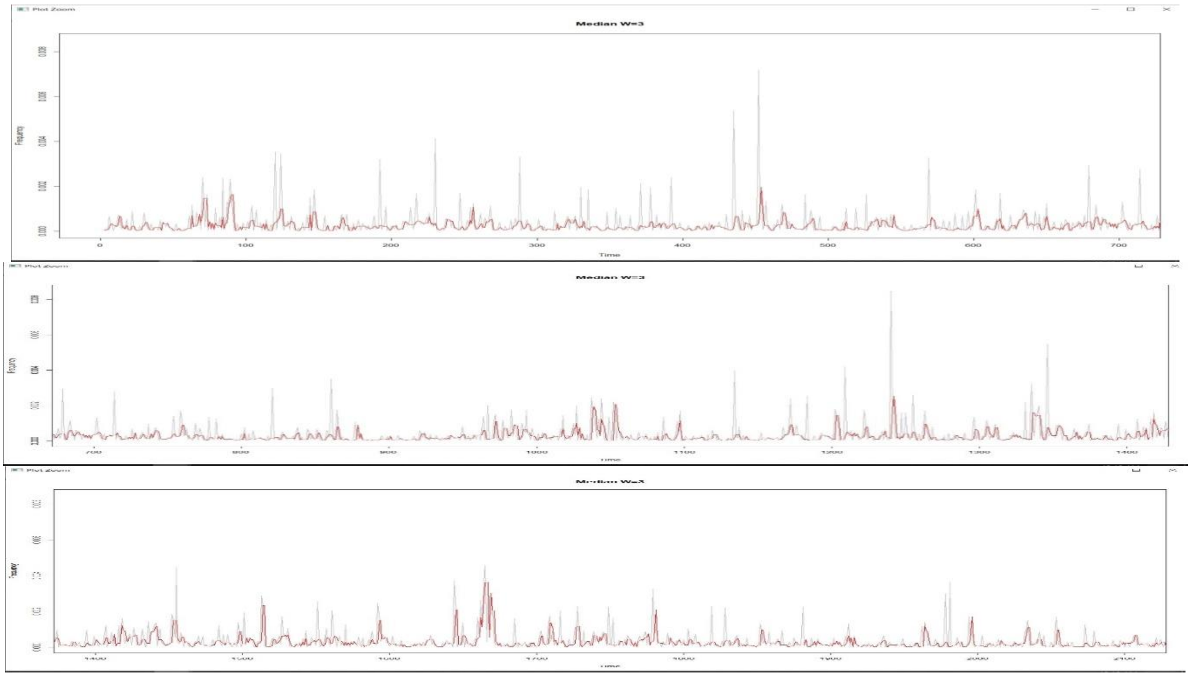


Figure 32: Median smoothing w = 3 of Article 1 for the interval 0-700, 700-1400, 1400-2148

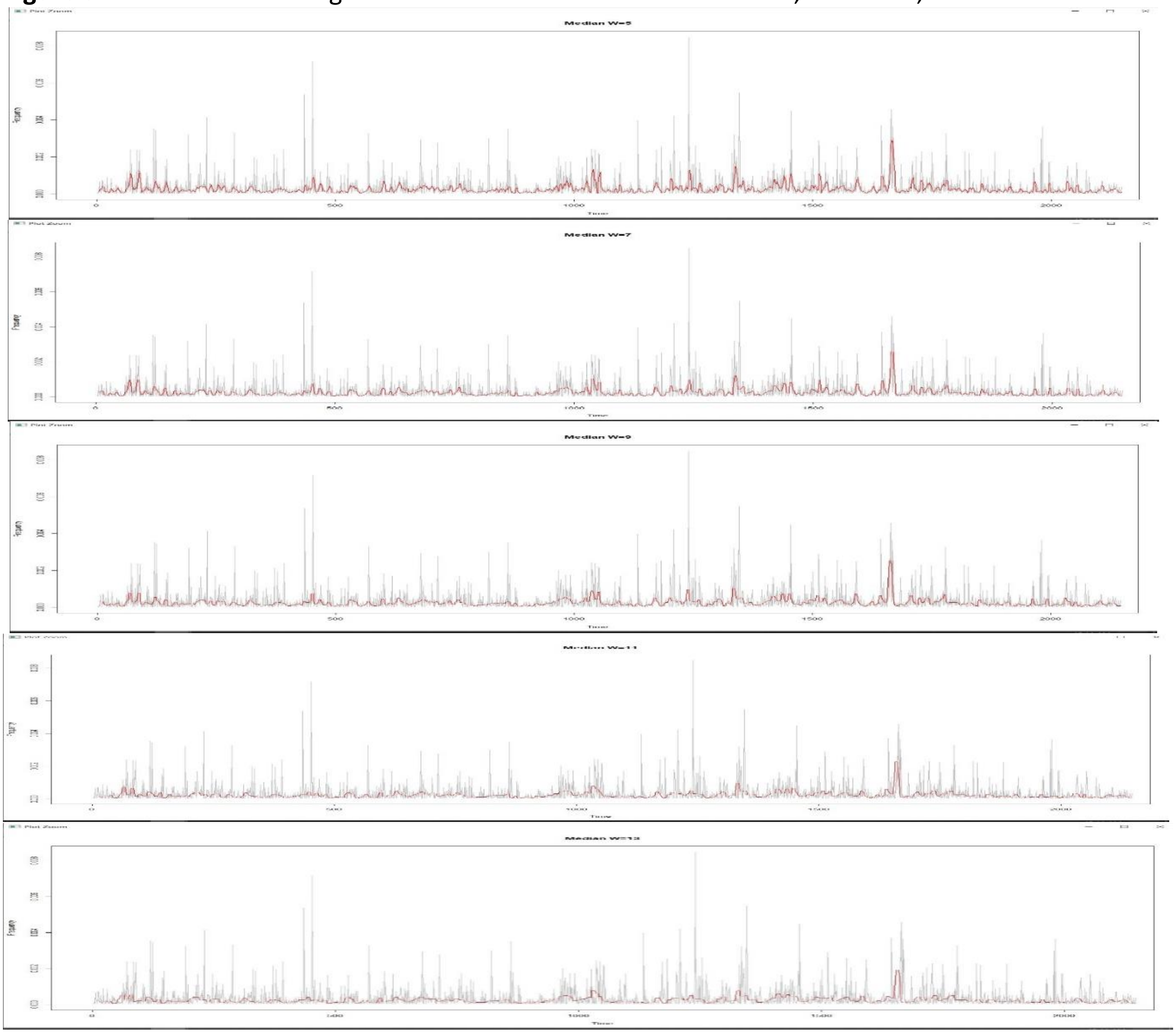


Figure 33 Median smoothing of Article 1 for w = 3, 5, 7, 9, 11, 13

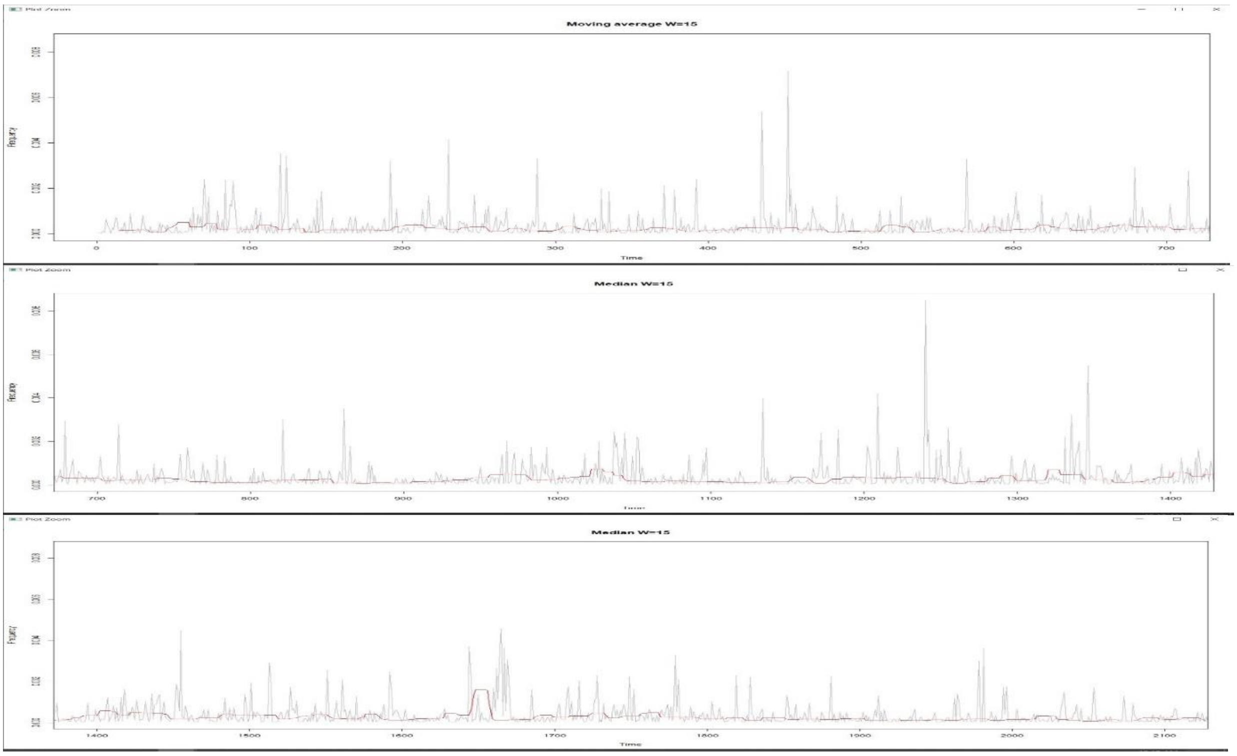


Figure 34: Median smoothing w = 15 of Article 1 for the interval 0-700, 700-1400, 1400-2148

6. Discussions

Graphical representation of the relationship between two studied sequences is called a correlation field or scatter plot. The graphical method provides a visual representation of the form of communication between these sequences. So, it is needed to construct a correlation field for Article 1 and 2 (Fig. 49), Article 1 and 3 (Fig. 50), Article 2 and 3 (Fig. 51).

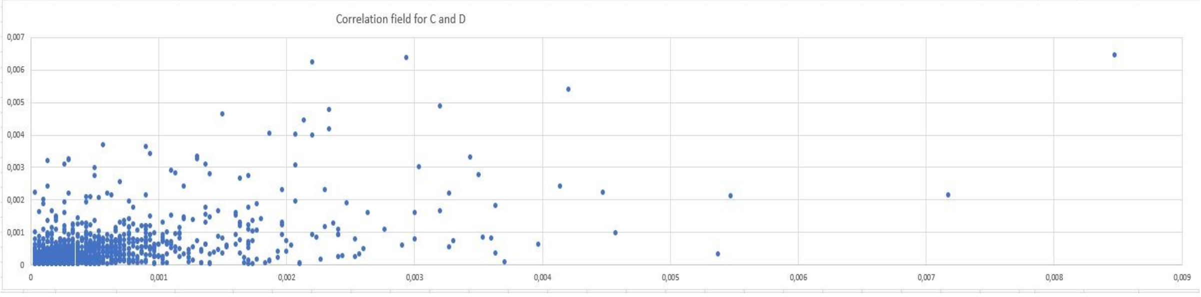


Figure49: Correlation field for Articles 1 and 2

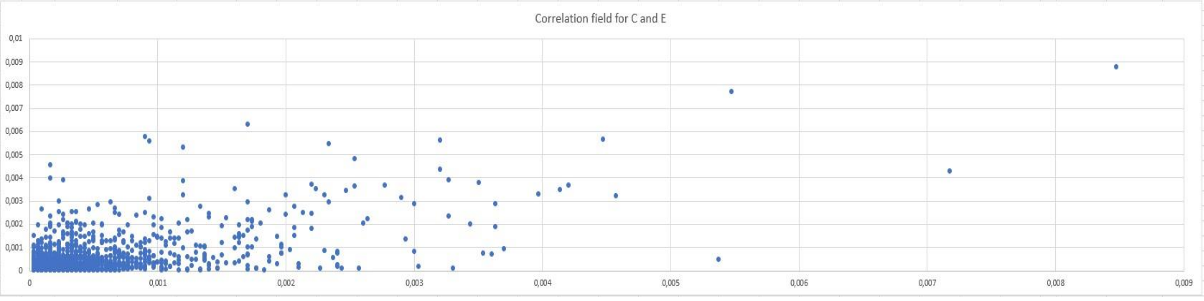


Figure50: Correlation field for Articles 1 and 3

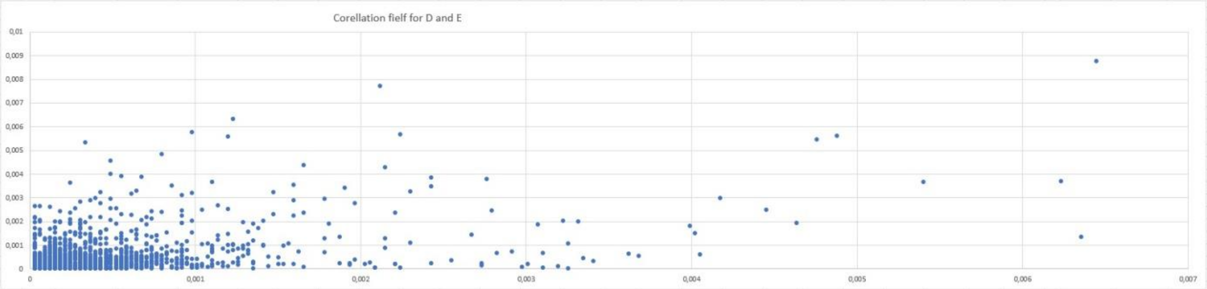


Figure51: Correlation field for Articles 2 and 3

Visually assessing the nature of the relationship, it can be stated that there is a linear relationship in all three fields. Also evaluating the visual data of the field, we see that the correlation is present, so we can assume that these Ukrainian articles can be written by one author or are based on one topic. But visual assessment is not enough, so it is worth finding the value of the correlation coefficient for more accurate research results. The correlation coefficient characterizes the degree of closeness of the linear dependence. Therefore, there is a calculation of the correlation coefficients for Articles 1 and 2 (Correlation coefficient 0.575. Coefficient of determination 33%); for Articles 1 and 3 (Correlation coefficient 0.63023. Coefficient of determination40%); for Articles2 and 3 (Correlation coefficient 0.49038. Coefficient of determination24%). Correlation coefficients that are less than 0.7 but greater than 0.5 modulus indicate a medium-strength relationship (the coefficients of determination are less than 50% but more than 25%). It is worth noting that in the first two cases we received a connection of medium strength and in the third case we have a connection of weak force very close to the average, so it can also be attributed to the average. It is obvious that having three different Ukrainian articles the 100% correlation is unlikely to be. So, given the average connection, the assumption that these articles may have been written by the same author or are based on the similar topics has been confirmed. When the pair statistical dependence on the linear correlation is rejected, the correlation coefficient loses its meaning as a characteristic of the degree of closeness of the connection. In this case, such a measure of communication as the correlation ratio is used. Since there is a linear relationship between the pair of studied features, the correlation ratio does not need to be calculated.

Autocorrelation function is a correlation of function with itself shifted by a certain amount of independent variables. Autocorrelation is used to find patterns in a number of data, such as periodicity.

The graph of the autocorrelation function is also called the correlogram (Fig. 52).

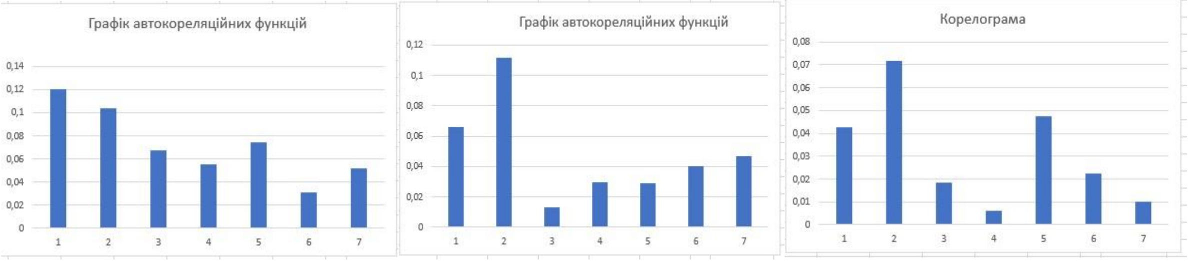
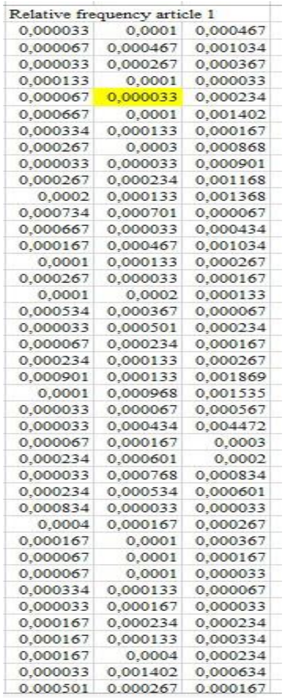


Figure 35: Correlogram

Fig. 52 shows that the studied series are not stationary, as in the case of fixed time series the graph of autocorrelation functions should be decreased rapidly after the first few values.

It is needed to divide the sequence of Relative frequency Article 1 into three equal parts of 715 values. For convenience, we take the data into a separate table (Fig. 53). The correlation matrix is a square table in which the correlation coefficient between the corresponding parameters is located at the intersection of the corresponding row and column. Correlation matrix for column divided into 3 parts and has been constructed and the results are obtained: correlation coefficients, that are less than 0.5, the absolute value or modulus indicate a weak relationship. On the correlation matrix it is seen that all values are close to 0, so we can conclude that there is no connection at all. It can be said that this is quite an expected result, as the data do not depend on each other and have different values. We find the coefficients of multiple correlation (Fig. 54-55).



Relative frequency article 1		
0,000033	0,0001	0,000467
0,000067	0,000467	0,001034
0,000033	0,000267	0,000367
0,000133	0,0001	0,000033
0,000067	0,000033	0,000234
0,000667	0,0001	0,001402
0,000334	0,000133	0,000167
0,000267	0,0003	0,000868
0,000033	0,000033	0,000901
0,000267	0,000234	0,001168
0,0002	0,000133	0,001368
0,000734	0,000701	0,000067
0,000667	0,000033	0,000434
0,000167	0,000467	0,001034
0,0001	0,000133	0,000267
0,000267	0,000033	0,000167
0,0001	0,0002	0,000133
0,000534	0,000367	0,000067
0,000033	0,000501	0,000234
0,000067	0,000234	0,000167
0,000234	0,000133	0,000267
0,000901	0,000133	0,001869
0,0001	0,000968	0,001535
0,000033	0,000067	0,000567
0,000033	0,000434	0,004472
0,000067	0,000167	0,0003
0,000234	0,000601	0,0002
0,000033	0,000768	0,000834
0,000234	0,000534	0,000601
0,000834	0,000033	0,000033
0,0004	0,000167	0,000267
0,000167	0,0001	0,000367
0,000067	0,0001	0,000167
0,000067	0,0001	0,000033
0,000334	0,000133	0,000067
0,000033	0,000167	0,000033
0,000167	0,000234	0,000234
0,000167	0,000133	0,000334
0,000167	0,0004	0,000234
0,000033	0,001402	0,000634
0,000501	0,000267	0,000167

1		
0,055450017	1	
-0,027816863	-0,06816395	1

Figure53: The column is divided into 3 equal parts and Correlation matrix

1	0,120173631	1	0,066002628	1	0,042590192
2	0,104051246	2	0,111779282	2	0,071599169
3	0,067186819	3	0,012826123	3	0,018317132
4	0,055053896	4	0,029412894	4	0,006056347
5	0,074429898	5	0,028723036	5	0,047593571
6	0,030903496	6	0,040345677	6	0,022404806
7	0,0515167	7	0,046982752	7	0,010037348

Figure54: Multiple correlation coefficients

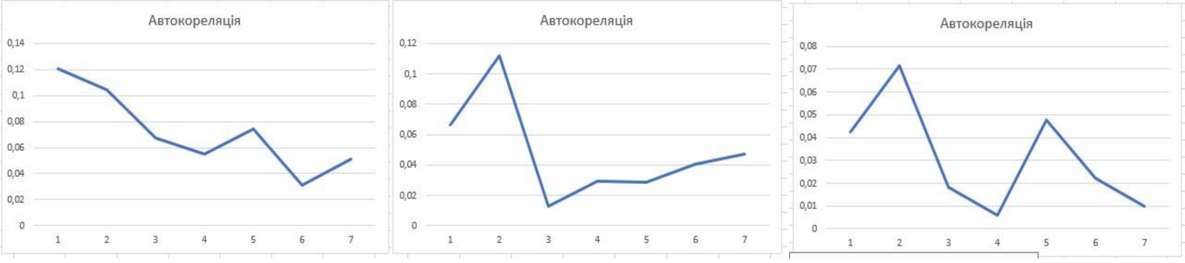


Figure 36: Autocorrelation

According to these graphs, Article 1 and Article 2 were more likely to have been written by one author, although Article 1 and Article 3 could also have been written by one author (but this is not true). But Articles 2-3 were definitely written by different authors. The application of linguistic and statistical analysis of 3-grams to a set of articles will allow to form a subset of similar linguistically characteristic publications. Imposing additional conditions on this subset in the form of statistical and quantitative analyzes (sets of keywords, stable phrases, stylistic, linguometric, etc.) will significantly reduce this subset, clarifying the list of more likely author works. Thus, the analysis of the content and frequency of occurrence of only business words will separate Articles 1 and 3 into different subsets, Articles 1 and 2 in one the same. This study does not address the problem of identifying the author in full due to the fact that the difference in authorial traits is subjective and depends on the limitations imposed on the creative process of the author. However, as a result, a system that implements such methods is able to give recommendations on the degree of belonging of the text to a particular author. Further experimental research is needed to test the proposed method to determine the style of the author from other categories of texts - scientific humanities, art, journalism and more. Therefore, we compare the frequencies of all trigrams that begin with a particular letter (Fig. 56).

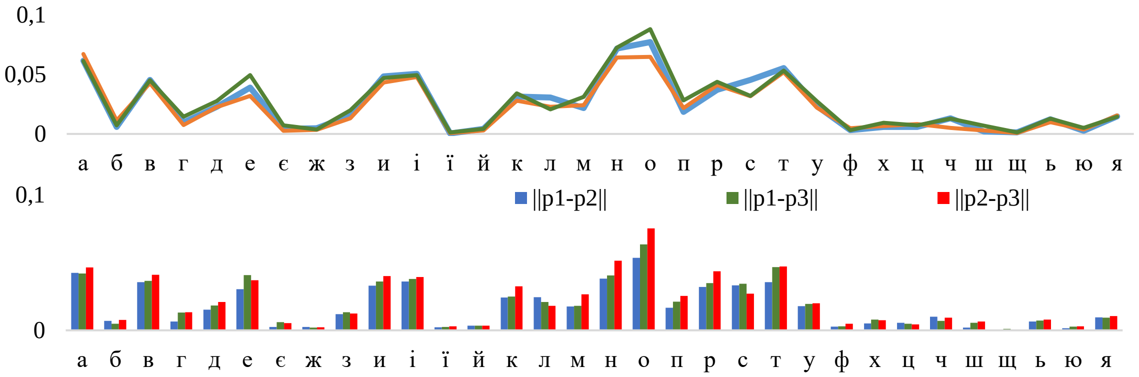


Figure56: The 3-gram usage that starts with a specific letter (Article 1 – blue, Article 2 – red, Article 3 – green)

According to these graphs, Article 1 and Article 2 were more likely to have been written by one author, although Article 1 and Article could also have been written by one author (but this is not true). But Articles 2-3 were definitely have been written by different authors. The application of linguistic and statistical analysis of 3-grams to a set of articles will allow to form a subset of similar linguistically characteristic publications. Imposing additional conditions on this subset in the form of statistical and quantitative analyzes (sets of keywords, stable phrases, stylistic, linguometric, etc.) will significantly reduce this subset, clarifying the list of more likely author works.

Thus, the analysis of the content and frequency of occurrence of only business words will separate Articles 1 and 3 into different subsets, Articles 1 and 2 in one the same.

This study does not address the problem of identifying the author in full due to the fact that the difference in authorial traits is subjective and depends on the limitations imposed on the creative process of the author. However, as a result, a system that implements such methods is able to give recommendations on the degree of the text belonging to a particular author. Further experimental research needs to test the proposed method to determine the style of the author from other categories of texts such as scientific humanities, art, journalism and others.

7. Conclusions

The article dwells upon the completed scientific research in the field of information technology in the part concerning computer linguistics, artificial intelligence and Machine Learning. Correlation analysis of text author identification results based on n-grams in Ukrainian technical and scientific texts have been made. The comparison between three articles have been done and the results have been obtained. Quantitative content analysis of textual scientific and technical content has been studied based on the fact that text authorship determination systems typically use plagiarism and rewrite its metrics of identification fully or partially. The article presents the method of determining the author by decomposition on the basis of the analysis of such speech coefficients as lexical diversity, degree of syntactic complexity, speech coherence, indices of exclusivity and concentration of the text. Also, the parameters of the author style such as words, sentences, prepositions, conjunctions numbers and quantity of words with defined frequencies have been analyzed. It is highlighted that in the algorithmic approach smoothing procedures are widely used. So, the relative frequency of 3-grams consumptions in the studied texts has been smoothed by the method of moving average, exponential and median smoothing. It is proposed to analyze the reference text in several stages for high-quality and effective analysis of content in determining the degree of text authorship. To achieve the research goal a system with the ability to select the language / languages of the analyzed content have been developed and implemented on the Victana Web-resource. It is said that in order to compare the texts with each other it is necessary to compare the text with some numerical characteristic that was close to the texts of the same author and would different in the works of various authors that uses the distribution function density of letter combinations of three consecutive characters. So, rapid distribution of text documents in electronic form has caused the importance of using automatic methods to analyze the content including the necessity of documents classification and clustering by various criteria.

8. References

[1] S. O. Sushko, L. Y. Fomychova, Y. S. Barsukov, Chastotypovtoryuvanosti bukv i bihram u vidkrytykh tekstakh ukrayins′koyumovoyu [Frequency of repetition of letters and digrams in open texts in Ukrainian]. Zakhyst informatsiyi [Protection of information] 12(3(48)) (2010).

[2] P. S. Dyurdeva, Authorship definition based on the frequency distribution of letter combinations, 2015. URL: https://www.math.spbu.ru/SD_AIS/documents/2015-12-441/2015-12-b-07.pdf.

[3] A. Likasa, N. Vlassisb, J. J. Verbeekb, The global k-means clustering algorithm, Pattern Recognition 36(2) (2003) 451–461. URL: https://www.cs.uoi.gr/~arly/papers/PR2003.pdf

[4] A. P. Reynolds G. Richards, Rayward-Smith V. J. The Application of K-medoids and PAM to the Clustering of Rules, Lecture Notes in Computer Science 3177 (2004). doi: 10.1007/978-3-540-28651-6_25.

[5] 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.

[6] 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.

[7] 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.

[8] K-Means Clustering. URL: https://people.revoledu.com/kardi/tutorial/kMean/#google_vignette.

[9] A. S. Romanov, Methodology and software package for identifying the author of an unknown text, 2010. URL: https://www.dissercat.com/content/metodika-i-programmnyi-kompleks-dlya-identifikatsii-avtora-neizvestnogo-teksta

[10] L. A. Borisov, Yu. N. Orlov, K. P. Osminin, Identification of the author of the text by the frequency distribution of letter combinations, Applied Informatics 26(2) (2013) 95–108.

[11] Yu. N. Orlov, K. P. Osminin, Determining the genre and author of a literary work by statistical methods, 2010. URL: https://keldysh.ru/papers/2013/prep2013_27.pdf.

[12] Yu. N. Orlov, K. P. Osminin, Methods of statistical analysis of literary texts, LIBROKOM, 2012.

[13] V. A. Pavlov, P. S. Dyurdeva, D. S. Shalymov, Clustering of Russian Manuscripts Based on the Feature Relationship Graph, Computer tools in education 1 (2016) 24–35.

[14] T. V. Batura. Formal methods for determining the authorship of texts, NGU Bulletin of Series Information technologies 10(4) (2012). URL: https://cyberleninka.ru/article/n/formalnye-metody-opredeleniya-avtorstva-tekstov

[15] M. Romanyshyn, Intro to Natural Language Processing. Grammarly, Inc., 2017.

[16] A. V. Babash, G. P. Shankin, Cryptography, SOLON-R, 2002. URL: https://pub.flowpaper.com/docs/https://book.edu-lib.net/books1/Babash_Kriprografiya_1.pdf, https://pub.flowpaper.com/docs/https://book.edu-lib.net/books1/Babash_Kriprografiya_2.pdf.

[17] A. P. Alferov, A. Yu. Zubov, A. S. Kuzmin, A. V. Cheryomushki, Fundamentals of cryptography, Helios, 2002. URL: https://studfile.net/preview/6311470/

[18] A. M. Yaglom, I. M. Yaglom, Probability and information. Science, ed. Phys.-Math. lit., 1973.

[19] R. G. Piotrovsky, Information measurements of language, Nauka, 1968.

[20] I. M. Yaglom, R. L. Dobrushin, A. M. Yaglom, Information theory and linguistics, Questions of linguistics 1 (1960) 100–110.

[21] D. S. Lebedev, V. A. Garmash, On the possibility of increasing the speed of transmission of telegraph messages, Telecommunications 1 (1958) 68-69.

[22] C. E. Shannon, Prediction and entropy of the printed English, 1951. URL: https://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf.

[23] O. V. Verbitskyy, Vstup do kryptolohiyi [Introduction to cryptology], Vydavnytstvo Naukovo-tekhnichnoyi literatury [Publishing House of Scientific and Technical Literature], Lviv, 1998.

[24] V. I. Perebyynis, M. P. Muravytska, N.P. Darchuk, Chastotni slovnyky ta yikhvykorystannya [Frequency dictionaries and their use], Naukova dumka [Scientific opinion], 1983.

[25] I. Khomytska, V. Teslyuk, A. Holovatyy, O. Morushko, Development of methods, models, and means for the author attribution of a text, Eastern-European Journal of Enterprise Technologies. 3(2(93)) (2018) 41–46. doi: 10.15587/1729-4061.2018.132052.

[26] 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.

[27] I. Khomytska, V. Teslyuk, N. Kryvinska, I. Bazylevych, Software-Based Approach Towards Automated Authorship Acknowledgement – Chi-Square Test on One Consonant Group, Electronics 9(7) (2020) 1138. doi: 10.3390/electronics9071138.

[28] I. Khomytska, V. Teslyuk, I. Bazylevych, I. Shylinska, Approach for Minimization of Phoneme Groups in Authorship Attribution Attribution, International Journal of Computing 19(1) (2020) 55–62. Doi: 10.47839/IJC.19.1.1693.

[29] D. Jurafsky, J. H. Martin, N-gram Language Models. URL: https://web.stanford.edu/~jurafsky/slp3/3.pdf.

[30] D. Jurafsky, J. H. Martin, Speech and Language Processing. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf.

[31] D. Jurafsky, J. H. Martin, Regular Expressions, Text Normalization, Edit Distance. URL: https://web.stanford.edu/~jurafsky/slp3/2.pdf.

[32] O. S. Goh, C. C. Fung, A. Depickere, Domain knowledge query conversation bots in instant messaging (IM), Knowledge-Based Systems 21(7). (2008) 681-691.

[33] S. Buk, Osnovy statystychnoy lingvistyky, LNU n. I. Franko Publishing House,2008.

[34] V. Vysotska, V. B. Fernandes, V. Lytvyn, M. Emmerich, M. Hrendus, Method for Determining Linguometric Coefficient Dynamics of Ukrainian Text Content Authorship, Advances in Intelligent Systems and Computing 871 (2018) 132–151. doi: 10.1007/978-3-030-01069-0_10.

[35] P. Kravets, The control agent with fuzzy logic, in: Proceedings of the International Conference on Perspective Technologies and Methods in MEMS Design, Lviv, Ukraine, 2010, pp. 40–41.

[36] P. Kravets, The Game Method for Orthonormal Systems Construction, in: Proceedings of the 2007 9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics, Lviv, Ukraine, 2007. doi: https://doi.org/10.1109/cadsm.2007.4297555.

[37] P. Kravets, Game Model of Dragonfly Animat Self-Learning, in: Proceedings of the International Conference on Perspective Technologies and Methodsin MEMS Design, Lviv, 2016, 195–201.

[38] 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.

[39] I. Balush, V. Vysotska, S. Albota, Recommendation System Development Based on Intelligent Search NLP and Machine Learning Methods, CEUR WorkshopProceedings 2917 (2021) 584–617.

[40] 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 the 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 1, Lviv, Ukraine, 2021, pp. 184–188. doi: 10.1109/CSIT52700.2021.9648764.

[41] N. Shakhovska, K. Shakhovska, The Method of Text Tonality Classification, in: Proceedings of the Computer Sciences and Information Technologies (CSIT), 1, Lviv, Ukraine, 2020, pp. 19–23.

[42] I. Khomytska, V. Teslyuk, L. Bordyuk, The Kolmogorov-Smirnov's Test for Authorship Attribution on the Phonological Level, in: Proceedings of the Computer Sciences and Information Technologies (CSIT), 1, 2020, pp. 259–262. doi: 10.1109/CSIT49958.2020.9322042.

[43] Y. Hlavcheva, V. Bobicev, O. Kanishcheva, Language-independent features for authorship attribution on Ukrainian texts, CEUR Workshop Proceedings Vol-2833 (2021) 134–143.

[44] O. Bisikalo, O. Boivan, N. Khairova, O. V. Kovtun, V. Kovtun, Precision Automated Phonetic Analysis of Speech Signals for Information Technology of Text-dependent Authentication of a Person by Voice, CEUR Workshop Proceedings Vol-2853 (2021) 276–288.

[45] N. Dubey, A. A. Verma, S. R. S. Iyengar, S. Setia, Implicit Visual Attention Feedback System for Wikipedia Users, in: Proceedings of the 17th International Symposium on Open Collaboration, Association for Computing Machinery, NY, USA, 2021, pp. 1–11. 10.1145/3479986.3479993.

[46] V. Lytvyn, V. Danylyk, M. Bublyk, L. Chyrun, V. Panasyuk, O. Korolenko, The lexical innovations identification in English-languagee eurointegration discourse for the goods analysis by comments in e-commerce resources, in Proceedings of the 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies, Lviv, 2021, 85–97.

[47] 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 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 22–25 September, 2021, 138–146. doi: 10.1109/CSIT52700.2021.9648682.

[48] 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: Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 22–25 September, 2021, 101–112.

[49] N. Hrytsiv, T. Shestakevych, J. Shyyka, Corpus Technologies in Translation Studies: Fiction as Document, CEUR Workshop Proceedings 2917 (2021) 327–343.

[50] D. Koshtura, V. Andrunyk, T. Shestakevych, Development of a Speech-to-Text Program for People with Haring Impairments, CEUR Workshop Proceedings 2917 (2021) 565–583.

[51] O. Prokipchuk, L. Chyrun, M. Bublyk, V. Panasyuk, V. Yakimtsov, R. Kovalchuk, Intelligent System for Checking the Authenticity of Goods Based on Blockchain Technology, CEUR Workshop Proceedings 2917(2021) 618–665.

[52] 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.

[53] L. Chyrun, A. Kowalska-Styczen, Y. Burov, A. Berko, A. Vasevych, I. Pelekh, Y. Ryshkovets, Heterogeneous data with agreed content aggregation system development, CEUR Workshop Proceedings 2386 (2019) 35–54.

[54] B. Rusyn, L. Pohreliuk, A. Rzheuskyi, R. Kubik, Y. Ryshkovets, L. Chyrun, S. Chyrun, A. Vysotskyi, V.B. Fernandes, The mobile application development based on online music library for socializing in the world of bard songs and scouts' bonfires, Advances in Intelligent Systems and Computing 1080 (2020) 734–756. doi: 10.1007/978-3-030-33695-0_49.

[55] L. Chyrun, A. Gozhyj, I. Yevseyeva, D. Dosyn, V. Tyhonov, M. Zakharchuk, Web Content Monitoring System Development, CEUR Workshop Proceedings 2362 (2019) 126–142.

[56] N. Antonyuk, L. Chyrun, V. Andrunyk, A. Vasevych, S. Chyrun, A. Gozhyj, I. Kalinina, Y. Borzov, Medical news aggregation and ranking of taking into account the user needs, CEUR Workshop Proceedings 2488 (2019) 369–382.

[57] N. Antonyuk, M. Medykovskyy, L. Chyrun, M. Dverii, O. Oborska, M. Krylyshyn, A. Vysotsky, N. Tsiura, O. Naum, Online Tourism System Development for Searching and Planning Trips with User's Requirements, Advances in Intelligent Systems and Computing 1080 (2020) 831-863.

[58] V. Andrunyk, A. Vasevych, L. Chyrun, N. Chernovol, N. Antonyuk, A. Gozhyj, V. Gozhyj, I. Kalinina, M. Korobchynskyi, Development of information system for aggregation and ranking of news taking into account the user needs, CEUR Workshop Proceedings 2604 (2020) 1127–1171.

[59] 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.

[60] 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 Data Stream Mining and Processing, 2020, pp. 420–425. doi: 10.1109/DSMP47368.2020.9204160.

[61] 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 Data Stream Mining and Processing, 2020, pp. 432–437. 10.1109/DSMP47368.2020.9204321.

[62] 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 Data Stream Mining and Processing, 2020, pp. 414–419. doi: 10.1109/DSMP47368.2020.9204101.

[63] V. Vysotska, Victana Web-resource, 2022. URL: https://victana.lviv.ua/nlp/n-grams.