Cluster Analysis of Exclamations and Comments on E-Commerce Products

Oleh Veres[1], Yurii Matseliukh[1], Taras Batiuk[1], Sofiia Teslia[1], Alyona Shakhno[2], Tetiana Kopach[1], Yeva Romanova[1] and Inesa Pihulechko[1]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Kryvyi Rih National University, V. Matusevych Street, 11, Kryvyi Rih, 50027, Ukraine*

Abstract

A survey of consumers' opinions of women's clothing was obtained from reviews and comments during online sales. The high popularity of clothing and footwear as a segment of the electronic market is considered. Correlation analysis of survey data was performed, correlation coefficients were calculated, a correlation matrix was constructed, and autocorrelation was established, establishing how consumers perceive the offered products and services in the clothing sales segment. Cluster data analysis was performed. Dendrograms of clothing sales responses were constructed and analyzed due to the conclusions obtained from various research methods of the clothing sales segment on the Internet, recommendations for improving the clothing sales system, and proposals for developing new marketing measures.

Keywords

Cluster analysis, information technologies, business analysis, e-commerce products, exclamations, comments, data processing

1. Introduction

The problem of analyzing the opinion of women's clothing consumers, obtained from customer reviews and comments during online sales, has the high share in the world (more than half of all online sales), which belongs to the market for clothing and footwear. Solving this problem, according to the authors [1-6], will allow companies to develop a system of marketing activities to attract new customers who use the Internet to buy women's clothing to stimulate demand for goods and services. The unresolved problem remains the confirmation of the connection between positive feedback and company profits and business expansion.

As recommended by the authors [7-12], we can understand customers' moods and preferences with such data, which is paramount for marketing and business development in general. Such data analysis [13-17] will give companies an idea of how customers perceive their products and services and how to improve their offerings.

By analyzing this data, according to the authors [16, 18-23], business analysts and business owners will understand the relationship of different variables in customer feedback on clothing. Namely, it will be possible to track:

- which reviews predominate (positive/negative)
- what feelings arise in buyers
- what things do different age groups buy

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: Oleh.M.Veres@lpnu.ua (O. Veres); indeed.post@gmail.com (Y. Matseliukh); taras.batiuk.mnsa.2020@lpnu.ua (T. Batiuk); sofia.tesla.sa.2019@lpnu.ua (S. Teslia); ashakhno@knu.edu.ua (A. Shakhno); tetiana.m.kopach@lpnu.ua (T. Kopach); yeva.romanova.sa.2019@lpnu.ua (Y. Romanova); inessa.pihulechko.sa.2019@ lpnu.ua (I. Pihulechko)
ORCID: 0000-0001-9149-4752 (O. Veres); 0000-0002-1721-7703 (Y. Matseliukh); 0000-0001-5797-594X (T. Batiuk); 0000-0002-2591-2431 (S. Teslia); 0000-0003-0718-0051 (A. Shakhno); 0000-0002-1293-229X (T. Kopach); 0000-0003-0522-0806 (Y. Romanova); 0000-0003-2789-2902 (I. Pihulechko)

For example, if a customer comes with a request to update a website or create a new one, business analysts can [24-27]:

- to compile the initial User Persona, namely the target audience of the store
- to see which positions are most often bought
- to give tasks to the team so that it can develop the system so that it promotes these positions in the section "recommended" products
- make sure that the reviews that users find helpful are shown first
- show sections with clothes, guided by the section that has the most reviews of the product is the most popular
- things with negative reviews show last

At the end of the work, business analysts can improve user experience and make the business more profitable.

The aim work is following:

- To use visualization methods for graphical display
- To use primary statistical processing for numerical data on the feedback and comments of buyers of women's clothing during online sales;
- To analyze trends of the studied indicators;
- To conduct both data correlation analysis and cluster analysis;
- Building a dendrogram of feedback on clothing sales;

We will use the results to improve the website's user experience and thus increase the profitability of the online clothing store.

2. Literature review

The topic we have chosen is quite popular at the moment. We tried to choose the information sources that have been published in the last two years so that the information we received is not outdated. Researching information about it, we found several articles [6 - 19] and various sources [21 - 28] that helped us understand the relevance of the topic - Women's e-commerce clothing reviews [29, 30] and whether we can bring something new to its development.

In the article [18] we reviewed, the author analyzed a set of e-commerce women's clothing that contains numerical data and text reviews written by customers. The author [18] used a large number of methods of data analysis and visualization. He also used many graphs, tables, and charts. All this simplifies the understanding of the work, which may help in the future in our research.

The following article [19] we chose turned out to be very useful. The author of the work [19] justified the actions taken and made beneficial reviews that help to make, which reviews dominate in each department, customers of which age leave the most reviews, such as clothing aesthetics, position, and the quality of the material affects the rating and what you need to pick up to avoid problems. Companies can focus on what works and what doesn't. Knowing the demographics of reviewers, you can make marketing decisions [19, 20] (for example, advertising on the Internet on the sites most visited by people of a certain age).

In the following article [21], the authors discuss the importance of mood analysis and how it can be used to understand customer choices. The authors [22] tries to find out the age of groups of customers who are satisfied with buying a particular thing online. The authors [21, 23] first tries to analyze non-textual review functions, such as age, class of items purchased, etc. and then finds a relationship between them and the recommended product. They try to determine whether the review text recommends the purchased product or not.

In the following article [24], the author used five popular machine learning algorithms to solve the problem, including logistic regression, vector support machine (SVM), Random Forest, XGBoost, and LightGBM. Based on natural language processing (NLP), these algorithms elucidate the relationship between review functions and product recommendations based on natural language processing (NLP) [24]. The authors achieved the best result with the LightGBM algorithm with the highest AUC value and accuracy. Thus, authors [24-28] helped us determine which algorithm is the most effective, thus bringing something new to our study.

So, we can summarize the advantages and disadvantages of our chosen topic.

Benefits are following:

- Enough relevant information that helps us make our work better.
- The volumes of our dataset make it possible to analyze data by various methods and algorithms.
- This topic is relevant today, which is very important for research.
- Ability to process large amounts of information from the business analysis.
- Disadvantages:
- The difficulty of choosing the method that will give high accuracy to the study.

So, we can conclude that our topic is still relevant and needs new research in this area. It will help companies better understand customer preferences and how they should go. We have processed research information from four authors, and we can say that we can explore this topic more deeply and bring something new to the field.

3. Methods

According to paper [31-34], it is advisable to reduce the number of single records for statistical and analytical data processing by combining them into clusters with a similar set of properties. Designing this process does not make sense without initially establishing a basis for analysis. Namely, the researcher may be interested in the analysis regarding the reviewer's age, product ID, etc.

We focused on the hierarchical agglomerative cluster analysis of multidimensional data to systematize this analysis [35-37].

Our problem has no time sequence, which requires moving average, weighted moving average, median filtering, and normalization of time sequences. However, we are dealing with a multidimensional dataset [30, 31] that combines important data (Age, Clothing ID, Class Name, Rating) and less critical data for analysis (Title and Review Text, etc.). This division comes to mind due to the availability of systematic data set analysis. The presence of such a significant dimension, in our opinion, is due to the principle of maximum use of the work of the reviewer, who agreed to give an assessment. However, "garbage" data for analysis are not attributes of the subject area. For example, the Review Text reflects the verbal arsenal and temperament of the reviewer, which are highly subjective. Therefore, if it is necessary to study not statistics on the product but people who have agreed to be reviewers, it is necessary first to analyze all their textual reviews and, to a lesser extent, conclude the person by his preferences in clothing. A modest person will not look for a biker coat and write a review but rather choose a strict dress or coat.

Thus, the essence of our chosen method is to divide the data into relatively homogeneous groups - clusters, by determining the criteria for the acceptance of attributes [38-41]. Of course, it is necessary to determine the depth of sampling of the data set of exclamations, posts and comments [42-46], for example, for e-commerce products [47-69], i.e., to determine the number of clusters to which it is necessary to sort the records.

Regardless of the subject of the study, cluster analysis includes:

- Selection for clustering, data presentation in the table "object property."
- Rationing of table data.
- Reasonable choice of metric for the formation of the proximity matrix.
- Construction of a matrix near-bone based on a normalized table "object property."
- Aggregation strategy for cluster analysis procedure.
- Cluster analysis according to the procedure on the proximity matrix.
- Dendrogram, as a result of research and selection of the necessary clusters.

Cluster analysis, of course, has some limitations and shortcomings. Still, its advantages are crucial for our case analysis because we need to process a significantly large amount of sample data, which provides higher accuracy than in the case of small samples. It is also worth noting the "ubiquity" of this method for any set of parameters.

4. Experiments

Our data consist of consumers' opinions of women's clothing obtained from reviews and comments during online sales. We have a massive array of data (23,486 records) with a ten-dimensional attribute size as a dataset. Attributes in the study of consumer feedback included:

- Clothing ID: serial number of the item
- Age: age to find out the age category.
- Title: review title.
- Review Text: review text.
- Rating: product reviews from 1 to 5.
- Recommended IND: value 0 if not recommended by the user, one is recommended.
- Positive Feedback Count: The number of users who found the review helpful.
- Division Name: name of the department (intimates / general).
- Department Name: clothing category (top / bottom).
- Class Name: the name of the item (pants/blouse).

Table 1 presents only part of the 23,486 records of women's clothing reviews and comments during online sales.

Table 1
The data set structure of women's clothing reviews and comments during online sales

Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
767	33	-	Absolutely …	4	1	0	Initmates	Intimate	Intimates
1080	34	-	Love this dress! …	5	1	4	General	Dresses	Dresses
1077	60	Some major …	I had such high hopes …	3	0	0	General	Dresses	Dresses
1049	50	My favorite …	I love, love, love this …	5	1	0	General Petite	Bottoms	Pants
847	47	Flattering shirt	This shirt is very …	5	1	6	General	Tops	Blouses
1080	49	Not for the …	I love tracy reese …	2	0	4	General	Dresses	Dresses
1077	53	Dress looks ….	Dress runs small …	3	0	14	General	Dresses	Dresses

The next step was to generate a report table with the minimum number of empty cells. Then we plotted data graphs in Cartesian (Fig.1) and polar coordinate systems (Fig.2). We determined the descriptive statistics of quantitative dataset characteristics (Table 2).

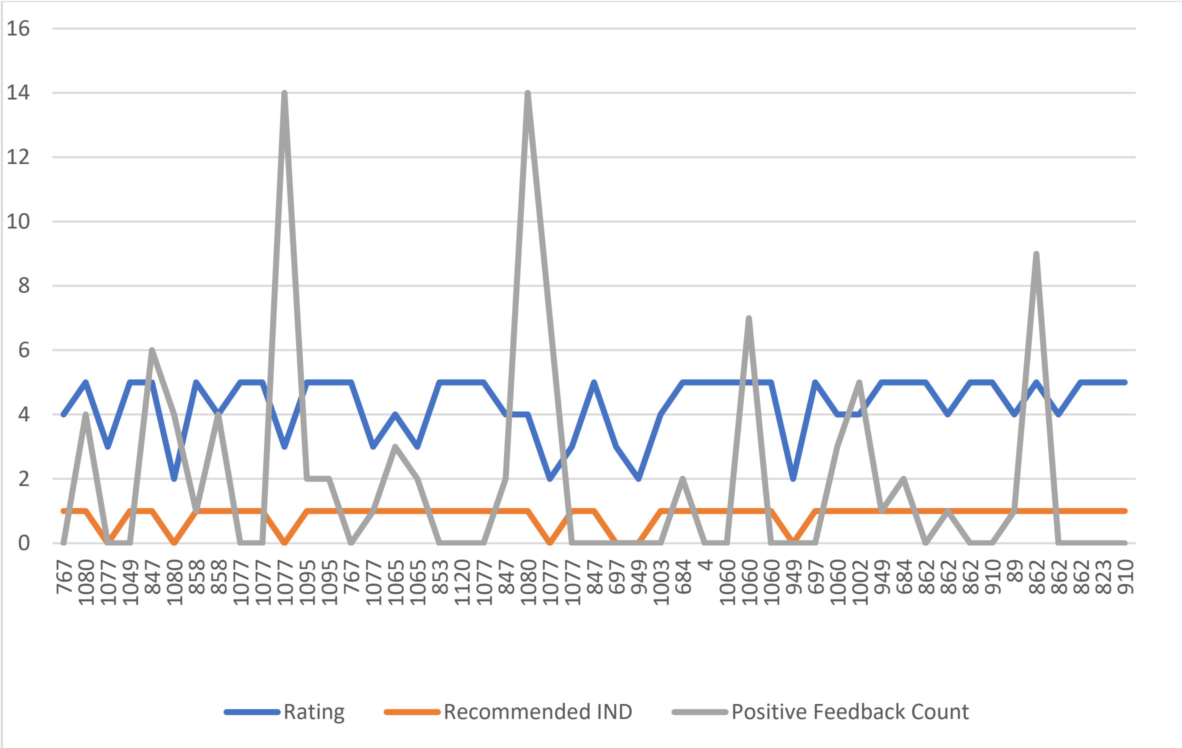


Figure 1: The data graphs in Cartesian coordinate systems

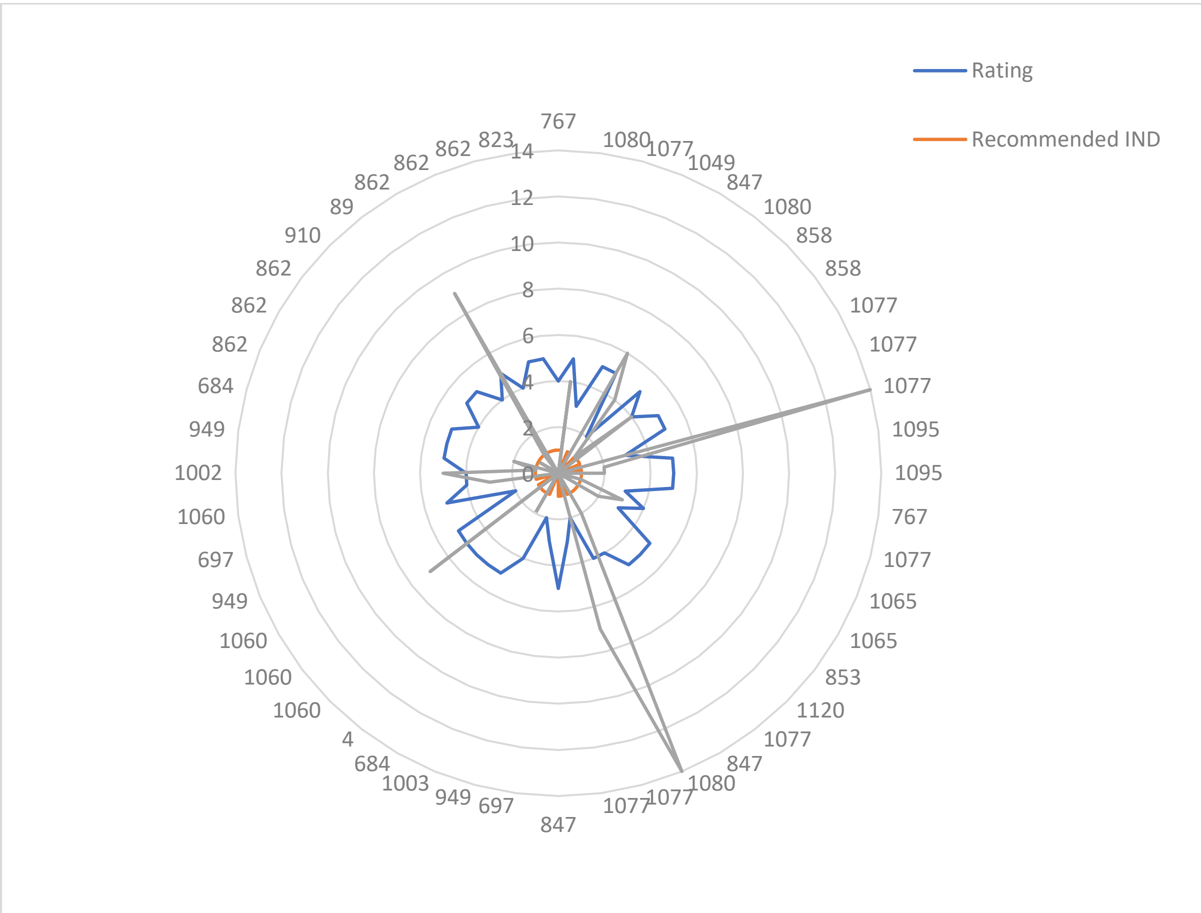


Figure 2: The data graphs in the polar coordinate system

Table 2
Results of descriptive statistics

Indexes	Age	Rating	Recommended IND	Positive Feedback Count
Average	43.1985438	4.1960317	0.822362258	2.535936302
Standard error	0.08012678	0.0072432	0.002494043	0.037208146
Median	41	5	1	1
Moda	39	5	1	0
Standard deviation	12.2795436	1.1100307	0.382215639	5.702201502
Sampling variance	150.787191	1.2321682	0.146088795	32.51510197
Kurtosis	-0.11182071	0.8041359	0.845878968	71.69317868
Asymmetry	0.52561451	-1.313529	-1.686951968	6.472997729
Interval	81	4	1	122
Minimum	18	1	0	0
Maximum	99	5	1	122
Sum	1014561	98548	19314	59559
Amount	23486	23486	23486	23486
Reliability level (95.0%)	0.1570537	0.0141971	0.004888486	0.072930385

We submit data on age in a histogram (Fig.3). The histogram's cumulative age data of women's clothing e-commerce review is shown in Fig.4.

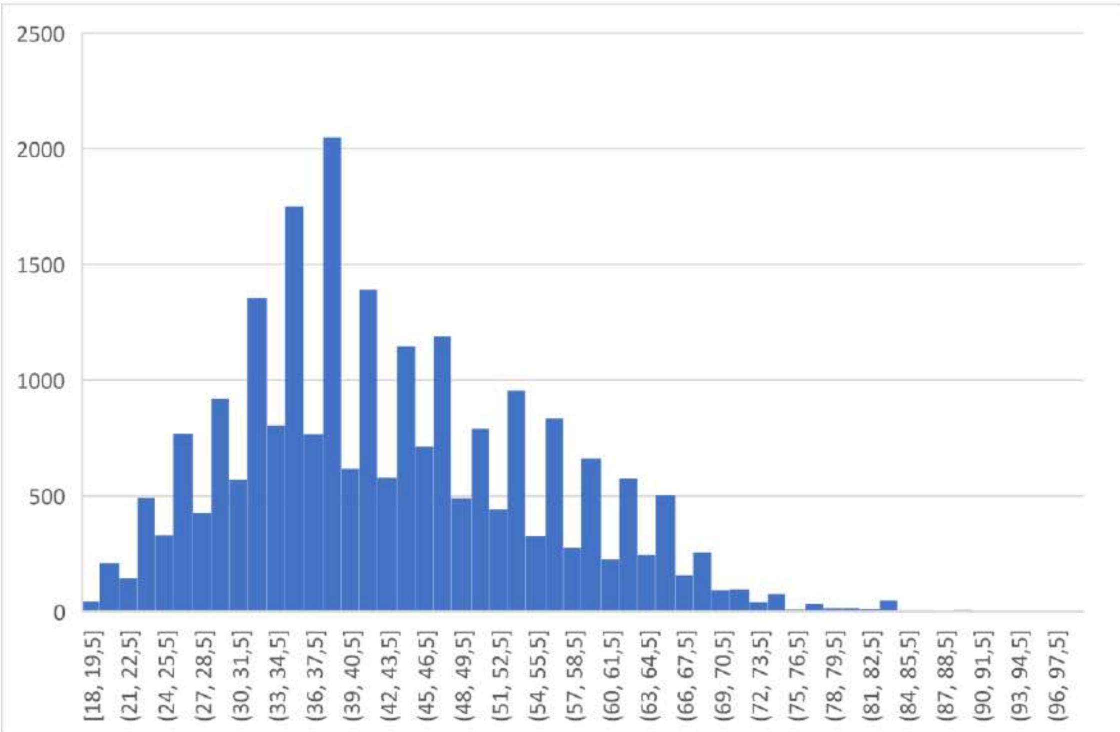


Figure 3: The histogram of age data of women's clothing e-commerce review

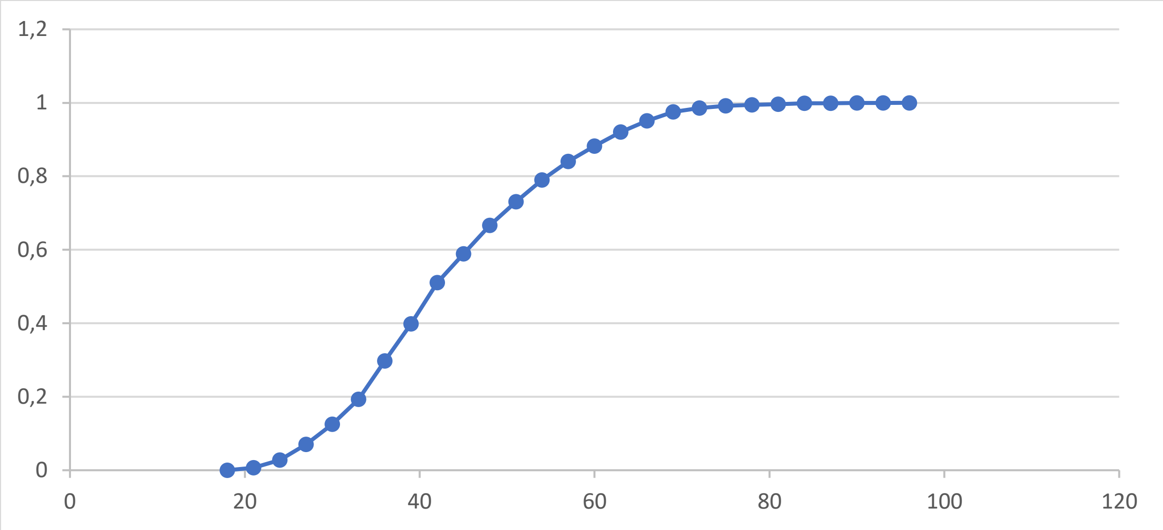


Figure 4: The cumulative age data of women's clothing e-commerce review by the histogram

5. Discussion

5.1. Smoothing time series

To achieve this goal, namely: acquaintance with the main methods of highlighting the trend of the studied indicator, which is represented by the nature of its trend, using methods of smoothing time series and presenting the results using an MS Excel spreadsheet, we conduct such research. We opened a new Excel workbook and entered our data on the new worksheet. We completed each task on one worksheet.

Smoothing according to Kandel formulas - simple moving average.

We smooth the data using the size of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 (Fig.5-Fig.7). We have to get seven columns in a row. Then we smooth the data using the smoothing interval w = 3, then smooth the obtained smoothed data again, but use the size of the smoothing interval w = 5. Continue smoothing the obtained data with a smoothing interval of w = 7 and w = 15. We must get seven in a row-column (Fig.8-Fig.9).

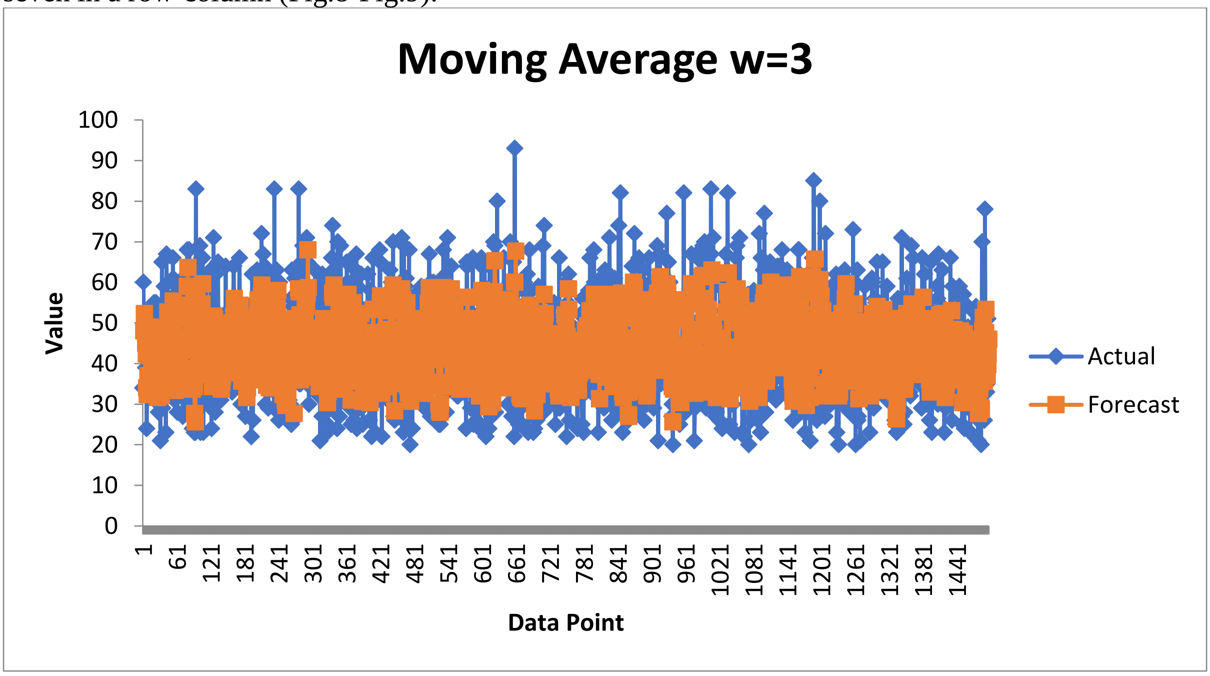


Figure 5: Moving average method for women's clothing e-commerce review at w = 3

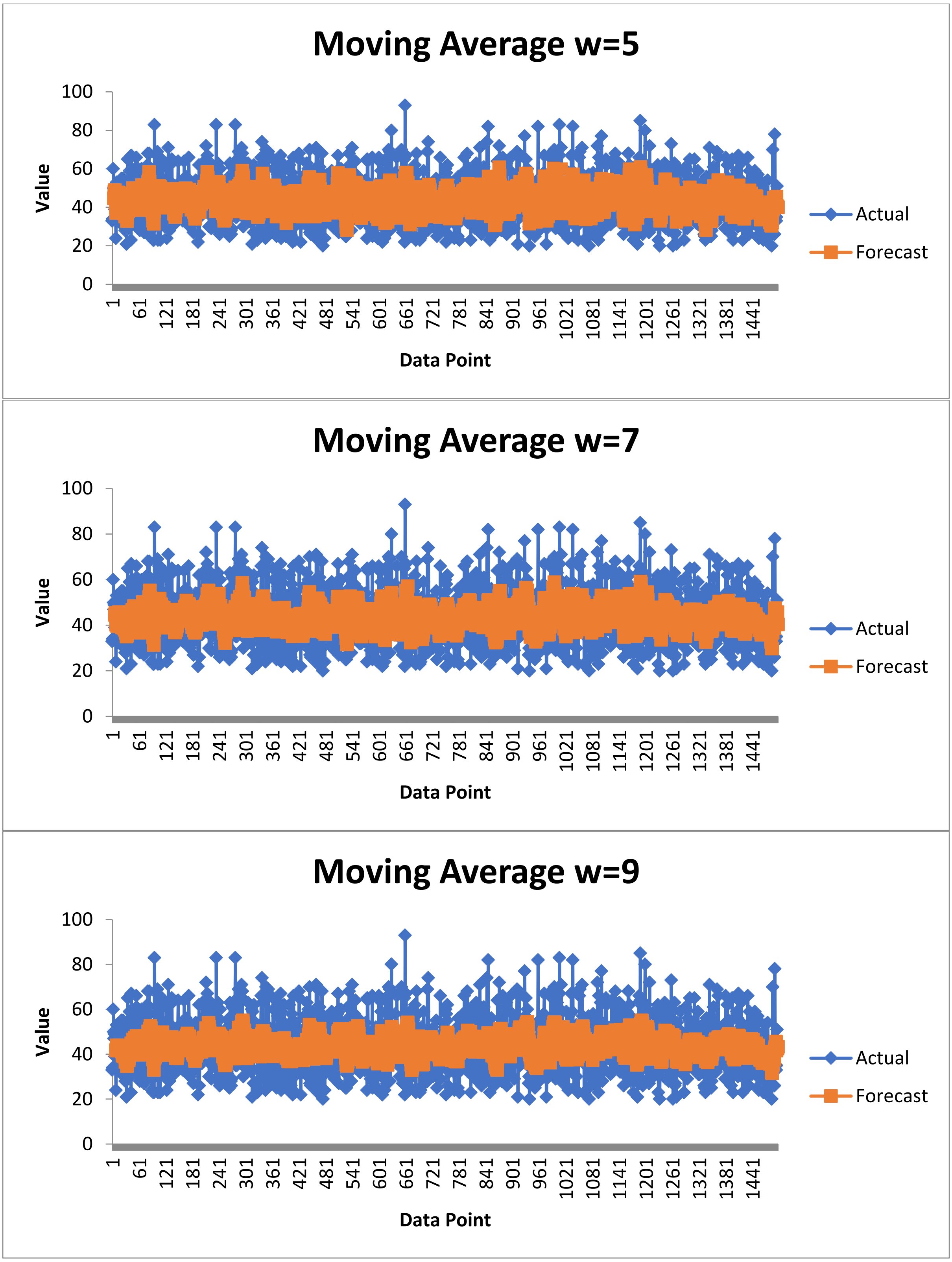


Figure 6: Moving average method for women's clothing e-commerce review at w = 5-9

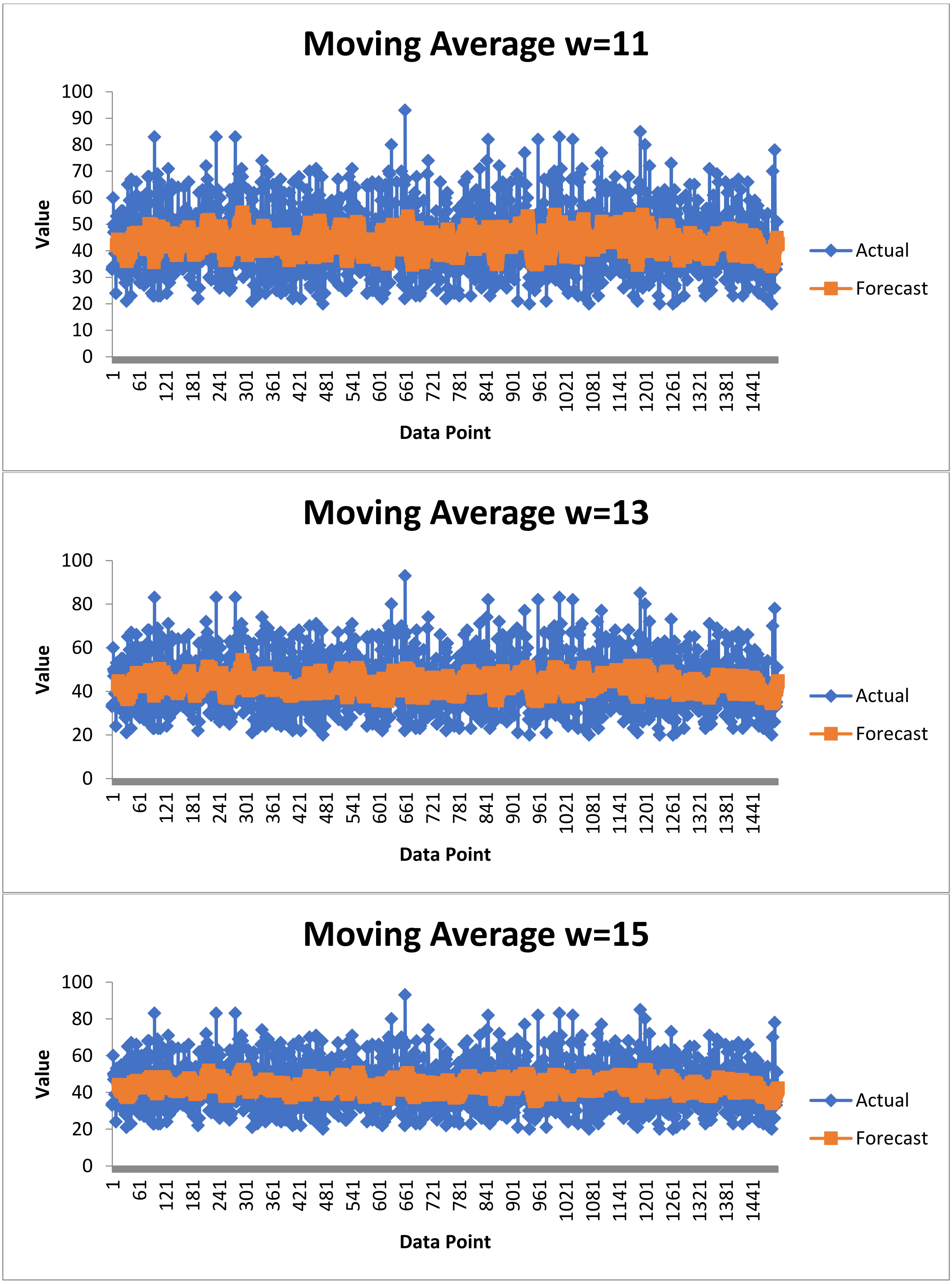


Figure 7: Moving average method for women's clothing e-commerce review at w =11- 15

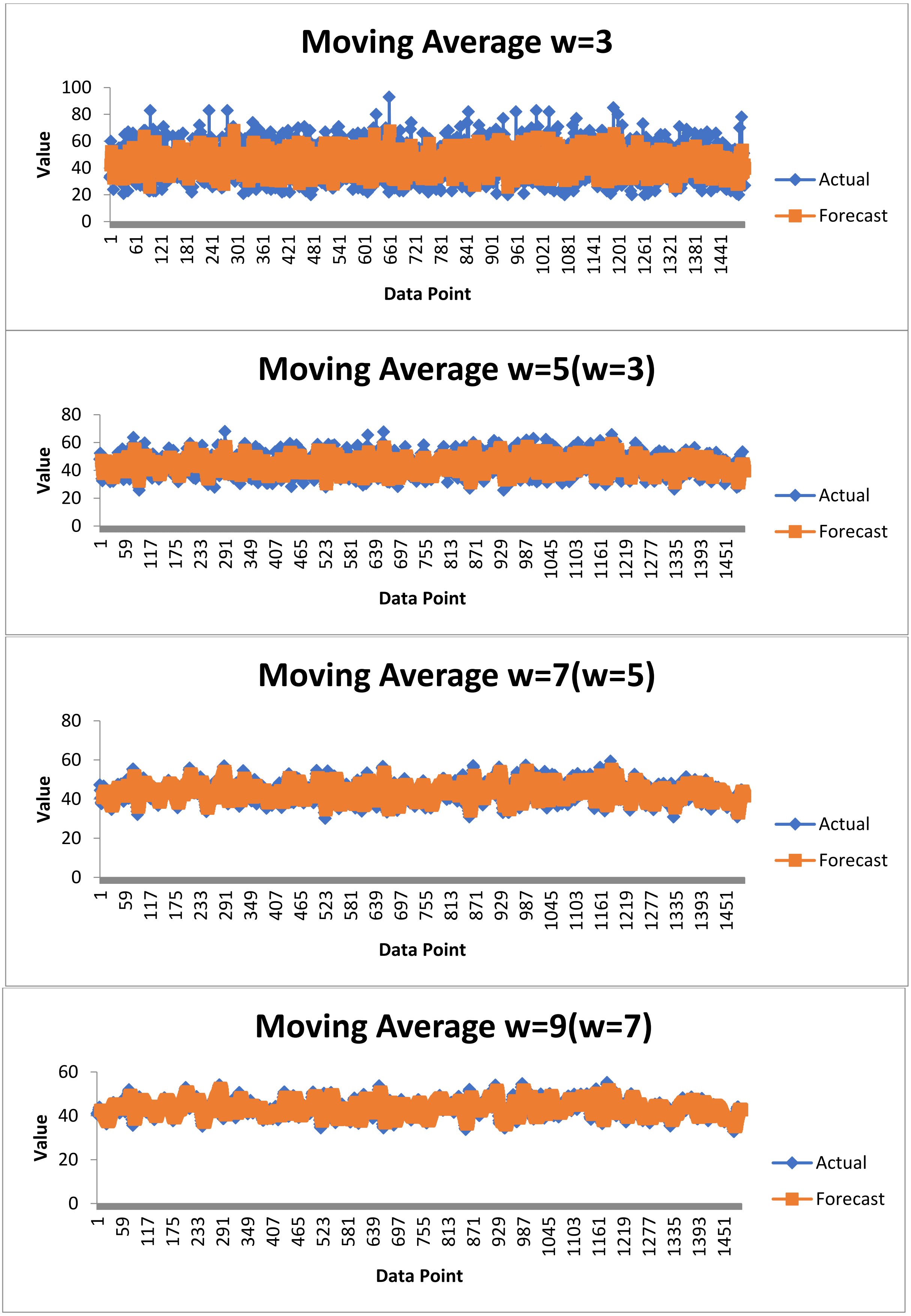
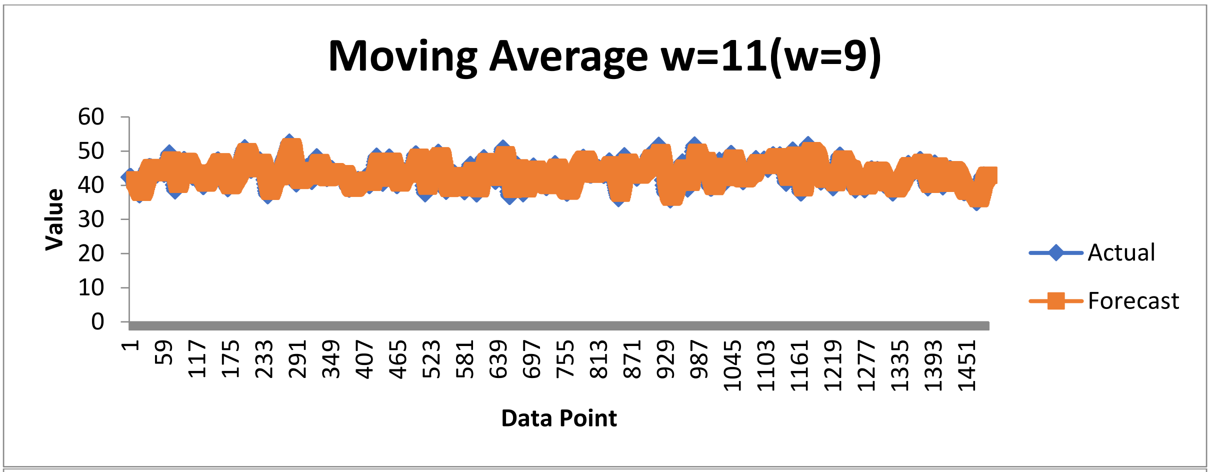
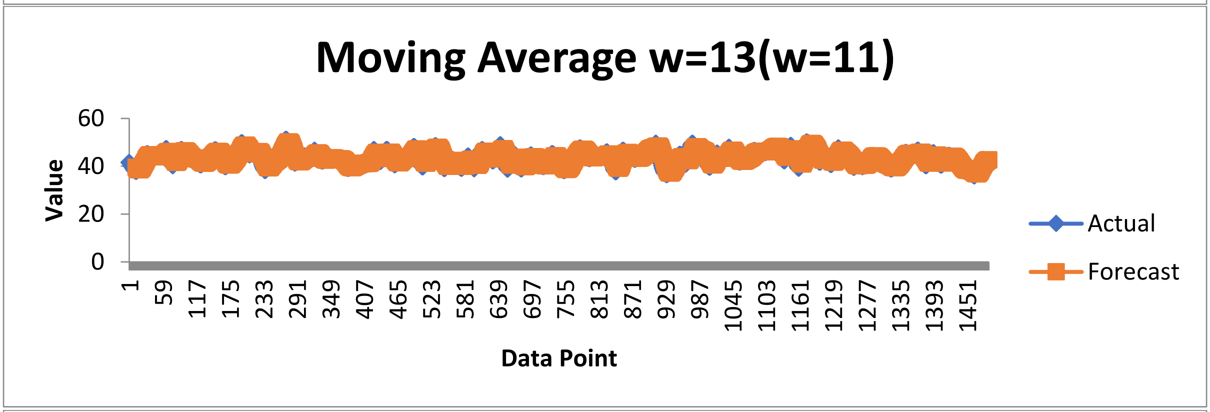


Figure 8: Moving average method for women's clothing e-commerce review at w = 3, w=5 (3), w=7 (5), w=9 (7)





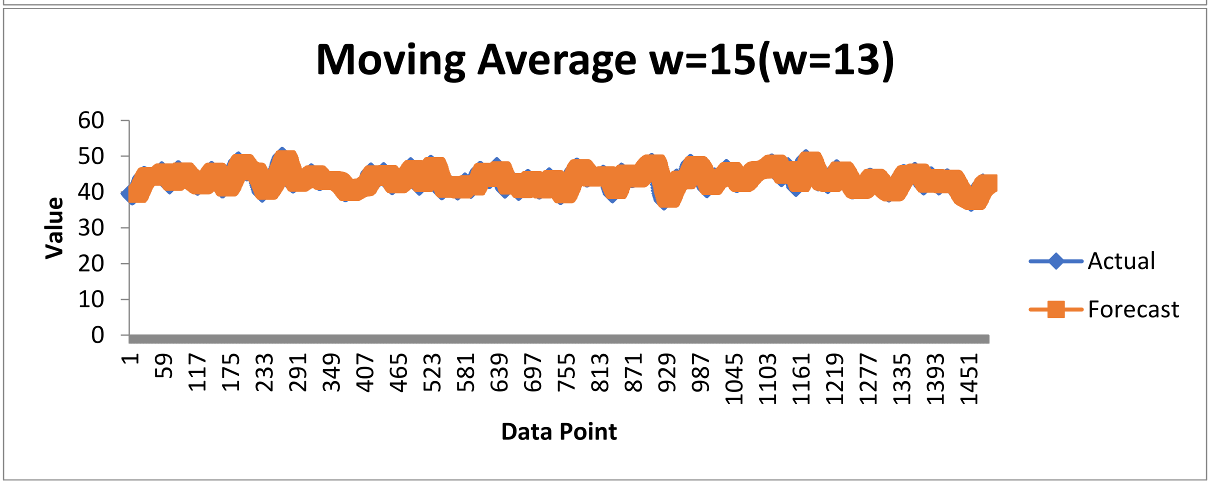


Figure 9: Moving average method for women's clothing e-commerce review at w=11 (9), w=13 (11), w=15 (13)

5.2. Smoothing according to formulas from Pollard

Depending on the size of the smoothing interval, the weight for the mid-level varies. Smoothing is carried out in the same way as in the previous paragraph. Smooth the data using the size of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15 (Fig.10-Fig.11). We have to get seven columns in a row.

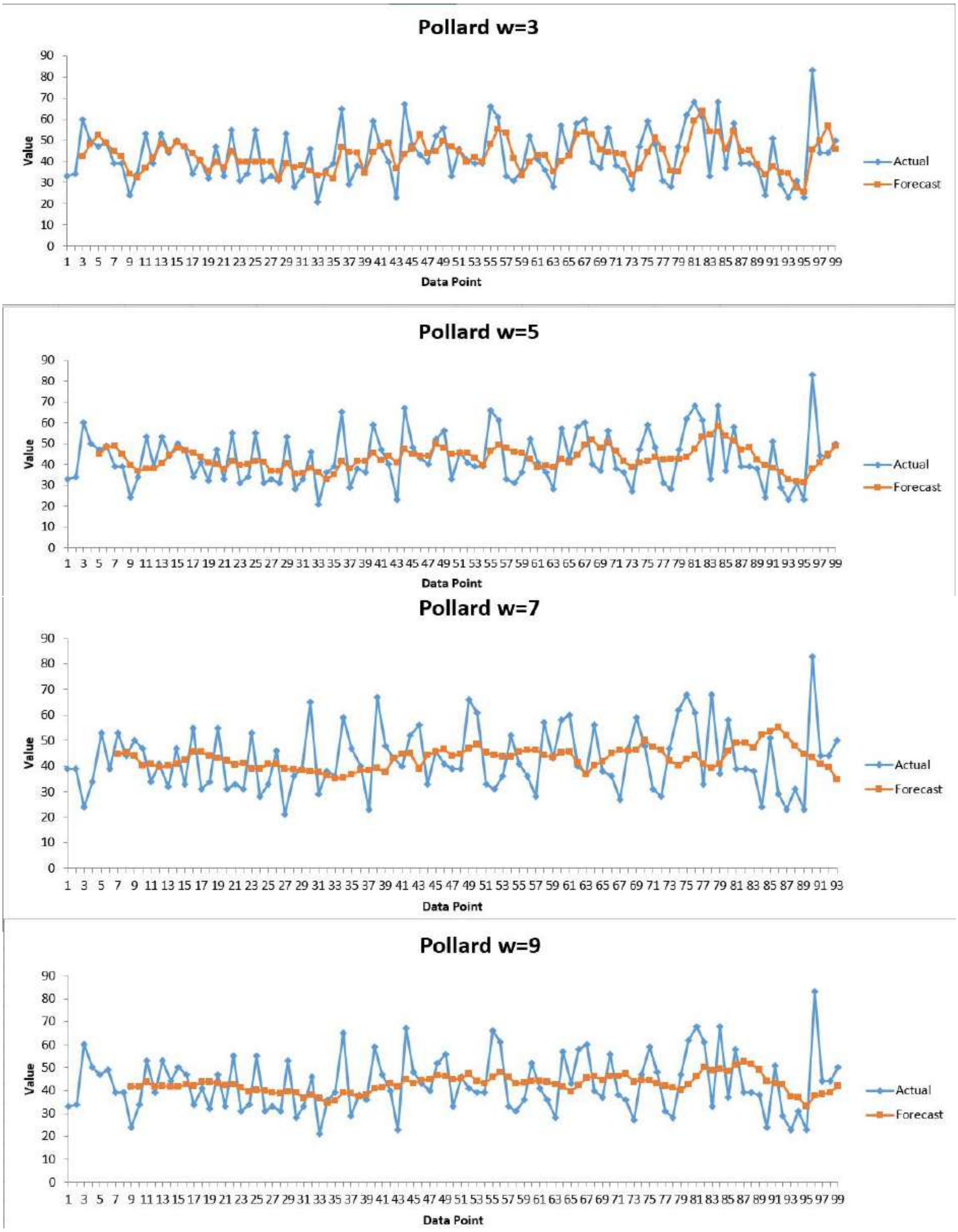


Figure 10: Pollard smoothing graph by formulas w = 3-9

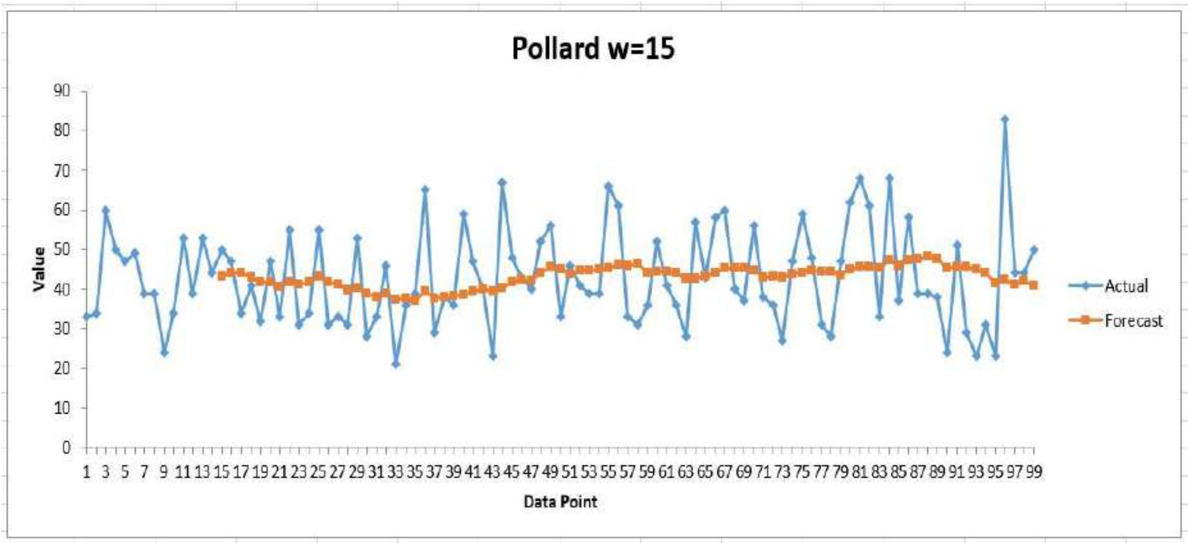


Figure 11: Pollard smoothing graph by formulas w = 15

We smooth the data using the smoothing interval w = 3, then smooth the obtained smoothed data again, but use the size of the smoothing interval w = 5. Continue smoothing the obtained data with a smoothing interval of w = 7 and w = 15. We must get seven in a row-column.

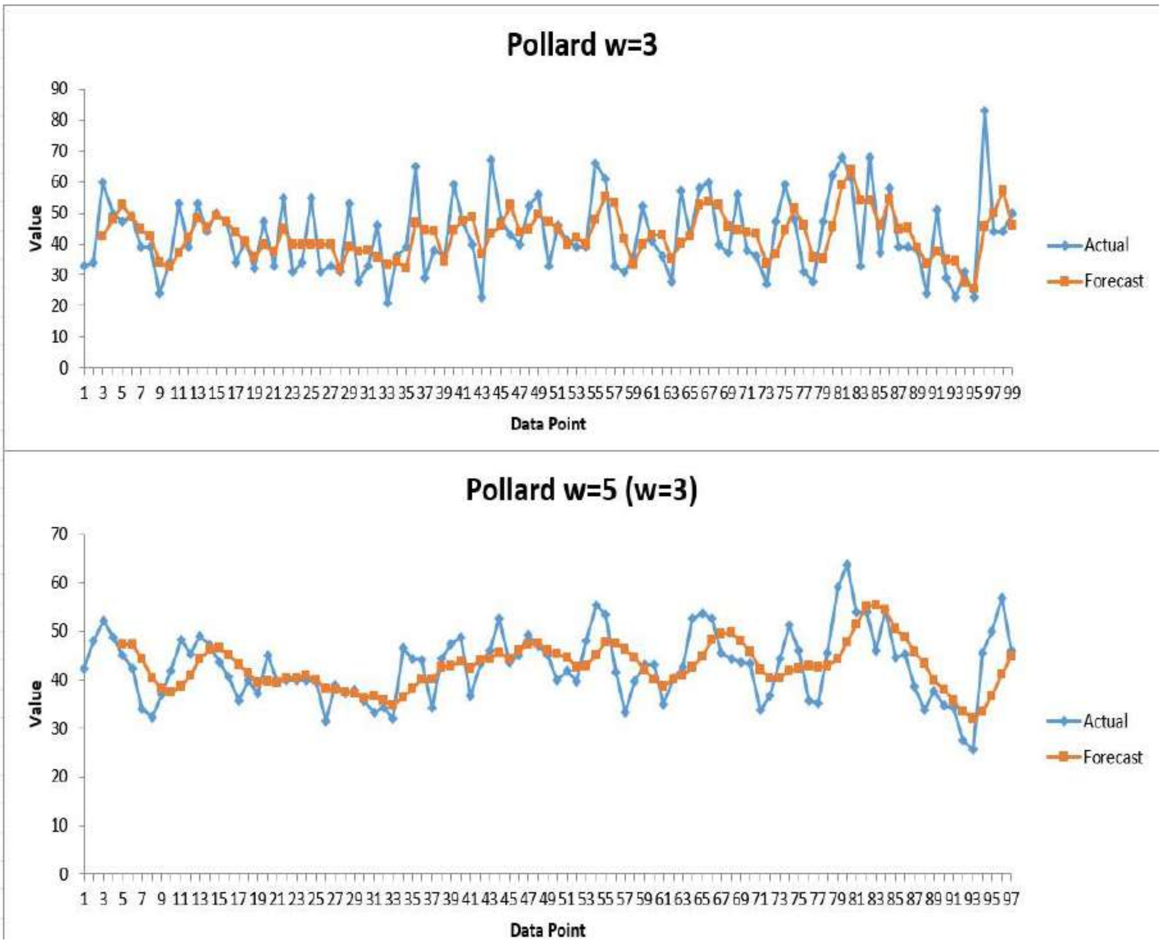


Figure 12: Pollard smoothing graph by formulas w = 3, w = 5 (w =3).

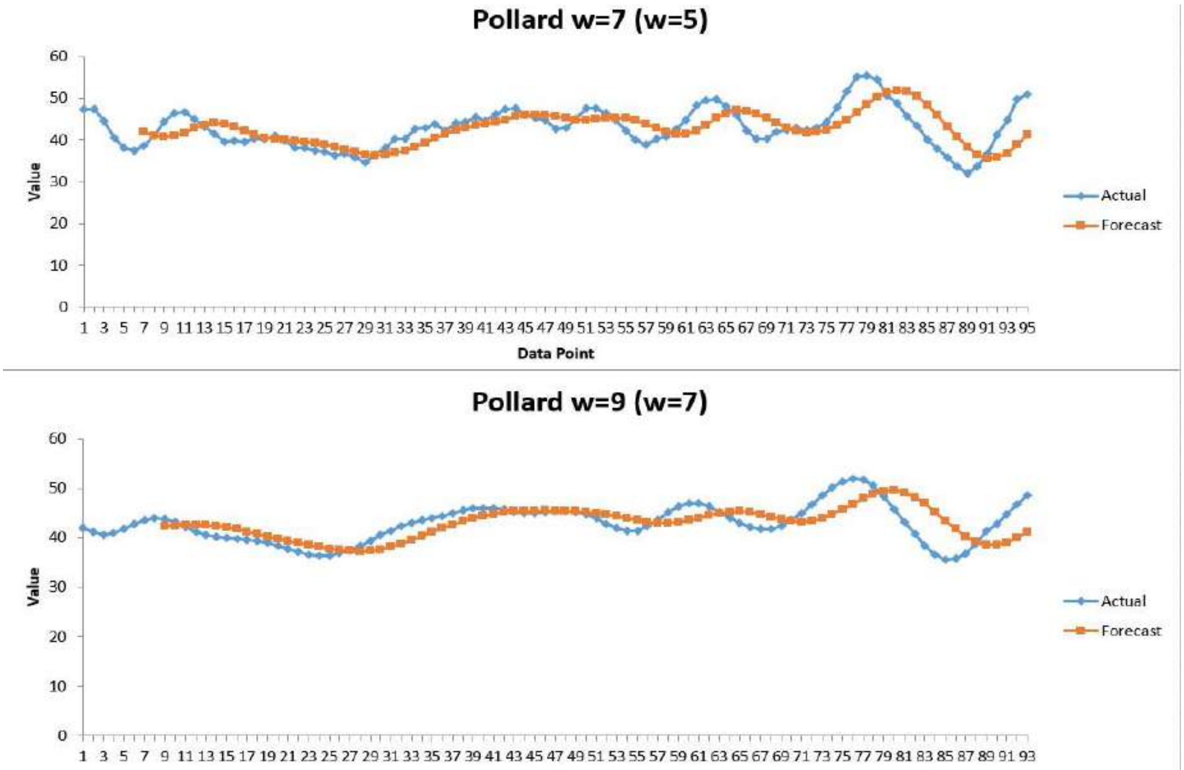


Figure 13: Pollard smoothing graph by formulas w = 7 (w = 5), w = 9 (w = 7)

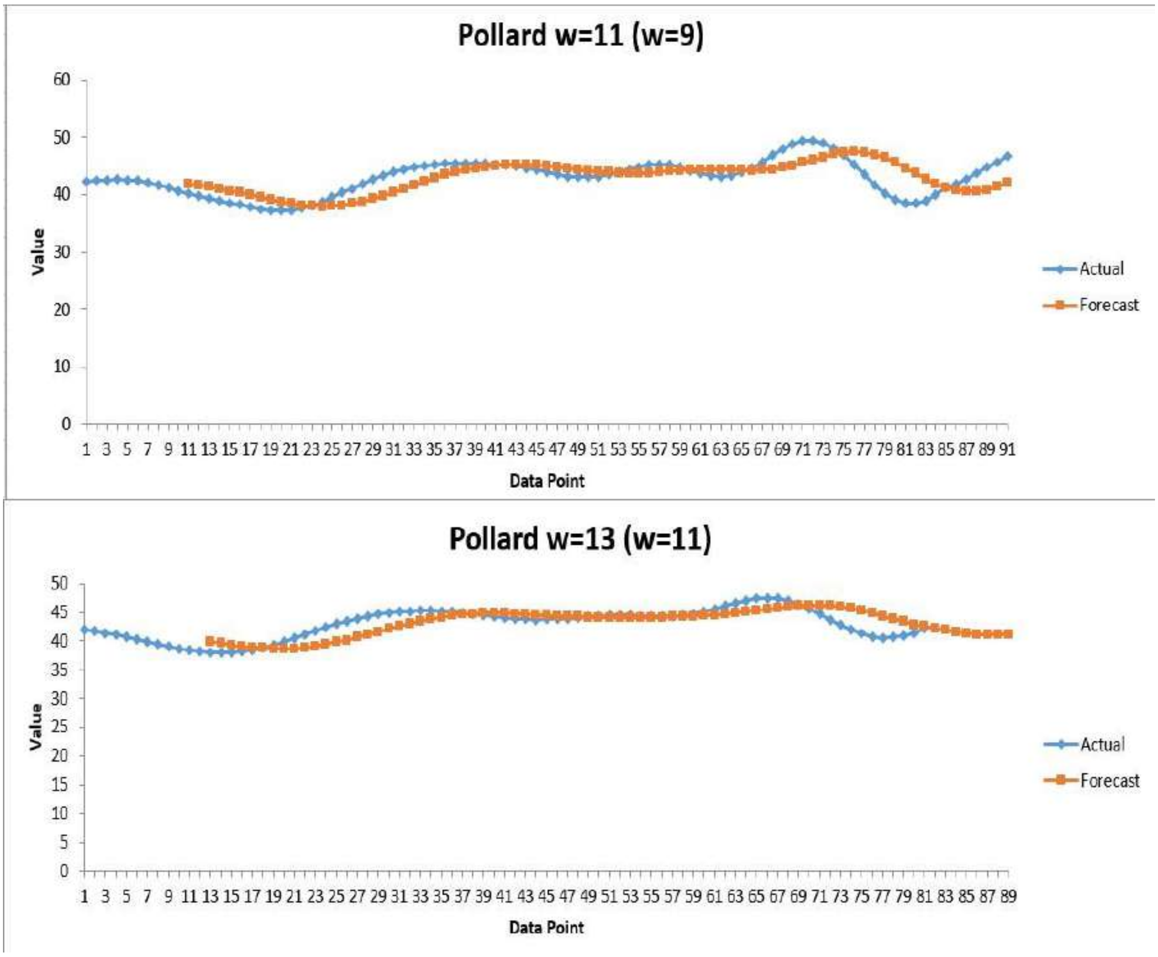


Fig. 14. Pollard smoothing graph by formulas w = 11 (w = 9), w = 13 (w = 11)

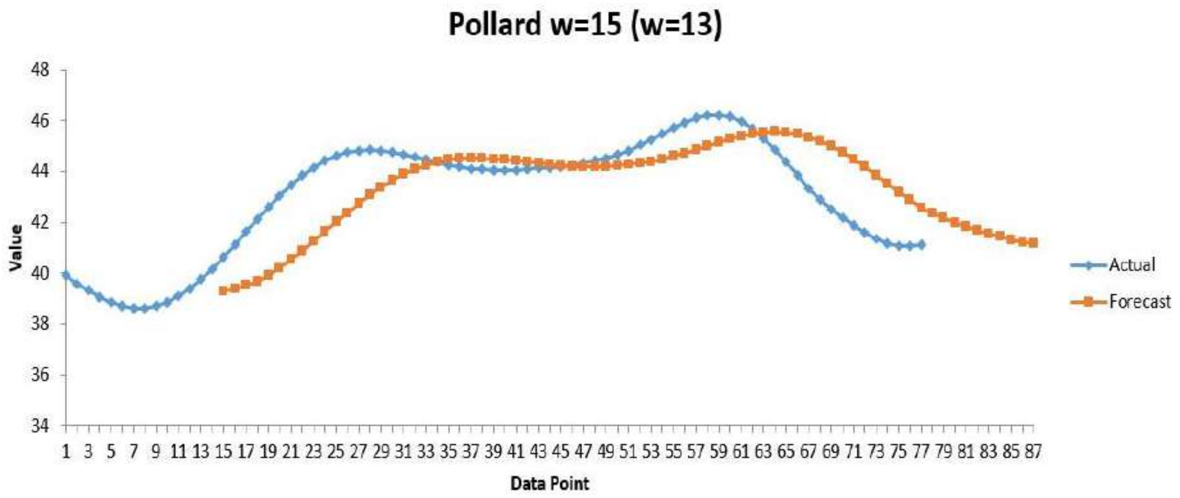


Figure 15: Moving average method for women's clothing e-commerce review at w = 15(w=13)

5.3. 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 smoothing the same series with the parameter values α= 0.1, 0.15, 0.2, 0.25, 0.3. To find the number of turning points and correlation coefficients between the original values and smoothed in all these cases (Fig.16 - Fig.17).

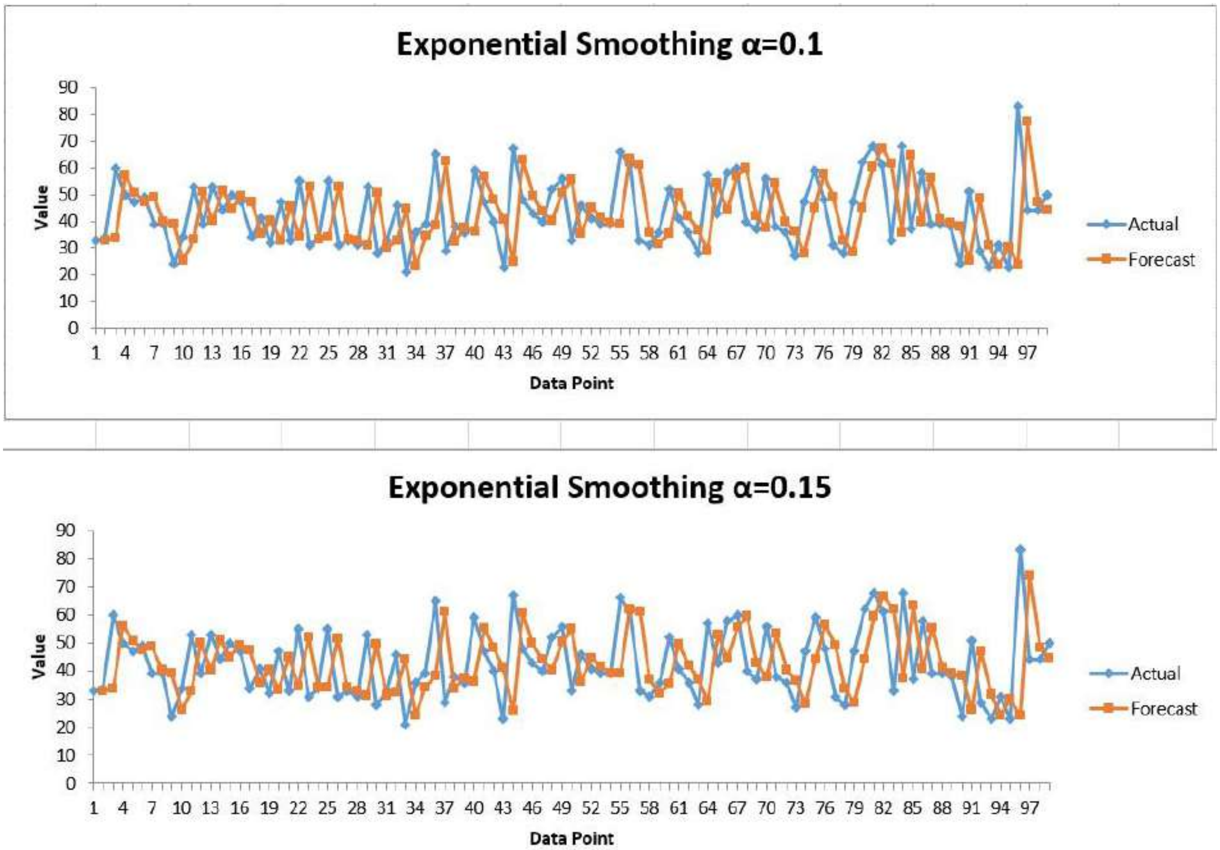


Figure 16: Graphs of exponential smoothing at α =0.15

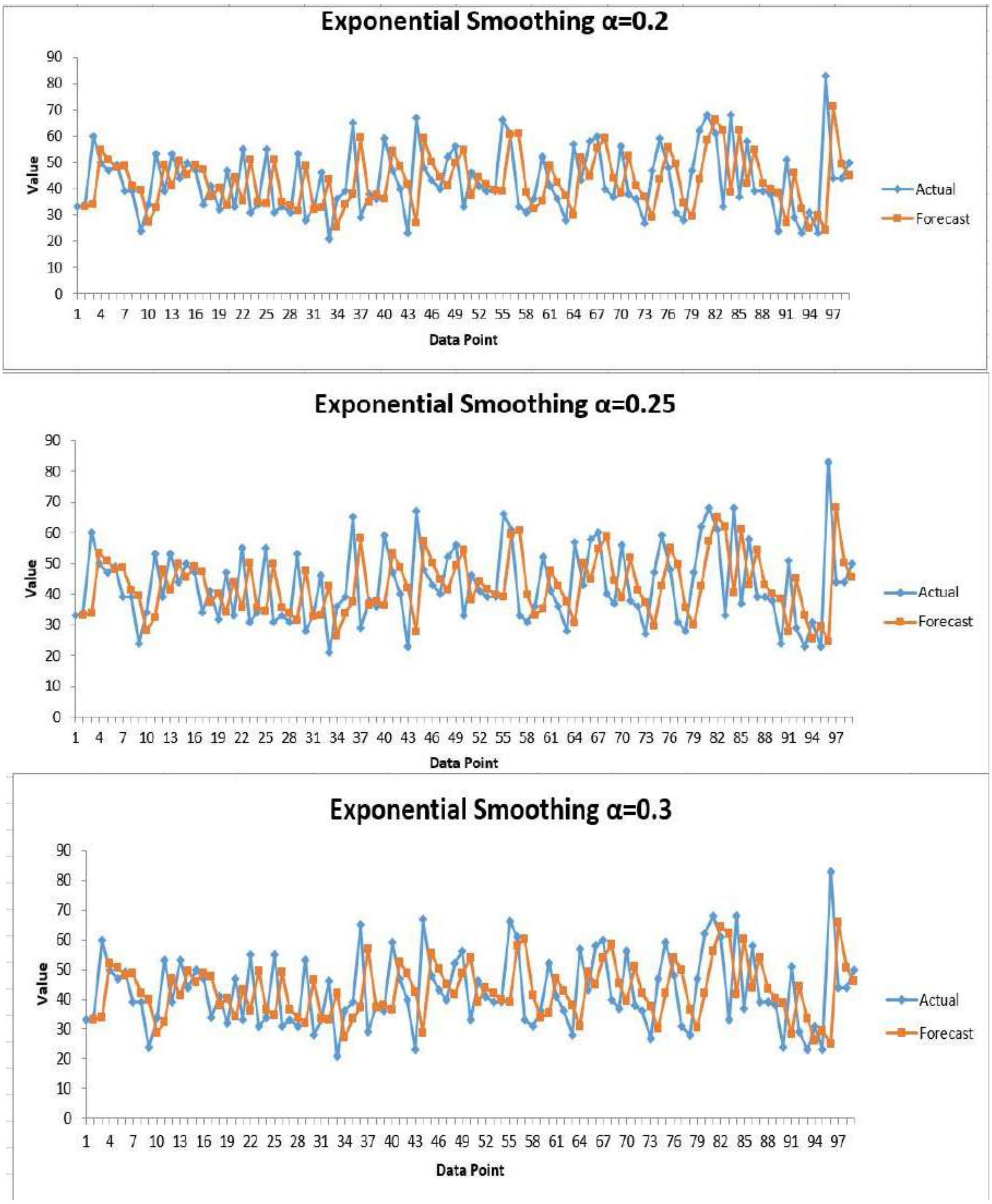


Figure 17: Graphs of exponential smoothing at α =0.2, 0.25 and 0.3

5.4. Median smoothing

Median smoothing. Use the exact dimensions of the smoothing interval and the operation like previously. We smooth the data using the size of the smoothing interval w = 3, 5, 7, 9, 11, 13, 15. We have to get seven columns (Fig.18 - Fig.19).

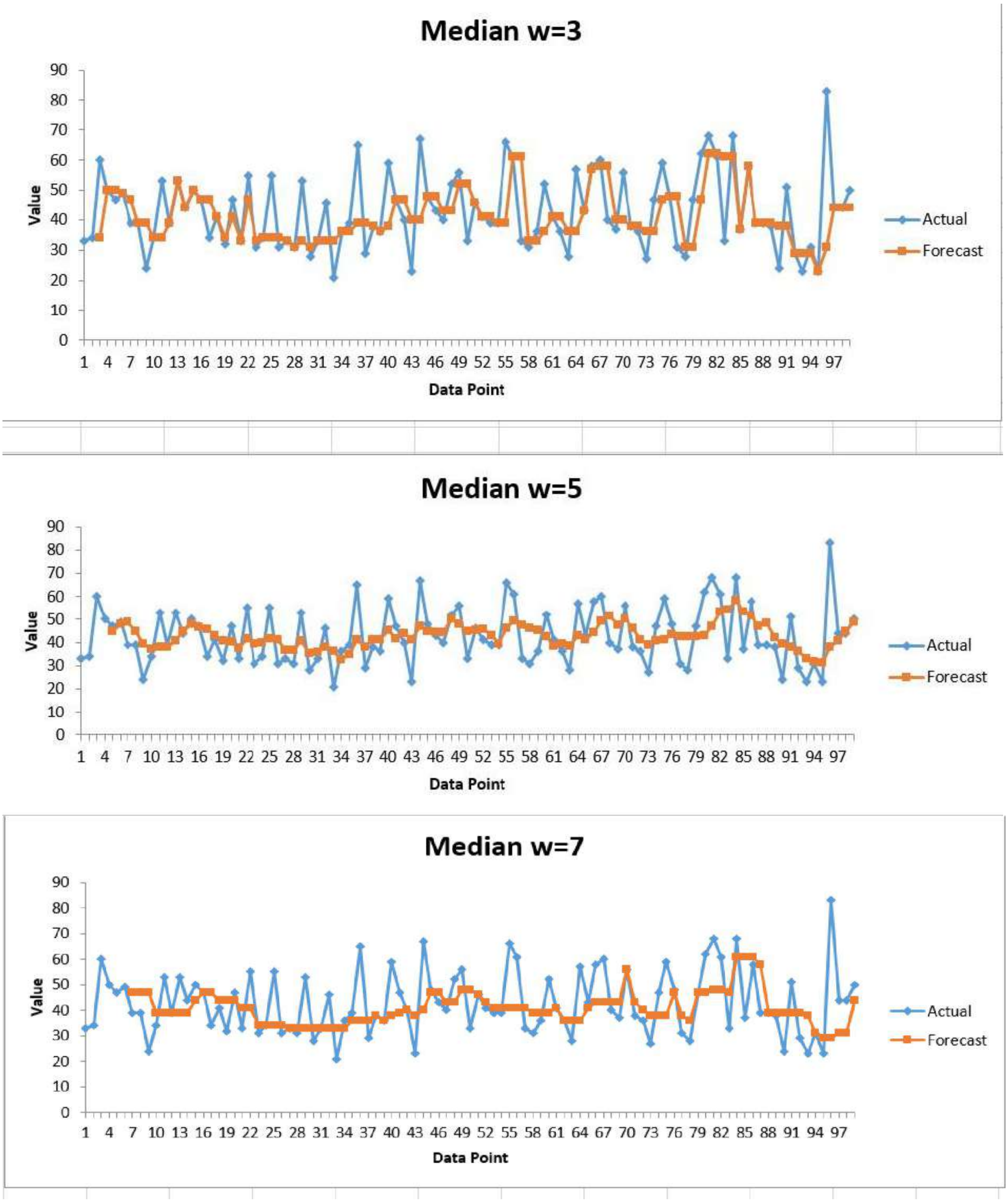


Figure 18: Median smoothing for women's clothing e-commerce review at w = 3-7

We smooth the data using the smoothing interval w = 3, then smooth the obtained smoothed data again, but use the size of the smoothing interval w = 5. Continue smoothing the obtained data with a smoothing interval of w = 7 and w = 15. We must get seven in a row-column (Fig.20 - Fig.21).

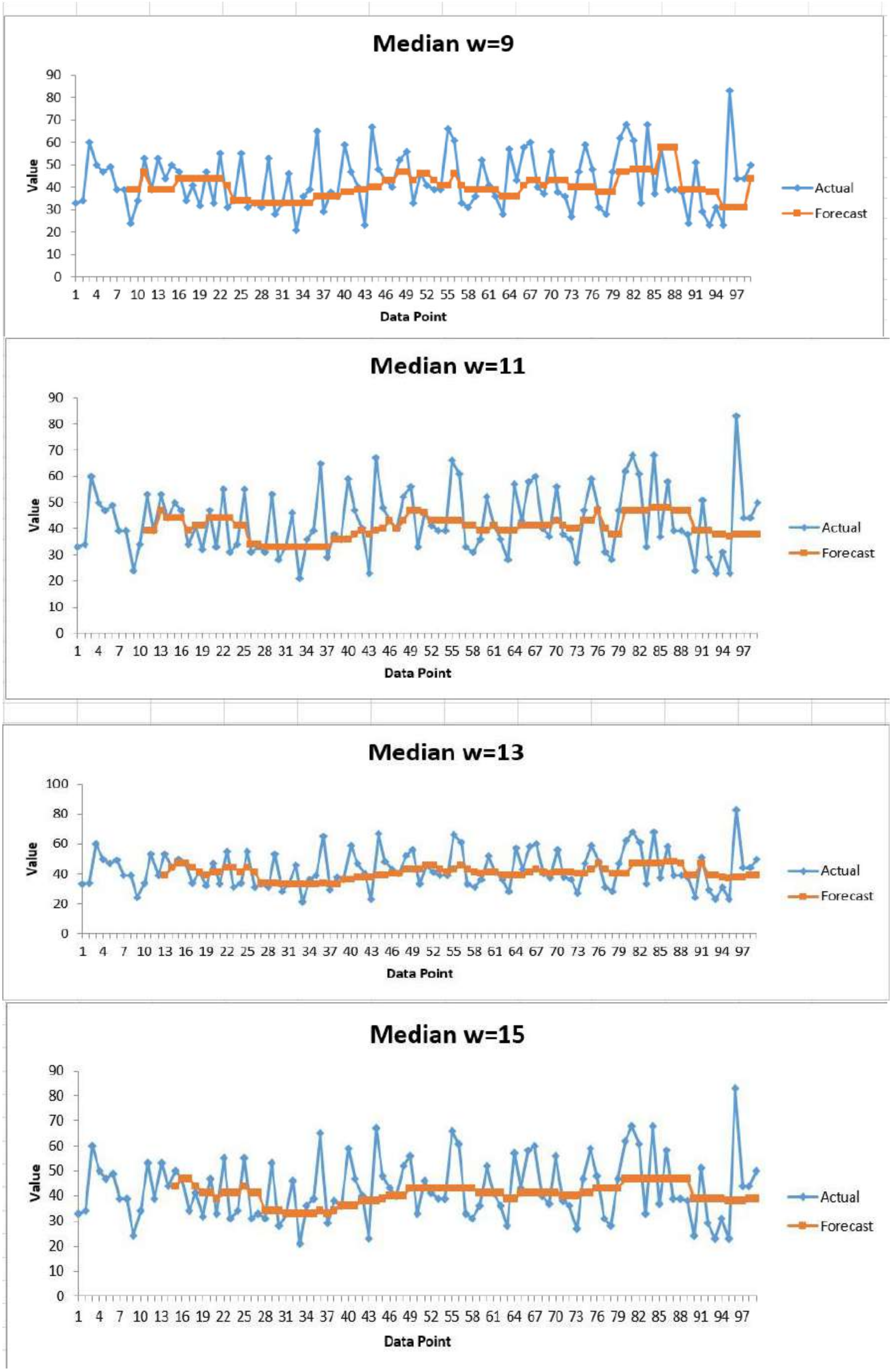


Figure 19: Graphs of median smoothing at w = 9, w = 11, w = 13, w = 15

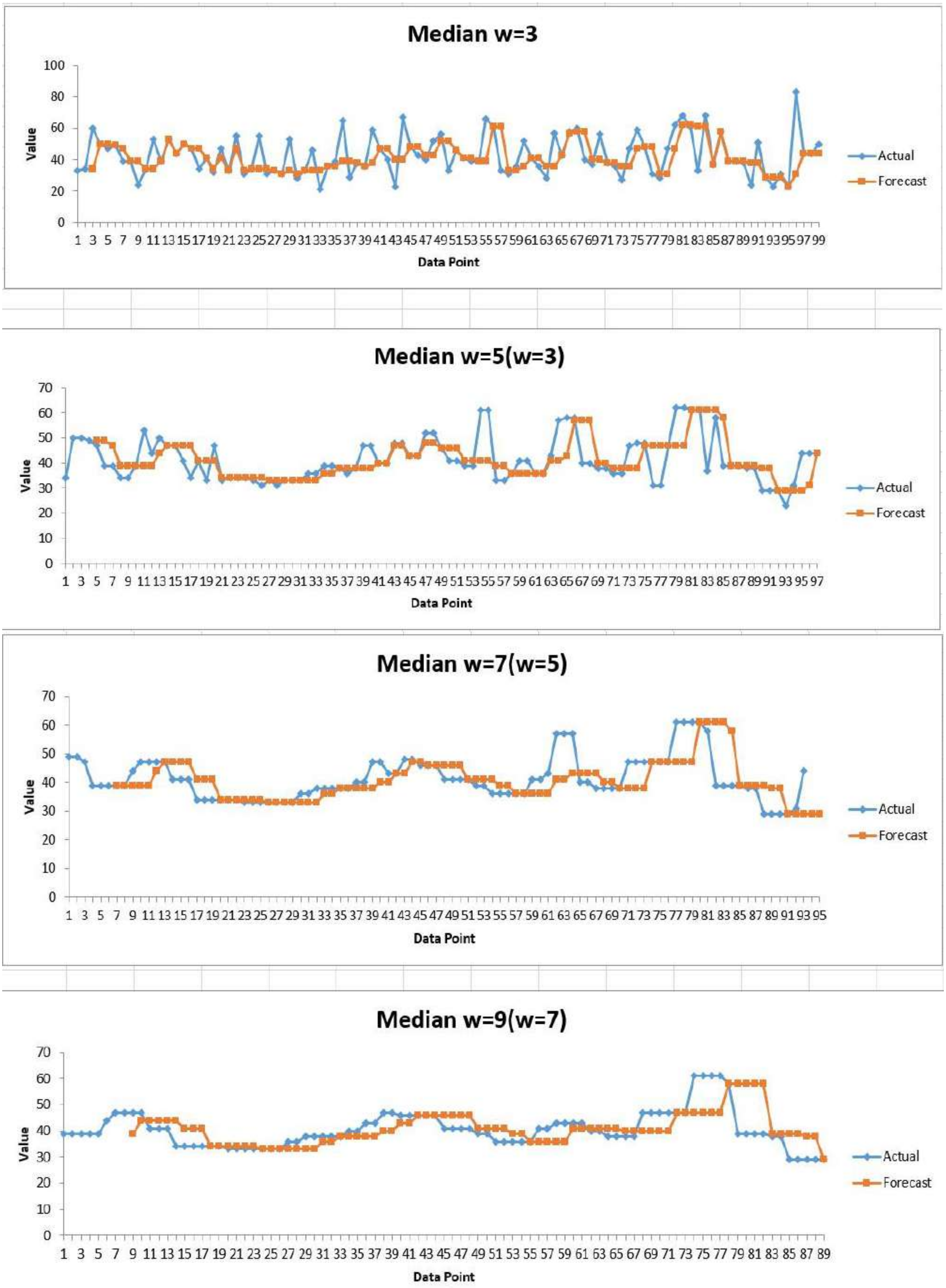


Figure 20: Graphs of median smoothing at w = 3, w = 5 (w = 3), w = 7 (w = 5), w = 9 (w = 7)

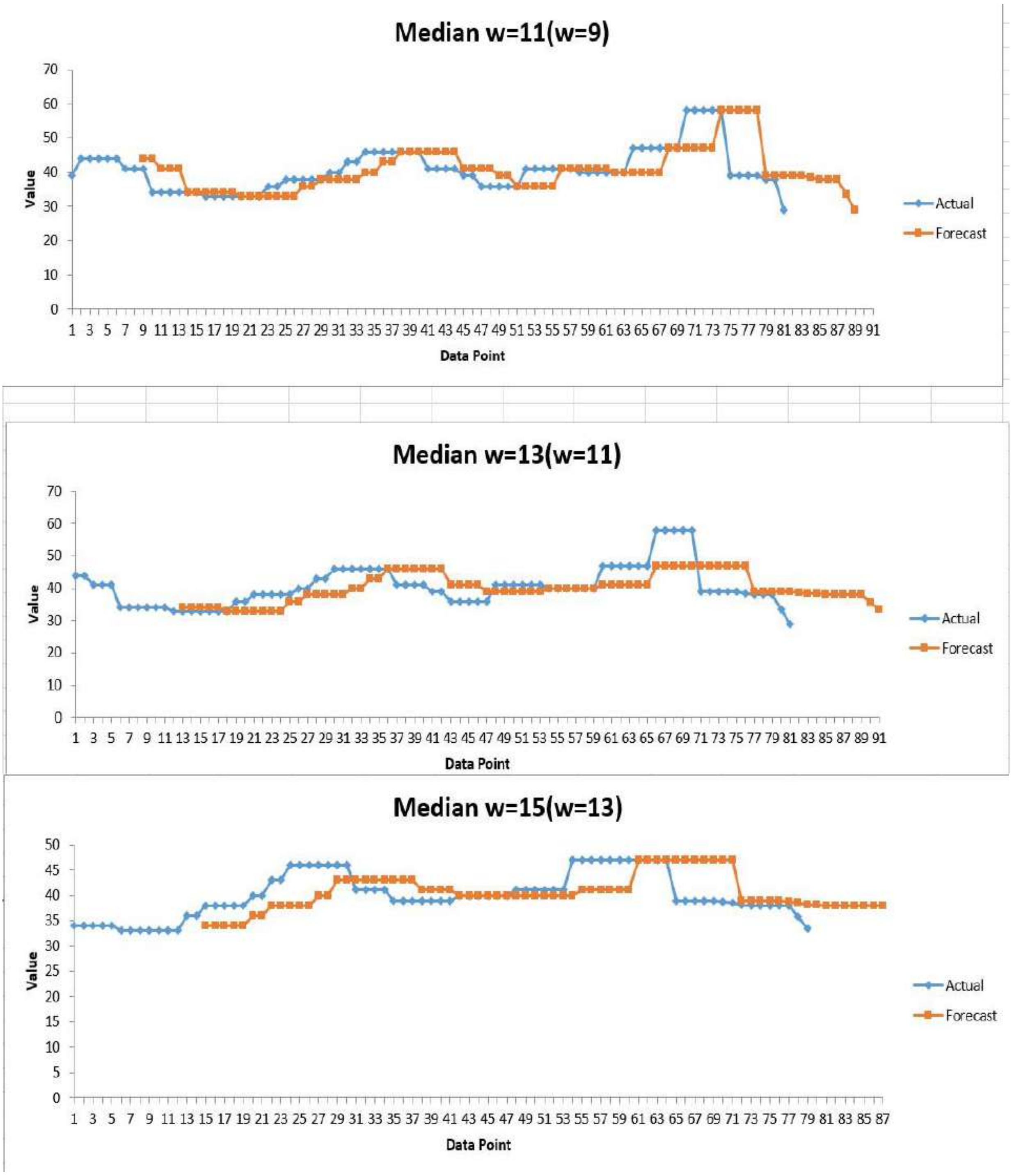


Figure 21: Graphs of median smoothing at w = 11 (w = 9), w = 13 (w = 11), w = 15 (w = 13)

5.5. Data correlation

We constructed a correlation field to visually understand the relationship between our studied traits. We chose such features as - Rating and Recommended IND to build the field. Where a rating is a rating of 1 to 5 for a specific product and Recommended, IND is a binary variable, where 0 means that the product is not recommended and 1 is recommended (Fig. 22). We also built correlation fields such as Age and Rating (Fig. 22) and Clothing ID vs Age (Fig. 22). To understand whether there is a relationship between the data, you need to calculate the correlation coefficient. The correlation coefficient characterizes the degree of closeness of the linear dependence.

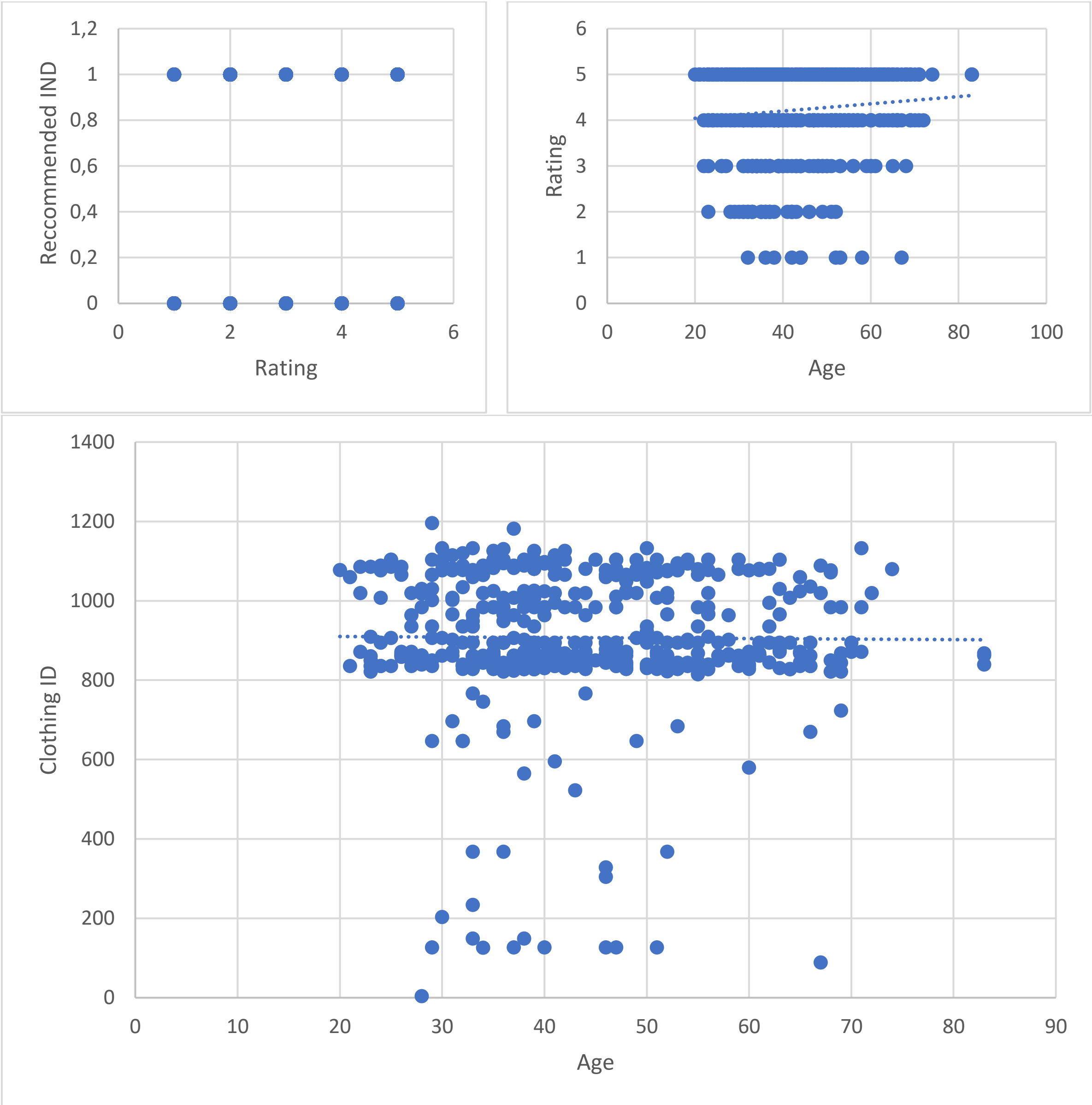


Figure 22: Correlation fields for Recommended IND vs Rating, Rating vs Age, Clothing ID vs Age

Calculate the correlation between the rating and the recommended IND. Calculate it by the formula: = CORREL (F2: F23487; G2: G23487). The correlation result is shown below: correlation coefficient R=0,792336288, and determination coefficient R2= 0,627736288. Thus, our correlation coefficient is about 0.79. They are significantly correlated because their correlation is close to 1. It is considered that the correlation coefficients, which are modulo more than 0.7, indicate a strong relationship between these features. We can conclude that clothes with higher ratings are more recommended for people. Calculate the correlation between rating and age. Calculate it by the formula: = CORREL (F2: F23487; C2: C23487). The correlation result is shown below: the correlation coefficient is R=0,026830575. Calculate the correlation between rating and clothing id. Calculate it by the formula: = CORREL (F2: F23487; B2: B23487). The correlation result is shown below: correlation coefficient R=-0,018879437. Correlation coefficients that are less than 0.5 modulo indicate a weak relationship. In the last two cases, our values do not correlate at all.

When the pairwise statistical dependence on the linear one is correlated, the correlation coefficient loses its meaning as a characteristic of the degree of closeness of the connection. In this case, use such a measure of communication as the correlation ratio.

Since there is a linear relationship between the pair of studied features, the correlation ratio does not need to be calculated.

5.6. Build of autocorrelation functions

An autocorrelation function correlates a function with itself shifted by a certain amount of independent variable. Autocorrelation is used to find patterns in several data, such as periodicity. The graph of the autocorrelation function is also called the correlogram. In Fig.23, we can see the result of autocorrelation.

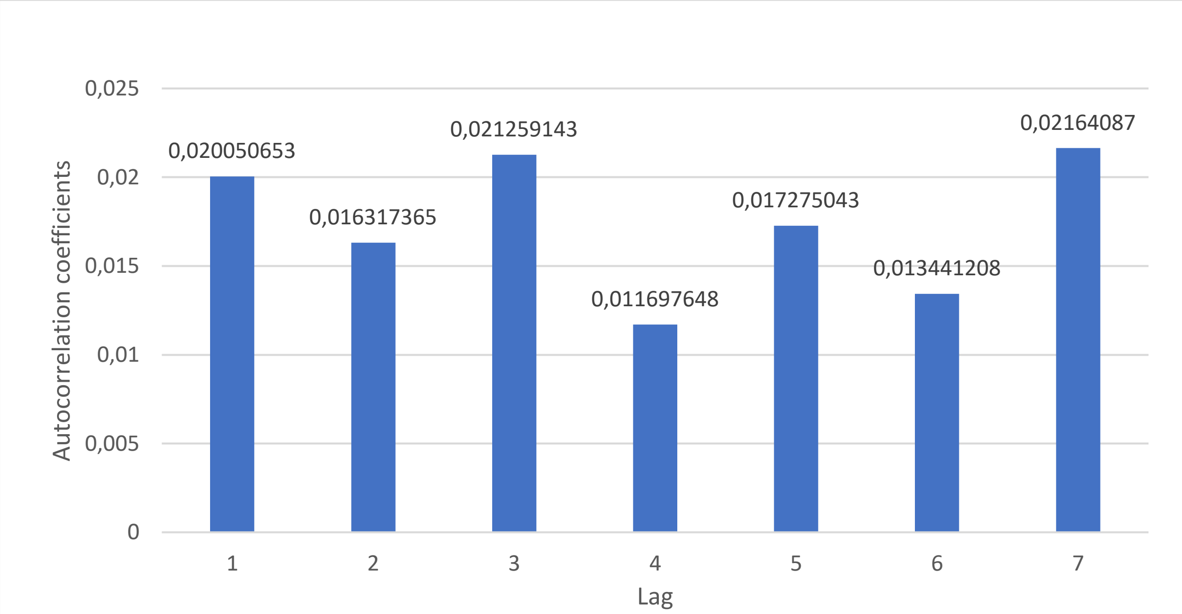


Figure 23: The graph of autocorrelation functions

Fig.23 shows that the studied series is not stationary. In the case of a stationary time series, the graph of autocorrelation functions should decline rapidly after the first few values.

We divided one of the sequences into three equal parts. For partitioning, we chose the sequence Rating and divided it into three equal parts with an interval of 7828. The result can be seen in Table 3. For convenience, we made it in a separate table. The correlation matrix is a square table where the correlation coefficient between the corresponding parameters is located at the intersection of the corresponding row and column (Table 4). The correlation matrix is a square table where the correlation coefficient between the corresponding parameters is located at the corresponding row and column intersection.

Table 3

The result of the division of the Rating sequence into three equal parts

Name	Part 1	Part 2	Part 3
Interval	[1;7829)	[7829;15658)	[15658;23487]
Range	7828	7828	7828

Table 4

A correlation matrix for three equal parts of the Rating

Name	Rating 1	Rating 2	Rating 3
Rating 1	1		
Rating 2	0,000290262	1	
Rating 3	-0,001842432	0,00664732	1

We use the CORREL function to calculate the autocorrelation coefficient in Excel to find the coefficients of multiple correlations. Assume that the base variable includes the range F1: F23487. Then the autocorrelation coefficient is presented in Table 5 and Fig. 24.

Table 5
Autocorrelation coefficients for the Rating vs Lag

Lag	Autocorrelation coefficient
1	0,020050653
2	0,016317365
3	0,021259143
4	0,011697648
5	0,017275043
6	0,013441208
7	0,02164087

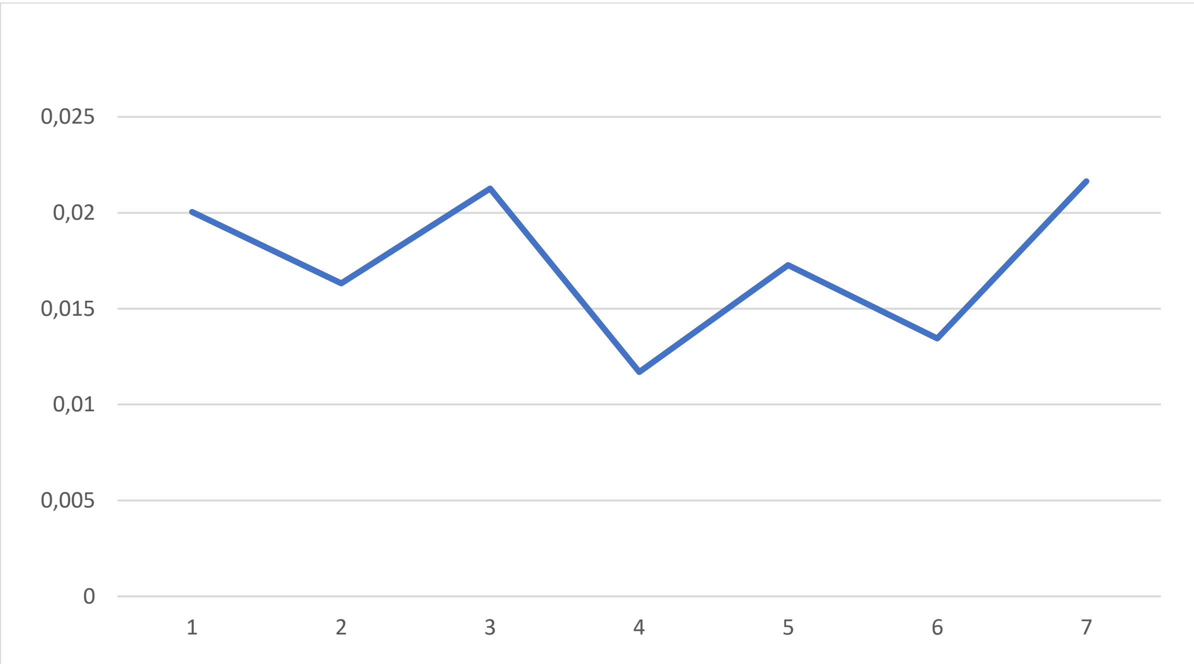


Figure 24: The graph of autocorrelation functions for Rating

5.7. Cluster data analysis

To conduct cluster analysis, we use an integrated data analysis and management system - Statistica, one of the most popular statistical programs for finding patterns, forecasting, classification, and data visualization. Before moving to Statistica, you need to prepare our data set using Excel. Namely, to create a table "object-property" by deriving the averages (Age, Rating, Recommended IND, Positive Feedback Count) for each type of clothing. Using data consolidation and applying the "average" function for indicators: Age, Rating, Recommended IND, and Positive Feedback Count (Table 6). For convenience, the data in the table have been sorted alphabetically.

The next step is to normalize the resulting Table 7. For this, use the formula. An Example of equation

$$z = \frac{x}{x_{max}}, \quad (1)$$

where x is the initial value and z is the normalized value.

Table 6
The table "object-property"

Class Name	Age	Rating	Recommended IND	Positive Feedback Count
Blouses	44.2525	4.15402	0.810138844	2.725217953
Casual bottoms	26.5	4.5	1	0
Chemises	38	4	1	0
Dresses	42.11489	4.150815	0.8081975	3.087513847
Fine gauge	44.73091	4.260909	0.837272727	2.013636364
Intimate	39.15584	4.279221	0.857142857	0.779220779
Jackets	43.81392	4.295455	0.845170455	2.826704545
Jeans	43.11595	4.360942	0.881429817	1.759372276
Knits	43.63081	4.161677	0.817674995	2.394796614
Layering	41.5274	4.376712	0.883561644	1.315068493
Legwear	41.54545	4.278788	0.860606061	1.272727273
Lounge	42.7178	4.301013	0.859623734	2.321273517
Outerwear	44.28659	4.198171	0.817073171	2.823170732
Pants	44.04755	4.26585	0.832853026	2.396974063
Shorts	40.72871	4.255521	0.839116719	1.675078864
Skirts	42.49206	4.245503	0.845502646	2.293121693
Sleep	43.10088	4.285088	0.855263158	1.750000000
Sweaters	45.06443	4.179272	0.800420168	2.208683473
Blouses	44.2525	4.15402	0.810138844	2.725217953

Table 7
The normalized table "object-property"

Class Name	Age	Rating	Recommended IND	Positive Feedback Count
Blouses	0.981983053	0.92311556	0.810138844	0.808730515
Casual bottoms	0.588046992	1	1	0
Chemises	0.843237195	0.888888889	1	0
Dresses	0.934548502	0.922403334	0.8081975	0.916244758
Fine gauge	0.992599114	0.946868687	0.837272727	0.597562911
Intimate	0.8688859	0.950937951	0.857142857	0.231240082
Jackets	0.972250721	0.954545455	0.845170455	0.838847484
Jeans	0.956762544	0.96909813	0.881429817	0.522107982
Knits	0.968187359	0.924817033	0.817674995	0.710675304
Layering	0.921511737	0.97260274	0.883561644	0.390257234
Legwear	0.921912436	0.950841751	0.860606061	0.377692133
Lounge	0.947927319	0.955780672	0.859623734	0.688856729
Outerwear	0.982739369	0.932926829	0.817073171	0.837798796
Pants	0.977435076	0.947966699	0.832853026	0.71132148
Shorts	0.90378843	0.945671223	0.839116719	0.497093229
Skirts	0.942918117	0.943445032	0.845502646	0.680502448
Sleep	0.956427969	0.952241715	0.855263158	0.519326683
Sweaters	1	0.928727046	0.800420168	0.655444722
Blouses	0.981983053	0.92311556	0.810138844	0.808730515

Now we can move on to cluster data analysis with Statistics. Let's transfer the normalized table to a separate sheet in Excel. We imported the sheet with the normalized table in Statistica. In our case, we choose the cluster method, Joining (tree clustering), i.e., hierarchical classification. We select all the values for analysis. Note that the file contains raw data, not a matrix of similarities, rows group clusters. We also choose the Euclidean distance as a metric for constructing a proximity matrix. Choose Single Linkage for the merger strategy. We build a dendrogram (Fig. 25).

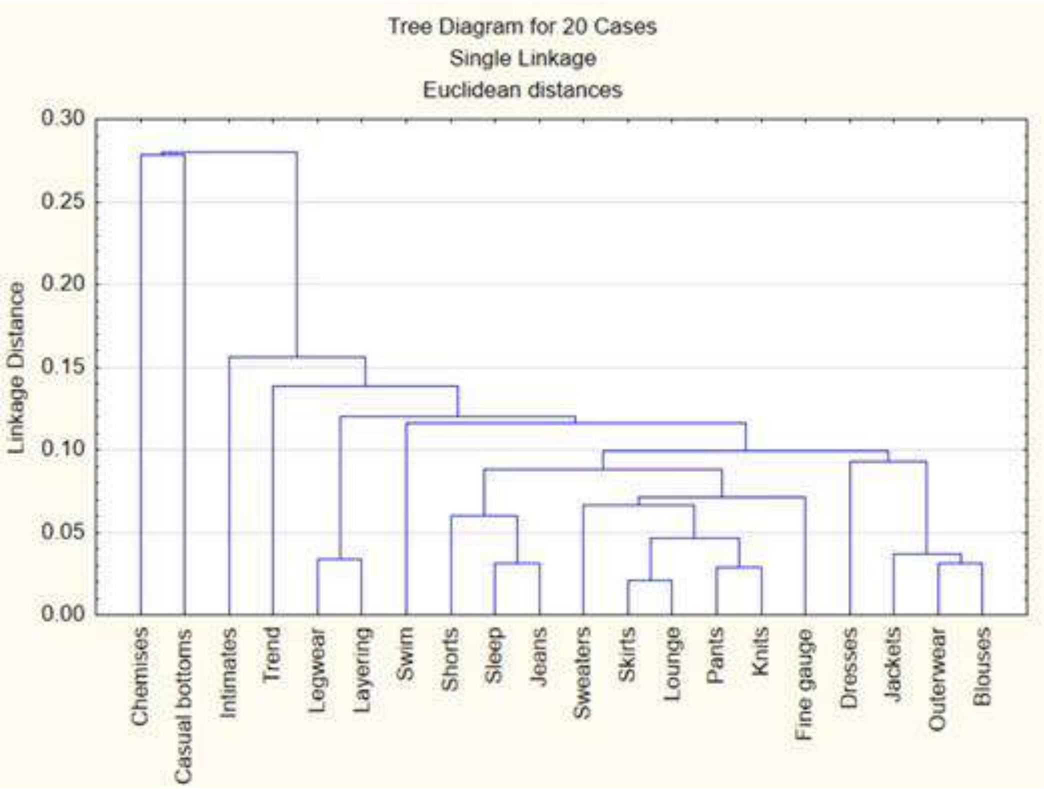


Figure 25: The dendrogram

Analyzing the resulting dendrogram, we can conclude that Skirts and Lounge have the most similar values for the variables Age, Rating, Recommended IND, and Positive Feedback Count, which is why they are combined into a standard cluster. By the same analogy, all other variables and clusters are merged until the last standard cluster is formed.

From the obtained dendrogram, we can conclude that customers who ordered Skirts most likely belong to the same age category as customers who ordered Lounge. They also most similarly evaluate the product. Next in similarity are Pants and Knits and so on.

This information helps us understand customers' needs and recommend the product concerning their previous purchases, which will help increase sales and profits of the online clothing store.

6. Conclusions

In this paper, we analyzed a dataset of opinions of consumers of women's clothing obtained as a result of reviews and comments during online sales. The study used various data analysis methods, using well-known software environments such as Excel and Statistica. It allows you to determine which clothes will bring more revenue to the company and which will increase the profitability of the online clothing store. The high popularity of clothing and footwear as a segment of the electronic market is considered.

Correlation analysis of survey data was performed. Correlation coefficients were calculated. A correlation matrix was constructed, and autocorrelation was established, which allowed establishing that very little data correlate with each other and therefore do not depend on each other entirely. A study of how consumers perceive the products and services offered in the clothing segment revealed that clothes with higher ratings are more recommended to buyers. After buying a product, it was also found that most people, about 80%, will recommend it and leave a positive response. Only 20% cannot recommend this product and remain dissatisfied. Since we have analyzed and understood which clothes are most often bought, we can conclude what we need to promote and emphasize to increase the store's popularity and profitability. Accordingly, the things that have the lowest reviews and are not recommended by buyers show the latter.

Cluster data analysis was performed, and dendrograms of clothing sales responses were constructed and analyzed. Due to the conclusions, we obtained from various research methods of the clothing sales segment on the Internet, recommendations for improving the clothing sales system, and proposals for developing new marketing measures.

7. References

[1] V. Vysotska, A. Berko, M. Bublyk, L. Chyrun, A. Vysotsky, K. Doroshkevych, Methods and tools for web resources processing in e-commercial content systems, in: Proceedings of 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 1, 2020, pp. 114-118.

[2] 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.

[3] 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.

[4] L. Chyrun, The E-Commerce Systems Modelling Based on Petri Networks, CEUR Workshop Proceedings Vol-2870 (2021) 1604-1631.

[5] A. Berko, M. Bublyk, L. Chyrun, Y. Matseliukh, R. Levus, V. Panasyuk, O. Brodyak, L. Dzyubyk, O. Garbich-Moshora, Models and methods for E-commerce systems designing in the global economy development conditions based on Mealy and Moore machines, volume Vol-2870 of CEUR Workshop Proceedings, 2021, pp. 1574 – 1593.

[6] J. Hammond, K. Kohler, E-commerce in the textile and apparel industries. Tracking a Transformation: E-commerce and the Terms of Competition in Industries, 2000.

[7] W. Jinfu, Z. Aixiang, E-commerce in the textile and apparel supply chain management: Framework and case study, in: 2009 IEEE Second International Symposium on Electronic Commerce and Security, 1, pp. 374-378.

[8] 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

[9] 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.

[10] 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.

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

[12] 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.

[13] V. Vysotska, Linguistic Analysis of Textual Commercial Content for Information Resources Processing, in: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 2016, pp. 709-713. doi: 10.1109/TCSET.2016.7452160.

[14] A. Dmytriv, V. Vysotska, M. Bublyk, The Speech Parts Identification for Ukrainian Words Based on VESUM and Horokh Using, in: Proceedings of 16th IEEE International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2021, 2, pp. 21–33. DOI: 10.1109/CSIT52700.2021.9648813
[15] 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.
[16] 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.
[17] 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
[18] K. Rastogi, Automation of EDA & Text processing. Exploring data and visualizing with Sentiment Analysis, 2021, URL: https://medium.com/analytics-vidhya/women-ecommerce-clothing-part-1-86b1acd19ffa.
[19] K. T. Cosine, Mining the Women's Clothing Reviews, 2022, URL: http://rstudio-pubs-static.s3.amazonaws.com/405423_5974e2e8473b486ea3d632edba55c 3fe.html
[20] 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.
[21] K. P. Anuj, P. Abhishek, K. J. Mradul, Sentimental Analysis on E-commerce Women's Clothing, in: International Journal for Research in Engineering Application & Management (IJREAM), 2020, 06 (05), DOI: 10.35291/2454-9150.2020.0555
[22] 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
[23] 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 on Data Stream Mining and Processing, DSMP, 2020, pp. 432-437.
[24] Xiaoxin Li, Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing, in: Proceedings of the 2020 2nd International Conference on Big Data and Artificial Intelligence, ISBDAI'20, 2020, pp. 32–36. DOI: 10.1145/3436286.3436293.
[25] 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.
[26] 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.
[27] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert systems with applications, 2009, 36(3), pp. 6527-6535.
[28] M. Bublyk; V. Lytvyn, V. Vysotska, L. Chyrun, Y. Matseliukh, N. Sokulska, The decision tree usage for the results analysis of the psychophysiological testing, volume Vol-2753 of CEUR Workshop Proceedings, 2020, pp. 458 – 472.
[29] Women e-commerce clothing reviews, Kaggle Dataset. URL: https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews
[30] Kaggle Dataset, URL: https://www.kaggle.com/nicapotato/womens-ecommerce-clothing.
[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Hierarchical agglomerative cluster analysis of one-dimensional asymmetrically distributed data in the MS EXCEL environment, 2022. URL: http://science.lp.edu.ua/sites/default/files/Papers/plugin-12_216.pdf.

[36] Hierarchical clusterig dendrogram, 2022. URL: https://quantdare.com/hierarchical-clustering/ dendrograma.

[37] M. Bublyk, A. Kowalska-Styczen, V. Lytvyn, V. Vysotska, The Ukrainian Economy Transformation into the Circular Based on Fuzzy-Logic Cluster Analysis. Energies 2021, 14, 5951. https://doi.org/10.3390/en14185951

[38] 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.

[39] 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.

[40] 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.

[41] S., Mashtalir, O., Mikhnova, M. Stolbovyi, Multidimensional Sequence Clustering with Adaptive Iterative Dynamic Time Warping. International Journal of Computing 18(1), (2019) 53-59.

[42] P. Zhezhnych, O. Markiv, Recognition of tourism documentation fragments from web-page posts, in: Proceedings of the14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET, 2018, pp. 948–951.

[43] P. Zhezhnych, A. Shilinh, V. Tkachuk, Computer-linguistic selection of potential entrances' motivation intentions from posts of education web-community, CEUR Workshop Proceedings 2392 (2019).

[44] A. Abdaoui, J. Aze, S. Bringay, N. Grabar, P. Poncelet, Analysis of Forum Posts Written by Patients and Health Professionals. In: Studies in Health Technology and Informatics 205 (2014) 1185.

[45] S. Albota, Linguistic and Psychological Features of the Reddit News Post, in: Proceedings of the IEEE 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2020, 1, pp. 295–299.

[46] S. Albota, Linguistically manipulative, disputable, semantic nature of the community reddit feed post, CEUR Workshop Proceedings 2870 (2021) 769–783.

[47] 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 IEEE 16th International conference on Computer science and information technologies, Lviv, 2021, pp. 85–97. doi: 10.1109/CSIT52700.2021.9648594.

[48] L. Chyrun, Y. Burov, B. Rusyn, L. Pohreliuk, O. Oleshek, A. Gozhyj, I. Bobyk, Web resource changes monitoring system development, volume 2386 of CEUR Workshop Proceedings, 2019, pp. 255-273.

[49] V. Vysotska, V.B. Fernandes, M. Emmerich, Web content support method in electronic business systems, CEUR Workshop Proceedings Vol-2136 (2018) 20-41.

[50] A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska, Uniform method of operative content management in web systems, volume 2136 of CEUR Workshop Proceedings, 2018, pp. 62-77.

[51] L. Chyrun, V. Andrunyk, L. Chyrun, A. Berko, I. Dyyak, N. Antonyuk, Online Business Processes Support Methods, in: Proceedings of the IEEE 15th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2020, 1, pp. 125–133.
[52] A. Berko, L. Chyrun, I. Dyyak, V. Andrunyk, L. Chyrun, N. Antonyuk, E-Commercial Systems Designing Methods for Virtual Enterprise, in: Proceedings of the Int. Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 2020, 1, pp. 119–124.
[53] R. Yurynets, Z. Yurynets, I. Myshchyshyn, N. Zhyhaylo, A. Pekhnyk, Optimal Strategy for the Development of Insurance Business Structures in a Competitive Environment, CEUR Workshop Proceedings Vol-2631 (2020) 79-94.
[54] R. Yurynets, Z. Yurynets, M. Kokhan, Econometric Analysis of the Impact of Expert Assessments on the Business Activity in the Context of Investment and Innovation Development, CEUR workshop proceedings Vol-2604 (2020) 680-694.
[55] A. Kopp, D. Orlovskyi, S. Orekhov, An Approach and Software Prototype for Translation of Natural Language Business Rules into Database Structure, CEUR Workshop Proceedings Vol-2870 (2021) 1274-1291.
[56] H. Lipyanina, A. Sachenko, T. Lendyuk, S. Nadvynychny, S. Grodskyi, Decision tree based targeting model of customer interaction with business page, CEUR Workshop Proceedings 2608 (2020) 1001-1012.
[57] I. Oksanych, I. Shevchenko, I. Shcherbak, S. Shcherbak, Development of specialized services for predicting the business activity indicators based on micro-service architecture, Eastern-European Journal of Enterprise Technologies 2(2-86) (2017) 50-55.
[58] A. Y. Berko, Methods and models of data integration in E-business systems, Actual Problems of Economics (10) (2008) 17-24.
[59] A. Berko, Consolidated data models for electronic business systems, in: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 2007, pp. 341-342.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] L. Chyrun, I. Turok, I. Dyyak, Information model of the tendering system for large projects, CEUR Workshop Proceedings 2604 (2020) 1224-1236.
[65] 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.
[66] 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.
[67] 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.
[68] 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 on Data Stream Mining and Processing, DSMP, 2020, pp. 432-437. doi: 10.1109 / DSMP47368.2020.9204321.
[69] 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.