IRSTI 83.77.29

Cluster analysis application in the compulsory insurance of civil-legal liability of the vehicles' owners

Sikhov M.B., Al-Farabi Kazakh National University,
Almaty, Kazakhstan, E-mail: mirbulats56@gmail.com
Beibitbekov A.B., Al-Farabi Kazakh National University,
Almaty, Kazakhstan, E-mail: beibitbekov.almas@gmail.com
Sapin A.M., Al-Farabi Kazakh National University,
Almaty, Kazakhstan, E-mail: azat.sapin@gmail.com

With an increase in flow of the processed and stored information in insurance organizations in Kazakhstan, associated with the building of customers' base, mergers and acquisitions processes and implementation of the new insurance products; the relevance of the problem of preliminary information processing for its structuring, allocation of distinctive attributed, generalization and sorting grows. Without appropriate scientific and methodological approach, data processing and analysis will be more difficult for insurance organizations and, may require the utilization of significant informational-computing and financial resources. In the present article as a modern scientific-research approach to the solution of this problem, it is suggested to apply a procedure of the cluster analysis by k-means algorithm, which makes it possible to simplify the processing and further analysis of data set by arranging data in relatively homogeneous groups. Particularly, the present article describes a process of the cluster analysis application by the k-means algorithm to the data on losses by a class of Compulsory insurance of civil-legal liability of the vehicles' owners. The purpose of the present article is to split the losses by this class of insurance into homogeneous qualitative groups (clusters) based on frequency and severity of losses and, to interpret acquired clusters. Results of the k-means algorithm confirm that each acquired cluster has statistically significant data with similar impact upon losses' process, which may be employed in the future for evaluation of losses of the insurance organization. Methodological approaches and results obtained in this article will, first of all, be interesting to the professional participants of insurance market of the Republic of Kazakhstan to conduct better underwriting research on the formation of the efficient structure of the insurance portfolio of Compulsory insurance of civil-legal liability of the vehicles' owners in accordance with tariff rates.

Key words: cluster analysis, unsupervised machine learning, k-means algorithm, insurance, underwriting analysis.

Көлік құралдары иелерінің азаматтық-құқықтық жауапкершілігін міндетті сақтандыруда кластерлік талдауды қолдану

Сихов М.Б., Әл-Фараби атындағы Қазақ ұлттық университеті,
Алматы қ., Қазақстан, E-mail: mirbulats56@gmail.com
Бейбітбеков А.Б., Әл-Фараби атындағы Қазақ ұлттық университеті,
Алматы қ., Қазақстан, E-mail: beibitbekov.almas@gmail.com
Сапин А.М., Әл-Фараби атындағы Қазақ ұлттық университеті,
Алматы қ., Қазақстан, E-mail: azat.sapin@gmail.com

Қазақстандағы сақтандыру ұйымдарында өңделетін және сақталатын ақпарат ағынының өсуімен, оның ішінде клиенттік базаны арттырумен, бірігу және сіңіру процестерімен және жаңа сақтандыру өнімдерін енгізумен байланысты, ақпаратты алдын ала өңдеу проблемаларының оны құрылымдау, сипатты белгілерді бөлу, қорыту және сұрыптау үшін өзектілігі өсуде. Тиісті ғылыми және әдіснамалық тәсілсіз деректерді өңдеу және талдау процесі сақтандыру ұйымдары үшін неғұрлым қиын болып табылуда және елеулі ақпараттық-есептеу және қаржы ресурстарын пайдалану талап етілуі мүмкін. Бұл мақалада осы проблеманы шешудің қазіргі заманғы ғылыми-зерттеу тәсілі ретінде салыстырмалы біртекті топтарға деректерді реттеу жолымен деректер массивтерін өңдеуді және одан әрі талдауды жеңілдетуге мүмкіндік беретін k-орташа әдісімен кластерлік талдау рәсімін пайдалану ұсынылады. Атап айтқанда, мақалада көлік құралдары иелерінің азаматтық-құқықтық жауапкершілігін міндетті сақтандыру сыныбы бойынша шығындар жөніндегі деректерге k-орташа әдісімен кластерлік талдауды қолдану процесі сипатталады. Мақаланың мақсаты сақтандырудың осы сыныбы бойынша шығындарды, шығындардың жиілігі мен ауырлығы негізінде біртекті сапалы топтарға (кластерлерге) бөлу және алынған кластерлерді түсіндіру болып табылады. K-орташа әдісінің нәтижелері әрбір бөлінген кластерде одан әрі сақтандыру ұйымының залалдарын бағалау үшін пайдаланылуы мүмкін шығындар процесіне ұқсас статистикалық маңызды деректер бар екенін куәландырады. Мақалада алынған әдіснамалық тәсілдер мен нәтижелер ең алдымен Қазақстан Республикасының сақтандыру нарығының қатысушыларына тарифтік ставкаларға сәйкес көлік құралдары иелерінің азаматтық-құқықтық жауапкершілігін міндетті сақтандыру бойынша сақтандыру портфелінің тиімді құрылымын қалыптастыру бойынша неғұрлым сапалы андеррайтингтік зерттеу жүргізу үшін қызықты болып табылады.

Түйін сөздер: кластерлік талдау, оқытушысыз машиналық оқыту, k-орташа әдісі, сақтандыру, андеррайтинг талдау.

Применение кластерного анализа в обязательном страховании гражданско-правовой ответственности владельцев транспортных средств

Сихов М.Б., Казахский национальный университет имени аль-Фараби,
г. Алматы, Казахстан, E-mail: mirbulats56@gmail.com
Бейбітбеков А.Б., Казахский национальный университет имени аль-Фараби,
г. Алматы, Казахстан, E-mail: beibitbekov.almas@gmail.com
Сапин А.М., Казахский национальный университет имени аль-Фараби,
г. Алматы, Казахстан, E-mail: azat.sapin@gmail.com

С ростом потока обрабатываемой и хранимой информации в страховых организациях в Казахстане, связанных, в том числе, с наращиванием клиентской базы, процессами слияний и поглощений и внедрением новым страховых продуктов, растет актуальность проблем предварительной обработки информации для ее структурирования, выделения характерных признаков, обобщения и сортировки. Без соответствующего научного и методологического подхода процесс обработки и анализа данных будет становится все более затруднительным для страховых организаций, и может потребоваться использование значительных информационно-вычислительных и финансовых ресурсов. В настоящей статье в качестве современного научно-исследовательского подхода к решению данной проблемы предлагается использовать процедуру кластерного анализа k-means algorithm, позволяющую упростить обработку и дальнейший анализ массивов данных путем упорядочивания данных в сравнительно однородные группы. В частности, в статье описывается процесс применения кластерного анализа k-means algorithm к данным по убыткам по классу обязательного страхования гражданско-правовой ответственности владельцев транспортных средств. Цель статьи состоит в том, чтобы разделить убытки по данному классу страхования на однородные качественные группы (кластеры) на базе частоты и тяжести убытков и интерпретировать полученные кластеры. Результаты k-means algorithm свидетельствуют о том, что в каждом из выделенных кластеров находятся статистически значимые данные со схожим влиянием на процесс убытков, которые могут быть использованы в дальнейшем для оценки убытков страховой

организации. Методологические подходы и результаты, полученные в статье, будут прежде всего интересны участникам страхового рынка Республики Казахстан для проведения более качественного андеррайтингового исследования по формированию эффективной структуры страхового портфеля по обязательному страхованию гражданско-правовой ответственности владельцев транспортных средств в соответствии с тарифными ставками.
Ключевые слова: кластерный анализ, машинное обучение без учителя, k-means algorithm, страхование, андеррайтинговый анализ.

1 Introduction

In the general insurance market of the Republic of Kazakhstan, one of the main classes of insurance is Compulsory Insurance of Civil Liability of Motor Vehicle Owners [1]. Due to the compulsory nature of insurance and the annual growth of automobile sales, this class of insurance prevails in the overall structure of premiums in Kazakhstan [2].

To date, in the general insurance branch the insurance companies of Kazakhstan have accumulated enough statistical data on this class of insurance required for underwriting research on the formation of an effective structure of the insurance portfolio and the allocation of target segments in accordance with certain tariff rates [3], which, in its turn, is necessary for the financial stability of insurance companies [4].

However, due to the growing flow of the information processed and stored in insurance companies, it is becoming more and more difficult to structure it accurately and highlight characteristic features, as well as generalize and draw rational conclusions [5].

As a modern scientific research approach to solving this problem, the authors of this article propose to use the k-means algorithm procedure, which allows to simplify the processing and further analysis of complex data by organizing it into relatively homogeneous groups [6].

In general, cluster analysis is one of the types of multidimensional classification in the absence of prior information about the number and type of classes into which the set of objects is divided [7]. In the framework of this article, the purpose of cluster analysis using the k-means algorithm is to split losses of the aforementioned insurance class into homogeneous qualitative groups (clusters) based on the frequency and severity of losses, each of which corresponds to a certain risk group.

Knowledge of the main descriptive characteristics in each cluster can be used further in the framework of underwriting to identify ineffective insurance portfolio and pricing errors, if any, in order to minimize the risks of losing funds and improve the financial stability of insurance organizations [8].

2 Literature Review

As already mentioned, cluster analysis provides an opportunity to learn about the structure of complex data by splitting them into similar objects (parts) [9]. In cluster analysis there are no pre-classified classes and no differences between dependent and independent variables. Cluster analysis algorithms detect similarities and group data into clusters.

Clustering methods are widely used in many areas such as marketing, pattern classification, biology, mathematics, etc. [10]. In business, clustering helps a marketer to characterize customer segmentation [11] and then direct marketing efforts to the most attractive segment. In biology, cluster analysis can be used with a view to classify genes [12] and to obtain

taxonomies of plants and animals [13]. Cluster analysis is used not only as a method of classification and segmentation, but also as a method of detecting fraudulent actions in the banking sector [14], property insurance [15] and health insurance [16].

There are many publications and research projects on the application of cluster analysis in the field of auto insurance. However, most of them approach the issue of application of cluster analysis from a marketing point of view and investigate the issue of identifying the most optimal customer segmentation for insurance organizations. For example, the authors Thakur S.S. and Sing J.K. [17] identified target customer segments for insurance companies in terms of customer interest in insurance by using cluster analysis. In another paper, the authors Kaveh K-D., Farshid A., and Shaghayegh A. [18] identified target segments for customers in terms of their profitability for insurance organizations using cluster analysis as well.

Among the articles dedicated to the study of losses of the insurance portfolio of auto insurance with the help of cluster analysis, we could highlight the article of the authors Ai C.Y., Kate A.S., Robert J.W. and Malcolm B. [19]. In this article, the authors consider the problem of forecasting claims and losses, taking into account the estimated risks for groups of policyholders. However, as noted above, only a small part of foreign clustering research is dedicated to the study and prediction of losses in the field of auto insurance. Moreover, in Kazakhstan, research on issues of claims and losses in the class of Compulsory Insurance of Civil Liability of Motor Vehicle Owners on the basis of cluster analysis has not been conducted to date, which only underlines the relevance and importance of this research direction for the local insurance market. The implementation of cluster analysis is not simple due to two factors [20]. First, the same clustering method can often produce different results. Thus, the final results in the framework of the same method will depend on the choice of parameters, such as the initial setting or the number of clusters [21]. Secondly, the interpretation of cluster structures is not simple. In this case, the detected clusters depend not only on the data, but also on the user's goal and the degree of granulation [22]. Ultimately, the resulting clusters should be considered as a representation of the data that can be used to restore the original data from aggregate clusters [23].

After an in-depth study of the available clustering methods [24], the authors of this article came to the conclusion that the k-means algorithm is the most optimal method for studying the nature of losses in the class of Compulsory Insurance of Civil Liability of Motor Vehicle Owners. The undoubted advantages of this method include the fact that it is relatively scalable and efficient in processing data sets [25]. Also, in favor of the choice of the k-means algorithm, its popularity among researchers is also due to its ease of use [26]. The k-means algorithm is used as a method of segmentation and classification more often than other clustering methods [27].

3 Materials and research methods

The following describes k-means algorithm – the most popular method of cluster analysis. In general, the k-means algorithm is a cluster analysis method, the objective of which is to split l observations from the multidimensional space R^n into k clusters, but each such observation relates to the cluster (group) whose centroid (average) it is closest to [28].

To begin with, let us take a detailed look at the following series of observations:

$$\left(x^{(1)}, x^{(2)}, \ldots, x^{(l)}\right), \quad x^{(i)} \in R^n \tag{1}$$

K-means algorithm splits l observations into k clusters ($k \leq l$), $G = (G_1, G_2, \ldots, G_k)$ in order to minimize the total squared deviation of cluster points from the centroids of these clusters:

$$min \left[\sum_{i=1}^{k} \sum_{x^{(j)} \in G_i} \|x^{(j)} - \mu_i\|^2 \right], \tag{2}$$

where, $x^{(j)} \in R^n$, $\mu_i \in R^n$ – centroid for a G_i cluster.

Thus, if the measure of the distance to the centroid is defined, then splitting objects into clusters is reduced to the determination of the centroids of these clusters. In this case, the number of clusters k is set in advance.

Let us consider the initial set of k centroids $\mu_1, \mu_2, \ldots, \mu_k$ in clusters $G_1, G_2, \ldots, G_k$. At the first stage, cluster centroids can be selected randomly. Further, we will assign each observation to the cluster whose centroid is closest to it. Each such observation should belong to only one cluster, even if it can be attributed to two or more clusters.

After the first iteration, the centroid of each i-th cluster is recalculated using the following formula:

$$\mu_j = \frac{1}{G_j} \sum_{x^{(j)} \in G_i} x^{(j)}. \tag{3}$$

Thus, the k-means algorithm involves recalculation of the centroid at each iteration step based on the information obtained in the previous step.

In this case, the iterative process of the algorithm of the k-means algorithm stops when the values of μ_i stop changing $\mu_i^{(t)} = \mu_i^{(t+1)}$.

4 Application of cluster analysis in Compulsory Insurance of Civil Liability of Motor Vehicle Owners

Let us suppose that we have the following aggregated data for the class of Compulsory Insurance of Civil Liability of Motor Vehicle Owners (Tab. 1 - Aggregated insurance data):

Table 1. Aggregated insurance data

Number of insured persons	100 000
Insurance premium per 1 insured person in tenge	20 000
Total premium in tenge	2 000 000 000
Number of insured events	831
Monthly Calculation Index (MCI) in tenge [29]	2 525
Total insurance loss in tenge	1 013 105 750
Total loss in MCI	401 230

Further, let us suppose that the above insurance losses have a distribution that is described according to the following table (Tab. 2 - Distribution of insurance losses):

Table 2. Distribution of insurance losses. Insurance losses

Insurance losses in MCI		Average compensation in MCI (average severity of losses)	Frequency	Number of insured events
0	40	20	0,0000	1
40	80	60	0,0002	24
80	120	100	0,0004	35
120	160	140	0,0003	31
160	200	180	0,0002	24
200	240	220	0,0004	36
240	280	260	0,0009	88
280	320	300	0,0003	31
320	360	340	0,0003	33
360	400	380	0,0005	52
400	440	420	0,0003	34
440	480	460	0,0005	52
480	520	500	0,0005	52
520	560	540	0,0006	56
560	600	580	0,0011	113
600	680	640	0,0001	11
680	760	720	0,0002	17
760	840	800	0,0002	19
840	920	880	0,0005	45
920	1000	960	0,0006	62
1000	1100	1050	0,0000	0
1100	1200	1150	0,0000	1
1200	1300	1250	0,0000	2
1300	1400	1350	0,0000	4
1400	1500	1450	0,0000	1
1500	1600	1550	0,0000	1
1600	1700	1650	0,0000	2
1700	1800	1750	0,0000	1
1800	1900	1850	0,0000	0
1900	2000	1950	0,0000	3

For the convenience of the study, the data on insurance losses in Table 2 are divided into groups in multiples of MCI with an average compensation (average severity of losses) equal to the average value of insurance losses per group. Also, the frequency indicator presented in Table 2 is determined according to the following formula:

$$\text{Frequency} = \frac{\text{number of insured events}}{\text{number of insured persons}} \tag{4}$$

Based on the above data, the risks of Compulsory Insurance of Civil Liability of Motor Vehicle Owners can be divided into 2 significant interdependent factors: the frequency and average severity of losses. These factors are used further in the k-means algorithm cluster analysis.

Moreover, in view of the fact that the k-means algorithm in the framework of cluster analysis requires estimates of the distances between clusters according to formula (2), it is necessary to specify a certain measurement scale when calculating distances. Since the factors chosen by us use completely different types of scales, the data must be standardized (normalized), so the values of each factor will lie in the segment $[0; 1]$.

Below there is a representation of the correspondence of the normalized frequency values to the normalized mean severity of losses in MCI as a graph (Fig. 1 - Compliance of the normalized frequency values with the values of normalized severity of losses in MCI.).

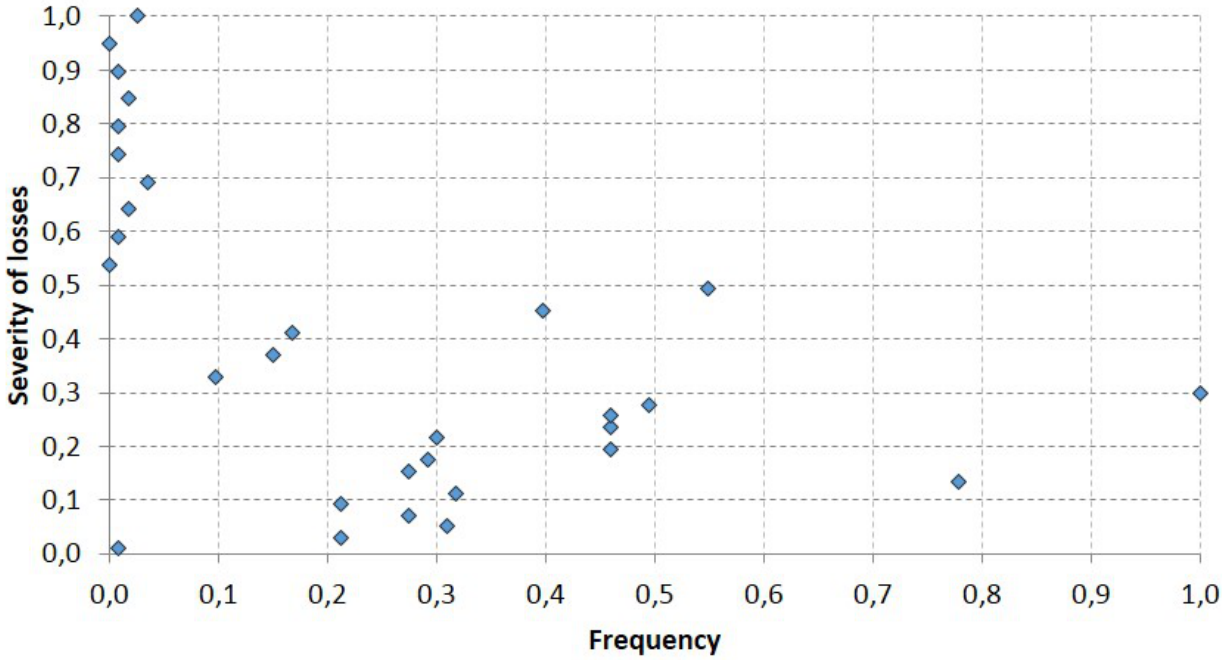

Figure 1: Correspondence of the normalized frequency values to the normalized values of the average severity of losses in MCI

Based on the visual presentation of the compliance results in Figure 1, it can be assumed that five natural clusters are formed. Having split the initial data into 5 clusters, we perform calculations according to the k-means algorithm and check the significance of the difference between the obtained clusters.

Let us set the values of the initial coordinates of the centroids in random order. In this case, the initial coordinates of the centroids are the diagonal points of the coordinate axis $[0; 1] \times [0; 1]$ (Tab. 3 - The initial coordinates of the centroids in random order):

Table 3. Initial coordinates of centroids in a random order

	Normalized frequency of losses	Normalized severity of losses
Centroids	x	y
C1	0,1000	0,1000
C2	0,2000	0,2000
C3	0,3000	0,3000
C4	0,4000	0,4000
C5	0,5000	0,5000

Further, in the framework of iteration No. 1, we will calculate the distances according to the Euclidean distances formula using the selected initial coordinates of the centroids. The results of calculations in the framework of iteration No. 1 are presented below (Tab.4 - Results of calculations in the framework of iteration No. 1):

Table 4. Results of calculations in the framework of the iteration No. 1

	x	y	Distance from C1	Distance from C2	Distance from C3	Distance from C4	Distance from C5	Cluster
1	0,0088	0,0103	0,1279	0,2693	0,4108	0,5522	0,6936	1
2	0,2124	0,0308	0,1320	0,1697	0,2831	0,4142	0,5504	1
3	0,3097	0,0513	0,2153	0,1848	0,2489	0,3602	0,4874	2
4	0,2743	0,0718	0,1766	0,1482	0,2296	0,3514	0,4840	2
5	0,2124	0,0923	0,1127	0,1084	0,2254	0,3604	0,4989	2
6	0,3186	0,1128	0,2190	0,1472	0,1881	0,2985	0,4276	2
7	0,7788	0,1333	0,6796	0,5826	0,5069	0,4632	0,4606	5
8	0,2743	0,1538	0,1825	0,0875	0,1484	0,2764	0,4132	2
9	0,2920	0,1744	0,2059	0,0955	0,1259	0,2501	0,3864	2
10	0,4602	0,1949	0,3725	0,2602	0,1916	0,2138	0,3077	3
11	0,3009	0,2154	0,2317	0,1021	0,0846	0,2095	0,3474	3
12	0,4602	0,2359	0,3850	0,2626	0,1725	0,1748	0,2671	3
13	0,4602	0,2564	0,3927	0,2662	0,1660	0,1557	0,2468	4
14	0,4956	0,2769	0,4333	0,3054	0,1969	0,1558	0,2231	4
15	1,0000	0,2974	0,9214	0,8059	0,7000	0,6087	0,5395	5
16	0,0973	0,3282	0,2282	0,1642	0,2046	0,3111	0,4378	2
17	0,1504	0,3692	0,2739	0,1763	0,1648	0,2514	0,3732	3
18	0,1681	0,4103	0,3177	0,2127	0,1719	0,2321	0,3438	3
19	0,3982	0,4513	0,4608	0,3201	0,1804	0,0513	0,1128	4
20	0,5487	0,4923	0,5960	0,4550	0,3144	0,1750	0,0493	5
21	0,0000	0,5385	0,4497	0,3931	0,3832	0,4233	0,5015	3
22	0,0088	0,5897	0,4982	0,4341	0,4108	0,4347	0,4993	3
23	0,0177	0,6410	0,5472	0,4772	0,4427	0,4519	0,5025	3
24	0,0354	0,6923	0,5958	0,5191	0,4732	0,4673	0,5028	4

25	0,0088	0,7436	0,6500	0,5762	0,5306	0,5206	0,5482	4
26	0,0088	0,7949	0,7008	0,6248	0,5742	0,5558	0,5729	4
27	0,0177	0,8462	0,7507	0,6714	0,6148	0,5875	0,5937	4
28	0,0088	0,8974	0,8026	0,7232	0,6646	0,6328	0,6318	5
29	0,0000	0,9487	0,8546	0,7750	0,7147	0,6790	0,6718	5
30	0,0265	1,0000	0,9030	0,8186	0,7515	0,7067	0,6886	5

The affiliation of a particular point to a particular cluster in the Table 4 is determined on the basis of the minimum distance between the point and the centroid:

$$\min(\text{Distance to}C_1; \text{Distance to}C_2; \text{Distance to}C_3; \text{Distance to}C_4; \text{Distance to}C_5) \quad (5)$$

Now let us find the new coordinates of the centroids for each cluster. For a particular cluster, they are defined as the average value of the coordinates of points (abscissas and ordinates) located in this cluster (Tab. 5 - Recalculated coordinates of the centroids after iteration No. 1).

Table 5. Recalculated centroid coordinates after iteration No. 1

	Normalized frequency of losses	Normalized severity of losses
Centroids	x	y
C1	0,1106	0,0205
C2	0,2541	0,1407
C3	0,1958	0,3994
C4	0,2035	0,5802
C5	0,3938	0,6282

Further, continuing the process of iteration, we come to the fact that by iteration No. 8 the values of the coordinates of the centroids cease to change. Below there are the tabular values (Tab. 6 - Recalculated coordinates of centroids after iteration No. 8, after which the coordinates of the centroids do not change) and a graph of the results (Fig. 2 - Results of cluster analysis after iteration No. 8 in accordance with the belonging of points to one of 5 clusters):

Table 6. The recalculated coordinates of the centroids after iteration No. 8, after which the values of the coordinates of the centroids do not change

	Normalized frequency of losses	Normalized severity of losses
Centroids	x	y
C1	0,2448	0,1014
C2	0,4705	0,3179
C3	0,0850	0,4472
C4	0,0155	0,8205
C5	0,8894	0,2154

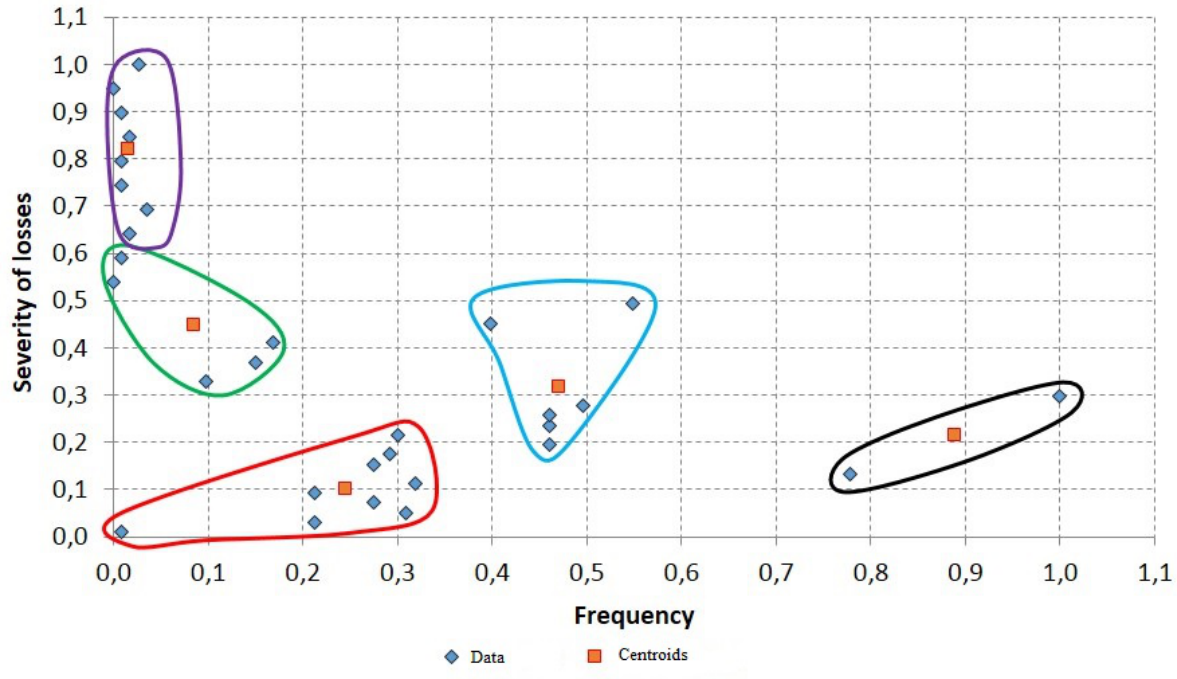

Figure 2: The results of cluster analysis after iteration No. 8 in accordance with the belonging of points to one of 5 clusters

We carried out identical calculations according to the k-means algorithm based on 2, 3, 4, 6, 7 and 8 clusters. The results are presented below (Fig. 3-8):

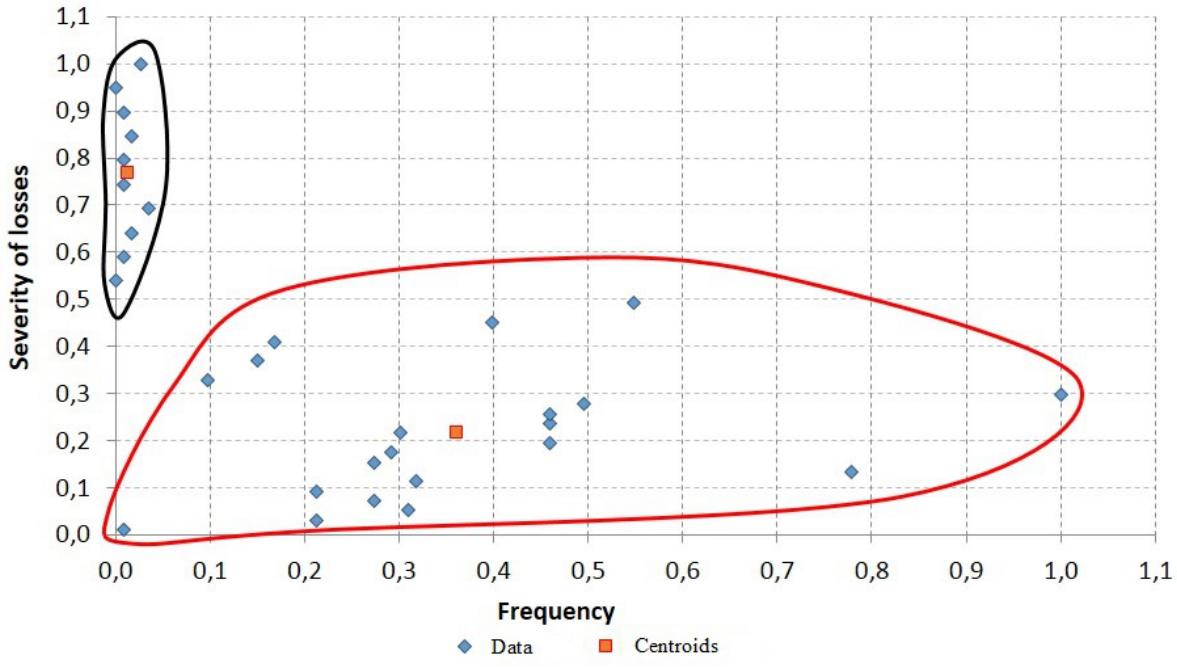

Figure 3: The results of cluster analysis in accordance with the affiliation of points to one of 2 clusters

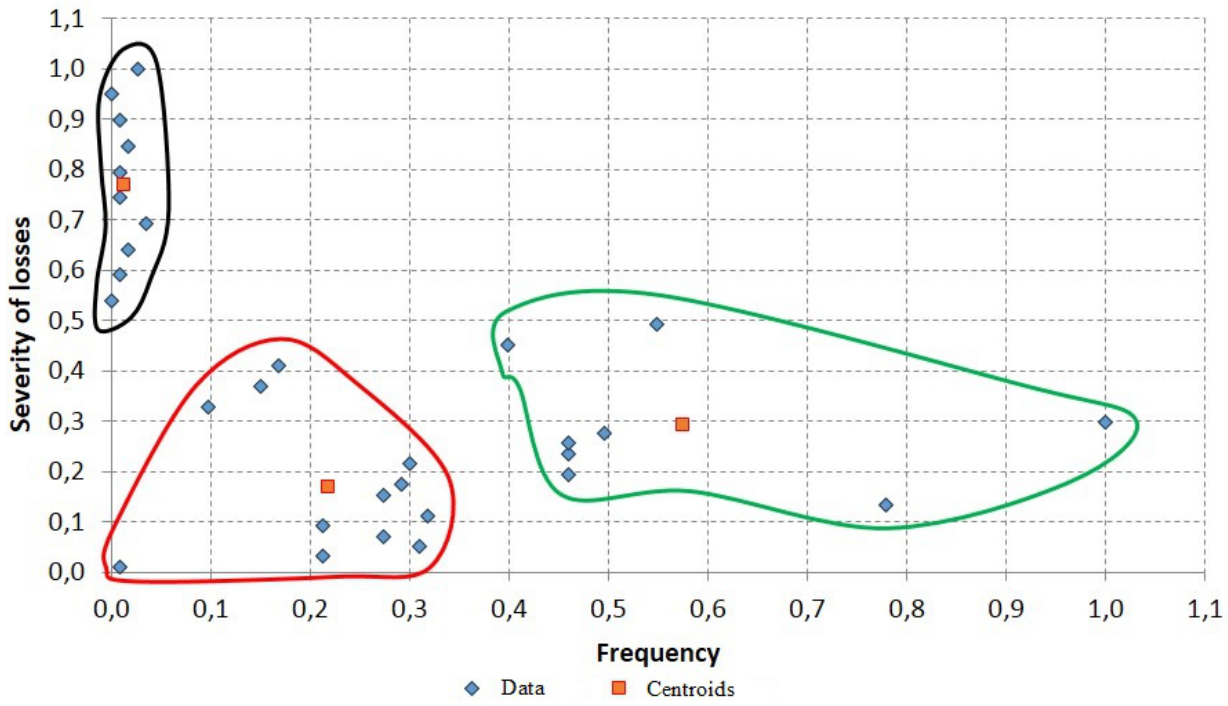

Figure 4: The results of cluster analysis in accordance with the affiliation of points to one of 3 clusters

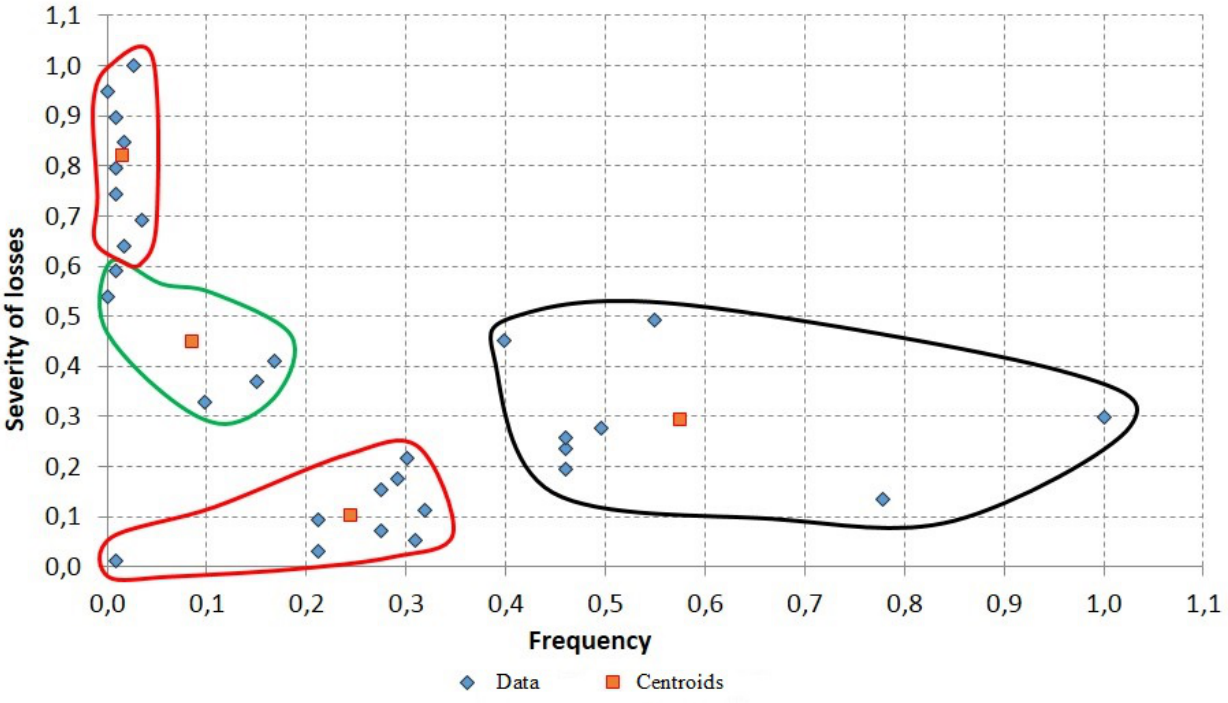

Figure 5: The results of cluster analysis in accordance with the affiliation of points to one of 4 clusters

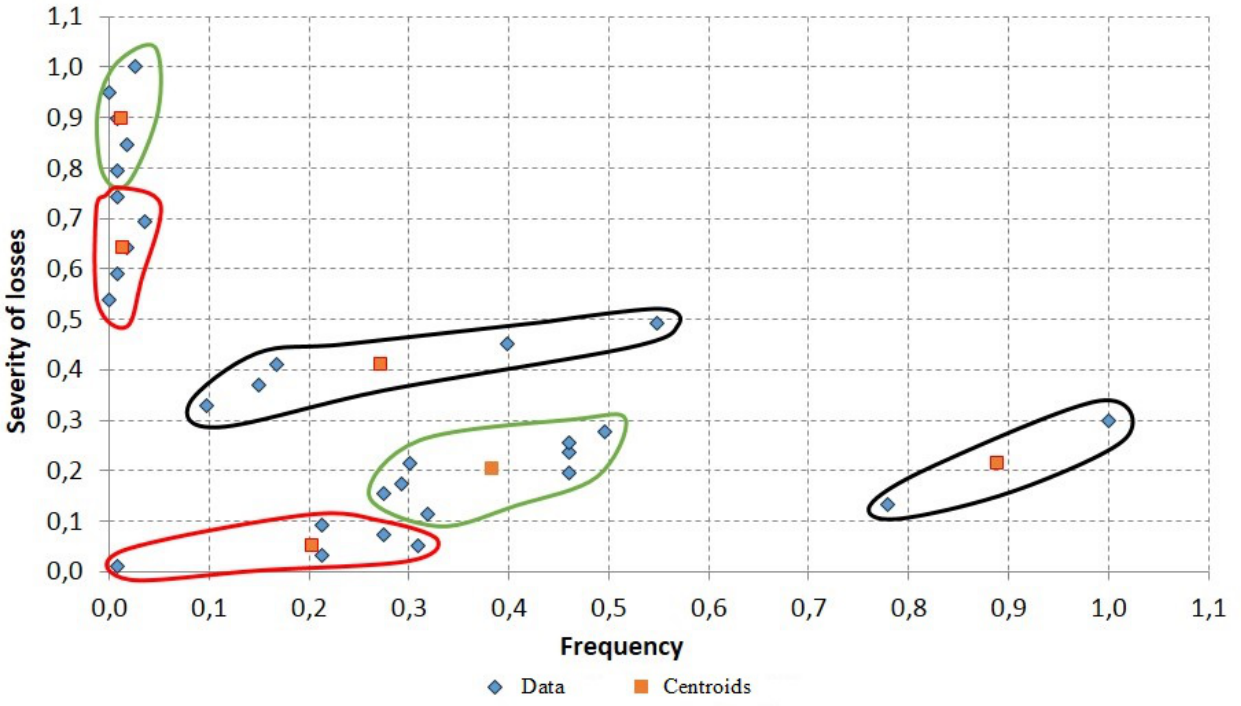

Figure 6: The results of cluster analysis in accordance with the affiliation of points to one of 6 clusters

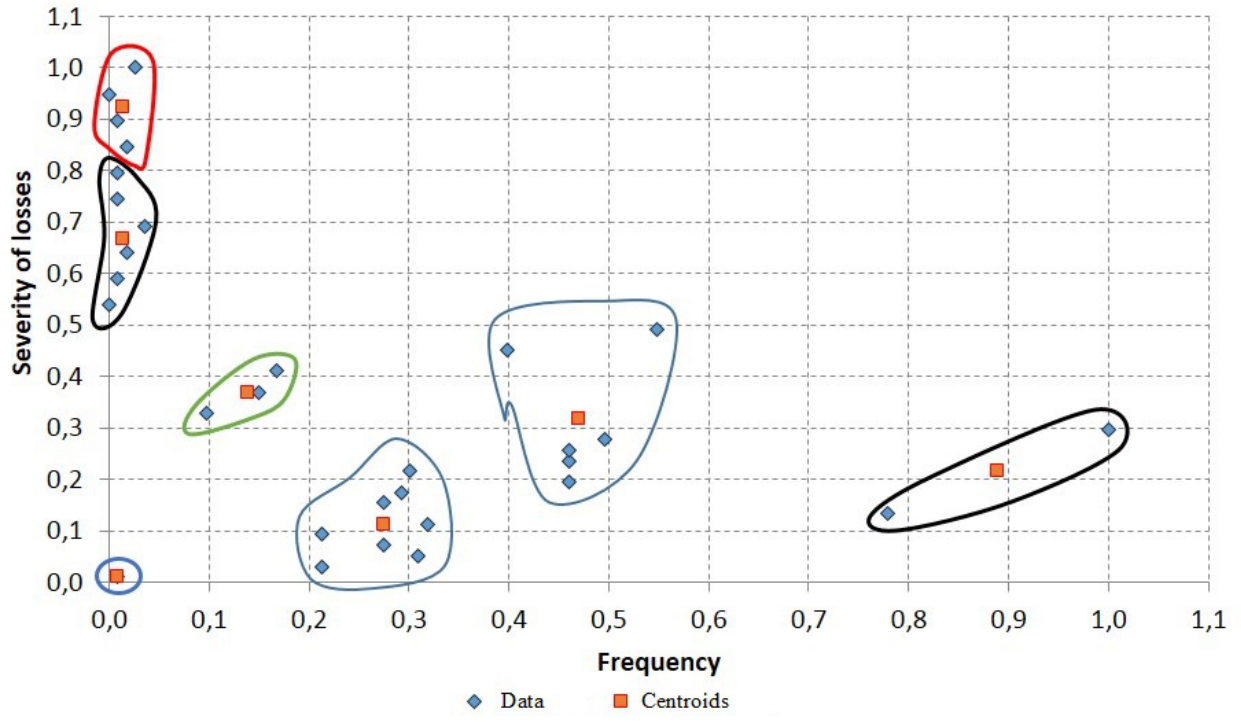

Figure 7: The results of cluster analysis in accordance with the affiliation of points to one of 7 clusters

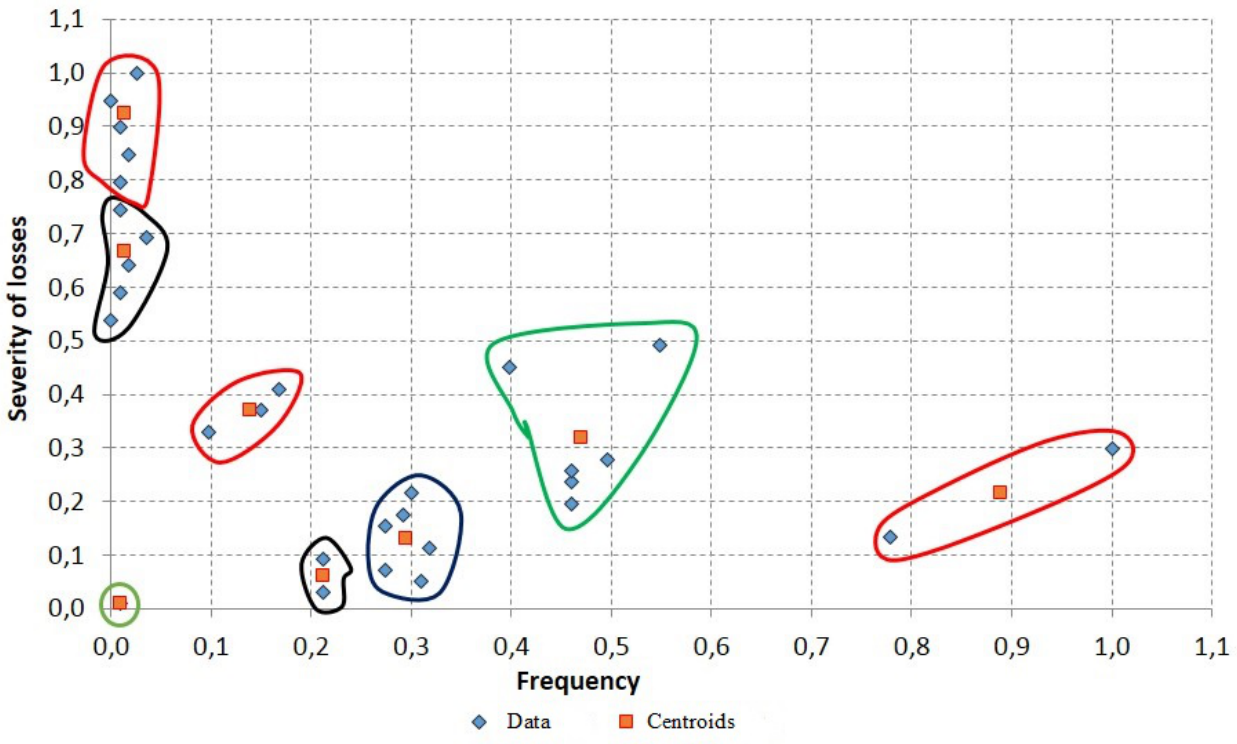

Figure 8: The results of cluster analysis in accordance with the affiliation of points to one of 8 clusters

In order to determine the optimal number of clusters, let us check the significance of the difference between the results obtained above. To do this, we use the criterion of deviation of the expectation of insurance losses $E[x]$, belonging to a specific cluster group, from the actual expected value of insurance losses in the amount of 401,230 MCI according to the data below (Tab. 7 - Deviation of the expected value of insurance losses $E[x]$ from the actual expected value of insurance losses):

Table 7. Deviation of the expected value of insurance losses $E[x]$ from the actual expected value of insurance losses

Number of clusters	Number of necessary iterations	Expected value of insurance losses $E[x]$	Deviation in %
2	7	369 300	-8,0%
3	5	416 087	3,7%
4	5	409 903	2,2%
5	8	395 703	-1,4%
6	5	378 040	-5,8%
7	5	393 120	-2,0%
8	6	394 987	-1,6%

5 Results and discussion

As we see, the use of the number of clusters below 5 in the calculations leads to an increase in the deviations. Also, splitting data into 6 or more clusters does not increase the accuracy of the estimates. Thus, the most optimal number of clusters in the above distribution of insurance losses in terms of accuracy and speed of data processing can be considered a quantity of 5.

So, after analyzing the results of the k-means algorithm with regard to 5 clusters, it can be noted that in each of the five clusters there are data with a similar effect on the loss process. Let us select the following distinguishing features of clusters:

1. The first cluster consists of 9 observations and includes insurance losses with an average frequency and low severity of 200 MCI. The first cluster is characterized by the lowest risk for the insurance company in a road traffic accident;
2. The second cluster consists of 6 observations and includes insurance losses with a frequency and severity above the average;
3. The third cluster consists of 5 observations and includes insurance losses with a frequency below the average and a severity above the average;
4. The fourth cluster consists of 8 observations and includes insurance losses with very low frequency and a very high severity of 1,600 MCI. The fourth cluster is characterized, first of all, by the greatest risk for the insurance organization in a road traffic accident;
5. The fifth cluster consists of 2 observations and includes insurance losses with a high frequency and severity below the average.

Thus, within the framework of the distribution of insurance losses, we can expect a decrease in the severity of losses from 1,600 MCI to 200 MCI as the frequency increases. Then, with a further increase in frequency, in general, we can expect an increase in the severity of losses.

In the future, data on the frequency and severity of losses can be combined with additional data, such as gender, age, driving experience of vehicle owners, in order to highlight target segments on Compulsory Insurance of Civil Liability of Motor Vehicle Owners in accordance with the tariff rates [30].

6 Conclusion

In this article, as a modern research approach to the qualitative underwriting analysis of insurance losses in the class of Compulsory Insurance of Civil Liability of Motor Vehicle Owners, it is proposed to use the k-means algorithm cluster analysis procedure, which allows to simplify the processing and further analysis of data by arranging it in a relatively homogeneous groups. In the framework of the proposed approach, losses for this class of insurance were divided into homogeneous qualitative groups (clusters) based on the frequency and severity of losses. As a result, calculations were made taking into account the optimally selected number of clusters, and clusters with similar effects on the process of losses were identified and interpreted. In the future, the data can be supplemented in order to highlight the target segments for Compulsory Insurance of Civil Liability of Motor Vehicle Owners in accordance with the tariff rates.

References

[1] The Law of the Republic of Kazakhstan "On compulsory insurance of civil liability of vehicle owners July 1, 2003, No. 446-II.

[2] "Current state of the insurance sector of the Republic of Kazakhstan National Bank of Kazakhstan, accessed May 15, 2019, https://nationalbank.kz/cont/%D0%A2%D0%A1%2001.04.2019%20eng.pdf.

[3] Resolution of Board of the Agency of the Republic of Kazakhstan on regulation and supervision of the financial market and the financial organizations, March 25, 2006, No. 85.

[4] Resolution of Board of National Bank of the Republic of Kazakhstan, December 26, 2016, No. 304.

[5] "Big Data v Kazahstne: O krupnom zakazchike, kadrah i perspektivah"[Big Data in Kazakhstan: On a large customer, personnel and prospects], accessed May 15, 2019, https://kapital.kz/tehnology/71257/big-data-v-kazahstane-o-krupnom-zakazchike-kadrah-i-perspektivah.html.

[6] Cherezov D.S., Tyukachev N.A., "Obzor osnovnyh metodov klassifikacii i klasterizacii dannyh [Overview of basic data classification and clustering methods]" , *The Bulletin of Voronezh State University* 2 (2009): 27.

[7] Atapina N.V., "Upravlenie processom anderrajtinga v imushestvennom strahovanii [Property Insurance Underwriting Management]" , *Molodoj uchenyj* 1 (2011): 84-87.

[8] Octaviani D, "Portfolio rule-based clustering at automobile insurance in Portugal" , *Internship report presented as partial requirement for obtaining the Master's degree in statistics and information management proposal, NOVA Information Management School* (2016).

[9] Berry M.J.A. and Linoff G.S., *Data Mining Techniques: for Marketing, Sales and Customer Relationship Management* (United States of America: Wiley Publishing, 2014), 1150.

[10] Kaufman L. and Rousseeuw P.J., *Finding groups in data: an introduction to cluster analysis* (United States of America: Wiley-Interscience, 2009), 3.

[11] Brito P.Q., Soares C., Almeida S., Monte A. and Byvoet M., "Customer segmentation in a large Data base of an Online customized fashion business" , *Robotics and Computer-Integrated Manufacturing* 36 (2015): 93-100.

[12] Hasan M.S. and Duan Z.H., "Hierarchical k-means: a hybrid clustering algorithm and its application to study gene expression in lung adenocarcinoma" , *Emerging trends in computational biology, bioinformatics and systems biology* (2015): 51-67. Accessed May 10, 2019. doi:10.1016/B978-0-12-802508-6.00004-1.

[13] Han J. and Kamber M., *Data mining concepts and techniques* (United States of America: Morgan Kaufmann Publishers, 2012), 600-703.

[14] He Z., Xu X., Huang J.Z. and Deng S., "Mining class outliers: concepts, algorithms and applications in CRM" , *Expert Systems with Applications* 27 (2004): 681-697.

[15] Ali Ghorbani and Sara F., "Fraud detection in automobile insurance using a data mining based approach" , *International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)* 8(27) (2018): 3764-3771.

[16] Yi P., Gang K., Alan S., Zhengxin C., Deepak K., Yong S. and Peter K., "Application of clustering methods to health insurance fraud detection"paper presented in the *International conference on service systems and service management, Troyes, France* (2006).

[17] Thakur S.S. and Sing J.K., "Mining Customer's Data for Vehicle Insurance Prediction System using k-Means Clustering-An Application" , *International Journal of Computer Applications in Engineering Sciences* 3(4) (2013): 148-153.

[18] Kaveh K-D., Farshid A. and Shaghayegh A., "Insurance customer segmentation using clustering approach" , *International Journal of Knowledge Engineering and Data Mining* (2016). Accessed May 09, 2019. doi:10.1504/IJKEDM.2016.082072.

[19] Ai C.Y., Kate A.S., Robert J.W. and Malcolm B., "Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry" , *Intelligent systems in accounting, finance and management* 10 (1) (2001): 39-50.

[20] Everitt B.S., Landau S. and Leese M., *Cluster Analysis* (London: Arnold, 2001): 260.

[21] Mirkin B., "Choosing the number of clusters" , *WIRE Data Mining and Knowledge Discovery* 3 (2011): 252-260.

[22] Romesburg C.H., *Cluster Analysis for Researchers* (North Carolina: Lifetime Learning Applications, Belmont, Ca. Reproduced by Lulu Press, 2004), 15-334.

[23] Mirkin B., *Clustering for Data Mining: a data recovery approach* (United States of America: Chapman & Hall/CRC, 2012), 93-137.

[24] Pang-Ning T., Michael S. and Vipin K., *Introduction to Data Mining* (United States of America: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, 2005), 125-147.

[25] Pérez-Ortega J., Almanza-Ortega N.N and Romero D., "Balancing effort and benefit of K-means clustering algorithms in Big Data realms" , *PLoS One* 13(9) (2018). Accessed May 07, 2019. doi:10.1371/journal.pone.0201874.

[26] Madhulatha T.S., "An overview on clustering methods" , *IOSR Journal of Engineering* 2(4) (2012): 719-725.

[27] Ghoreyshi S. and Hosseinkhani J., "Developing a clustering model based on k-means algorithm in order to creating different policies for policyholders in insurance industry" , *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* 4(2) (2015): 46-53.

[28] Jain A., Murty M. and Flynn P., "Data clustering: A review" , *ACM Computing Surveys* Vol. 31, No. 3 (1999): 264-323.

[29] The Law of Republic of Kazakhstan "On the republican budget for 2018-2020".

[30] Izakova N.B. and Kapustina L.M., "Primenenie metodov klasternogo analiza dlya segmentirovaniya promyshlennyh rynkov [Application of cluster analysis for segmentation of industrial markets]" , *Vestnik of Samara State University of Economics* 9(131) (2015): 100-107.