IDENTIFYING INSURANCE COMPANIES' BUSINESS MODELS IN UKRAINE: CLUSTER ANALYSIS AND MACHINE LEARNING

OLEKSANDR TARNAVSKYI[ab], VIKTOR KOLOMIIETS[a]

[a]National Bank of Ukraine
[b]National University of Kyiv-Mohyla Academy
E-mail: Oleksandr.Tarnavskyi@bank.gov.ua
Viktor.Kolomiiets@bank.gov.ua

Abstract This study examines the performance of the nonlife insurance companies that operated in Ukraine in 2019–2020. Specifically, we employ a set of clustering techniques, e.g. the classic k-means algorithm and Kohonen self-organizing maps, to investigate the characteristics of the Retail, Corporate, Universal (represented by two clusters), and Reinsurance business models. The clustering is validated with classic indicators and a migration ratio, which ensures the stability of the clusters over time. We analyze the migration of companies between the identified clusters (changes in business model) during the research period and find significant migration between the Reinsurance and Corporate models, and within the Universal model. Analysis of the data on the termination of the insurers' ongoing activity allows us to conclude that companies following the Universal business model appear to be the most financially stable, while their peers grouped into the Reinsurance cluster are likely to be the least stable. The findings of this research will be valuable for insurance supervision and have considerable policy implications.

JEL Codes G22, D22

Keywords cluster analysis, neural networks, business model, insurance

1. INTRODUCTION

Starting from 1 July 2020, the NBU began implementing the reform of the financial sector, extending the requirements of transparency, reliability, and efficiency to the nonbanking financial sector. The primary goal of the reform is to improve the quality of insurance services and protect policyholders' interests.

Effective supervision, control and implementation of reforms on the nonbanking financial market require an understanding of the market structure and how its participants conduct their business. For example, different business models may have quite different risk profiles. The identification of homogeneous groups of companies with similar risks allows for a more detailed analysis of the stability and solvency of insurance companies, and the effective prediction of crisis events. This research aims to contribute to the understanding of the Ukrainian insurance market's structure as well as its companies' operational and risk profiles. It identifies Ukrainian insurers' business models and their key features using quantitative indicators to assist supervision of the insurance market.

To achieve our goal, we attempted to answer these questions: Can homogenous and stable groups (business models) of insurance companies be identified through the analysis of regulatory data? What are the key characteristics of these business models? How did companies change their business models during the research period? Can certain business models be associated with increased risks?

For this paper, we conducted a cluster analysis of the Ukrainian insurance market to determine the business models used by insurers. We apply a number of clustering methods, including hierarchical, nonhierarchical, and machine-learning ones. We identify five clusters with the k-means method that correspond to four business models – Corporate, Retail, Universal (divided into two clusters), and Reinsurance. Before applying clustering algorithms, an artificial cluster named "Inactive" was formed (it comprised companies that were not very active or did not engage in

 Available at https://doi.org/10.26531/vnbu2021.252.02

insurance activities at all, but had a license and were present in the sample). A number of calculated coefficients, namely the migration ratio and the silhouette coefficient helped us assess the quality of our research.

We analyze the business models by using both the features by which the clustering took place and additional variables that are not used in the clustering algorithms. Thus, the companies with the Corporate business model mostly ensure legal entities, while those using the Retail model, on the contrary, work with individuals. Companies with the Universal business model tend to use sales offices widely as a sales channel, while those with the Reinsurance business model do not use them at all. The further text provides a more detailed description of the clusters.

Next, we show how insurers migrated between the clusters in the period from 2019 to 2020. We observed significant migration between the Reinsurance and Corporate business models and within the Universal business model. We also find significant migrations to the artificially created Inactive cluster, i.e. in cases when insurers terminated their insurance activity. Based on these migrations, it is possible to empirically draw conclusions about the riskiness of a particular business model. Thus, the largest share of the companies that left the market during the studied period belonged to the cluster using the Reinsurance business model; more than half of the companies in this cluster ceased operations in 2020. Significant migration to the Inactive group was also observed in relation to entities using the Corporate and Retail business models.

The paper is structured in the following way. The second section provides an overview of the relevant literature. The third section highlights the methodology, data, and software used in this analysis. The fourth section presents the key findings of the research and shows the riskiness of each of the identified business models. The fifth section briefly summarizes the results of the research and outlines promising directions for future research.

2. LITERATURE REVIEW

The development of research on banks' business models was facilitated by the Basel regulatory framework and the implementation of the Supervisory Review and Evaluation Process (SREP). Studying insurance companies' business models are also seen as a promising area of research.

Most of the existing work in the area is aimed at the analysis and segmentation of insurers' clients. Research by Wang and Keogh (2008), and Zaqueu (2019) is devoted to a clustering analysis for target group identification. Clustering techniques were used to identify customer profiles based on datasets derived from policy transactions and policyholder information. Wang and Keogh used self-organizing maps (SOMs) and the k-means algorithm. The k-means method was also used in a publication by Abolmakarem et al. (2016) that used segmentation to identify the most profitable customers for companies. Velykoivanenko and Beschastna (2018) use SOMs to rate insurance companies in terms of their financial performance into three groups. Then they combine clustering results with experts' ratings to arrive at integral indicator of company financial stability.

Most researchers use a combination of the two clustering methods. In particular, a study by Kramarić et al. (2018) groups European insurance companies into seven clusters. Unlike previous studies, her research groups companies, rather than their customers, into clusters. Using a combination of hierarchical clustering and k-means clustering, 119 insurers are divided into seven groups by country of origin and company type. Bach et al. (2020) use a Kohonen map in combination with a hierarchical cluster analysis to investigate fraud risks in the leasing industry. The neurons of the Kohonen self-organizing map are combined into five clusters using Ward's method, after which the risk characteristics of the clusters are analyzed.

A study by Ahmar et al. (2018) is an example of a cluster analysis outside of the financial sector. In their study, they use the k-means method to group the provinces of Indonesia. Such a grouping, according to the authors, should help to classify the regions of that country so that social problems can be tackled more effectively. Abbas et al. (2020) compared the k-means and k-medoids methods using data on women during pregnancy. The k-medoids method is inherently very similar to the k-means method, so they are often used in combination. The authors show that the k-medoids method is more accurate than k-means for specific data.

Rashkovan and Pokidin (2016) identified business models of banks in Ukraine using a Kohonen self-organizing map, and drew parallels between business models and indicators of various types of risk to which a bank may be exposed. In terms of methodology, our work is very similar to this study. However, unlike Rashkovan and Pokidin, who base their research findings on a Kohonen self-organizing map, we use this method in addition to the k-means method. Ferstl and Seres (2012) also used a cluster analysis to identify business models. Unlike previous researchers, they utilized the k-means algorithm based on the use of the Mahalanobis distance. Their work identifies five business models of banks, based on five indicators.

Most authors use the simplest clustering models, including the k-means method. In our work, we intend to develop a methodology that helps to determine the distribution of companies according to their business model. To implement the research, we used a wide set of clustering tools, but the conclusions were based on the k-means method. Kohonen self-organizing maps are a convenient visualization tool in our work. This research for the first time evaluates the quality of clustering through the use of the migration ratio – an indicator that characterizes the stability of clusters.

3. DATA AND METHODOLOGY

Description of the Data Used

To conduct a cluster analysis of insurance companies, we gathered data from the regulatory reports of 247 Ukrainian insurers for two years, from 2019 to 2020. During the research period, the number of active insurers decreased significantly. Thus, as of the end of 2020, the database consisted of entries for 185 insurers. The data were taken from a regulatory database.

To identify a business model, we aimed to select indicators that would answer the following questions about an insurance company:

Who are its target customers?

What types of insurance does it focus on, and how explicitly?

What sales channels does it use?

Table 1. Indicators Used for Clustering

No.	Indicator	Variable name	Formula
1	Return on assets	ROA	Net income / Total assets
2	Number of offices	Offices	Total number of used offices that are not the head office
3	Share of premiums from mandatory types of insurance in the total amount of collected premiums	% of mandatory	Amount of premiums from mandatory types of insurance / Total amount of premiums
4	Share of premiums from legal entities	Corporate	Amount of premiums from legal entities / Total amount of premiums
5	Share of inwards (assumed) reinsurance in premiums	Re-to-premiums	Amount of reinsurance premiums / Total amount of premiums

We selected indicators that would simultaneously help to find the answers to these questions, and which would allow the insurers to be optimally sorted into clusters according to certain quantitative metrics – the ratio of migration and the silhouette coefficient (as described further in the text). According to the values of the metrics obtained, an optimal set of indicators for clustering was selected. Next, a different set of indicators that allowed for a broader description of the clusters and the risks inherent in them was chosen separately. These indicators were not included in the model, as partitioning based on them led to worsened clustering quality. Rather, they were used for the broader characterization of the identified clusters. Table 1 and Table 2 describe these two groups of indicators.

After calculating the indicators, their values were standardized (to mean 0 and unit variance). This is necessary because of the clustering algorithms' sensitivity to variance in the data. We also detected outliers in the data. Observations that were more than three standard deviations away from the mean were rounded to the nearest value within the range of three standard deviations. The distribution of the observations before and after this procedure is given in Figure B.2.

Companies whose total premiums for the reporting period did not exceed UAH 5 million were grouped into an artificial cluster (group) named "Inactive." Such companies in 2020 accounted for less than 1% of total market share (in premiums).

Further are the descriptive statistics of the data for 2020.

We can see that most companies in 2020 were slightly profitable or unprofitable, in contrast to the higher levels of profitability observed in previous years (Table A.1). The reason for such a drop in profitability could be attributed to an increase in health insurance claims as an effect of the COVID-19 pandemic. Indeed, loss reserves for health insurance increased significantly in the periods that there were peaks in new COVID-19 cases.

Most of the market focuses its activities on low-priced contracts, which are most likely to be sold to individuals. The median "average check" is about UAH 3,000.

According to our data, most companies did not have sales offices. On the one hand, this may indicate the predominance in the market of business models that do not use offices as a sales channel. On the other hand, such a strong skew indicates a possible risk that some companies are misreporting. We are not able to verify this. We assume that any misreporting companies are evenly distributed across the clusters and do not significantly shift cluster centers. It is worth noting that a similar structure is observed for the data for 2019.

There is also a high concentration of one type of insurance on the market. About half of the companies had a share of premiums from one of the groups of insurance types that exceeded 60%. This indicates the presence of

Table 2. Indicators Used for Additional Description of Clusters

No.	Indicator	Variable name	Formula
1	Ratio of the share of reinsurers in insurance reserves to the total amount of insurance reserves	Re-to-provisions	Amount of reinsurance recoverables / Amount of insurance reserves
2	Loss ratio	Loss ratio	(Insurance claims paid + Expenses associated with claim settlement + change in loss reserves / (Premiums + change in unearned premium reserves)
3	Average size of premium collected	Mean premium	Amount of premiums collected / Number of insurance contracts
4	Ratio of wages to premiums collected	Wages/ Premiums	Amount of wage expenses / Amount of premiums collected for the reporting period
5	Maximum concentration on a group of types of insurance	Concentration	Maximum value of premiums among 7 categories* / Total amount of premiums

* List of categories: 1. Nuclear insurance; 2. Motor insurance (other); 3. Motor insurance; 4. Liability insurance; 5. Personal insurance (health, accident, pension insurance, etc.); 6. Property insurance; 7. Other.

specialization by companies in a certain insurance segment. Thus, the portfolios of many companies can be described as weakly diversified.

Research Methodology and Models

Clustering algorithms are a convenient tool for dividing observations into homogeneous groups based on given features. The literature review cites only some of the successful cases of using cluster analysis in the study of social and economic phenomena. An important advantage of such algorithms is to reduce the influence of a researcher's judgment about a phenomenon under study on the findings of the research. We use classic hierarchical and nonhierarchical clustering methods, along with machine learning methods, to study the business models of insurers. The following is a brief summary of the applied methods.

The k-means method is the most commonly used nonhierarchical method. It suggests iterative minimization of the distance between constituents of a cluster, while the number of clusters is set at the beginning – that is, the model does not determine the optimal number of clusters. The centroid coordinates, the number of which corresponds to the number of clusters, are set randomly at the initial stage. As a result, the division into clusters can be unstable and can depend on the initial centers.

We use the method of seeding the initial centers for k-means, called k-means++, proposed by Arthur and Vassilvitskii (2006) to avoid this problem. Denoting the input data sample $\mathcal{X}$, and the shortest distance between an element of the sample x_i and the closest center $D(x_i)$, the algorithm can be described stepwise:

1. Choose the initial center c_1 from $\mathcal{X}$ at random

2. Choose the next center c_j from $\mathcal{X}$, selecting each element with probability

$$p(c_j = x_i) = \frac{D(x_i)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$$

3. Repeat Step 2 until the required number of the centers has been chosen

4. Proceed with the classic k-means algorithm.

It can be seen from Step 2 that the elements $\mathcal{X}$ located farther from the initial center are selected with a higher probability. That is, the centers are located so that they are different from each other. Clusters based on the k-means++ procedure were evaluated 100 times to select the clustering with the minimum within-cluster sum of squares (WCSS).

In summary, the cluster centers are first chosen from the sample elements so as to be located farther from each other, then iteratively change their coordinates to describe the largest possible group (cluster) of the sample elements.

For the k-means method, it was decided to use five clusters, which is the optimal number of clusters with regard to the elbow method. The elbow method results are given in Figure B.2. A division into five clusters was used for all further methods.

The k-medoids method, first described by Kaufman and Rousseeuw (1990), is inherently very similar to the k-means method. The key characteristic of the k-medoids method is a partitioning technique of clustering to choose data points as centers. Such data points, which are exemplars for their cluster, are called medoids.

The Partitioning Around Medoids (PAM) algorithm is used to choose medoids iteratively in such a way as to reduce the average distance from data points to the centers of their clusters. Modern algorithms of the k-medoids method offer faster optimization, but PAM remains one of the most accurate algorithms for solving this problem. That is why we chose this method. We selected the initial medoids using the k-medoids++ algorithm, which is identical to k-means++ and ensures cluster stability.

Hierarchical methods do not require the number of clusters to be known before applying the algorithm. They build a tree-like structure called a dendrogram. First, each dataset forms a separate cluster. Further, datasets (clusters) based on the selected criterion are combined into new, larger clusters until they are all combined into one cluster, which includes all observations. Ward's method was chosen for our purpose. The number of clusters is determined by the researcher based on the dendrogram produced by applying the algorithm.

According to Ward's method, a separate cluster is combined with the cluster and their combination will lead to the smallest increase in the distance between data points within the cluster. This distance, which is similar to the WCSS metric of k-means, is displayed on the dendrogram along the vertical axis.

Table 3. Descriptive Statistics of the Models' Variables

	ROA	Offices	Re-to-premiums	% of mandatory premiums	Corporate
Mean	0.02	8.16	0.08	0.19	0.51
Std. Deviation	0.10	24.99	0.21	0.26	0.33
Min	-0.67	0.00	0.00	0.00	0.00
Q(25%)	0.00	0.00	0.00	0.00	0.22
Q(50%)	0.01	0.00	0.00	0.03	0.46
Q(75%)	0.05	0.00	0.03	0.35	0.86
Max	0.40	200.00	1.00	0.86	1.00

Table 4. Descriptive Statistics of Companies' Parameters

	Re-to-provisions	Loss ratio	Mean premium	Wages/ Premiums	Concen-tration
Mean	0.21	0.39	116.41	0.06	0.68
Std. Deviation	0.24	0.51	405.89	0.06	0.19
Min	0.00	-0.80	0.00	0.00	0.29
Q(25%)	0.02	0.11	0.89	0.02	0.54
Q(50%)	0.12	0.35	3.22	0.04	0.66
Q(75%)	0.36	0.52	29.79	0.07	0.83
Max	0.93	4.08	3,534.63	0.46	1.00

The following is a brief description of the Kohonen map, which is a machine learning method capable of clustering. It is described in Kohonen's work (1982). A self-organizing map is an artificial neural network consisting of two layers:

1. Sample data are present in the input layer. The dimensionality of this layer corresponds to the number of features used to cluster datasets into distinct groups.

2. The output layer, which is actually a map consisting of neurons arranged in two (in the case of this study) dimensions and has predetermined arbitrary dimensionality.

All of the neurons on the grid are connected to all of the inputs, and these connections have strengths, or weights, associated with them. That is, each neuron has a set of weights that can be interpreted as a description of the neuron in the features of the data in the input layer. The learning algorithm of the Kohonen map can be described step-by-step:

1. The weights of neurons are initialized to sufficiently small random values.

2. The feature vector x_i from $\mathcal{X}$ is supplied to the input layer and the distance is calculated (this study uses the Euclidean distance) between the vectors x_i and w_j, where w_j is the vector of the weights of the neuron j in the output layer of the grid.

3. The neuron that is closest to x_i based on Step 2 is called the best matching unit (BMU).

4. Taking the radius $\sigma(t)$, the neighborhood parameter is computed for each neuron of the map based on the Gaussian function

$$N(t)_{BMU,j} = exp\left(-\frac{D(BMU,j)^2}{2\sigma(t)^2}\right),$$

where $D(BMU,j)$ is the topographic distance between the BMU and the neuron j.

5. The weights of the neurons on the map are updated according to the formula $\Delta w_j = \alpha(t)N_j(t)(x_i - w_j)$, where $\alpha(t)$ is the learning rate, which is a decreasing function of time.

6. Steps 2-5 are repeated for a given number of epochs (training cycles), as determined by the researcher. At the same time, it is customary to pay attention to the quantization error, which reflects the average distance between the input data and the BMU, and to the topographic error, which reflects the number of data samples for which the first BMU (BMU1) is not an adjacent neighbor of the second BMU (BMU2).

As a result of training, the neurons become "similar" to the input data. As training proceeds, the parameters $\alpha(t)$ and $\sigma(t)$ gradually decrease. Thus, the further the training progresses, the slower the neurons adapt their weights and the less "interaction" they demonstrate. The decreasing function used in this study to describe the dynamics of the parameter $\alpha(t)$ has the following formula:

$$\alpha(t) = \frac{\alpha(0)}{1 + (t / (MI / 4))} \quad (1)$$

where $\alpha(0)$ is the initial value of $\alpha(t)$ set by the researcher; MI is the maximum number of epochs (iterations) set by the researcher;
t is a sequence number of an epoch.

We set $\alpha(0)$ for this study at 0.5, while MI equals 10,000. Thus, the learning rate gradually decreases from 0.5 to 0.1. The dynamics of the parameter $\sigma(t)$ in the process of learning for the model are similar, with an initial value of 1. Honkela (1998) describes the self-organizing map algorithm in more detail.

Given the number of observations in a dataset, a 10x10 map (100 neurons in total) with a rectangular topology was chosen for this study. It is common to initialize neurons' weights based on the principal components observed in the data. However, given that neurons are activated (become BMUs) evenly on the map (Figure B.3) and the learning time is acceptable, we do not use this approach. The dynamics of the topographic error and the quantization error are presented in Figure B.4.

After training, the neurons were clustered by applying the k-means method to their weights in order to be able to compare the findings of the Kohonen map with those of other methods.

The described methods were implemented with Python tools using open-source machine learning libraries such as Scikit-learn (Pedregosa et al., 2011) and MiniSom (Vettigli, 2019).

Evaluation of Clustering Results

Each of the methods has its advantages and disadvantages: assessments of them are given in Table 5.

Table 5. Comparison of Clustering Methods

Features	k-means	k-medoids	Ward's method	Kohonen maps
Ease of interpretation of findings	+	+	+	-
Availability of graphic tools	-	-	+	+
Resistance to outliers	-	+-	+-	+-
Applicability of evaluated model to different datasets	+	+	-	+
Ease of use	+	+	+	-

As we see, none of the methods stands out as the best. Therefore, when applying cluster analysis, the method chosen is most often the one best suited to the available data and numerical criteria for clustering quality.

To evaluate the quality of the models, we used a classic indicator, the Calinski-Harabasz score (CH score), which evaluates the quality of clustering into groups based on the distances between observations. The stability of clusters over time is also important for our purposes. Business models reflect stable behavior (a strategy), and so in order to draw conclusions about business models and their risks, it is important that the clusters do not change significantly over time. To assess stability, we used the migration ratio.

The CH score was first described by Calinski and Harabasz (1974). Also known as the Variance Ratio Criterion, it is the ratio of the sum of between-cluster dispersion and of inter-cluster dispersion for all clusters, both weighted by their respective degrees of freedom. The indicator is calculated as follows:

$$CH\ score = \frac{\left(\frac{BCSS}{k-1}\right)}{\left(\frac{WCSS}{n-k}\right)} \quad (2)$$

where $BCSS$ is the variance between clusters;
$WCSS$ is the variance between datasets within clusters;
k is the number of clusters;
and n is the number of datasets.

There is no critical value for this indicator. However, a larger value indicates a more definite grouping into clusters. The value of the CH score is larger when the centers of the clusters are farther from each other, and the datasets in the clusters are close to their centers.

High-quality clustering forms groups (clusters) that do not change significantly over time. In our example, this is fundamentally important because a business model is a stable feature of a company that does not change significantly under normal operating conditions. To assess the stability of clusters, the migration ratio was calculated:

$$\text{Migration ratio} = \frac{n_m}{n_{2019 \cap 2020}} \quad (3)$$

where n_m is the number of companies that, based on the model, migrated between clusters from 2019 through 2020. $n_{2019 \cap 2020}$ is the number of companies that were active in 2019 and 2020.

Migration between clusters occurs as a result of two factors – a change in a company's business model and clustering errors. Therefore, an overly large value of the migration ratio indicates inaccurate clustering, and an overly small value hints that a model is "overfit." There is no critical value of this indicator.

A pseudo migration ratio was calculated for Ward's method. Since the estimated model cannot be applied to data from another year, we calculated the pseudo migration ratio. To do this, the model was evaluated on the basis of the most recent data. Next, based on the centroid-based classification method described by Tibshirani et al. (2002), we identified clusters for data from the previous year and applied the formula (3).

Table 6 assesses the clustering quality for the applied models.

Table 6. Comparison of the Quality of Clustering Methods

Indicator	k-means	k-medoids	Ward's method	Kohonen maps
CH score	68.807	68.806	77.101	-
Migration ratio	15.8%	19.0%	20.6% (pseudo)	15.8% (between clusters), 76.9% (between neurons)

It can be seen that Ward's method gives the best clustering outcome according to the CH score criterion. However, the clusters are significantly less stable compared to all of the other methods. Given this criterion, which is of great importance from the point of view of the applicability of the model, we decided not to draw conclusions based on Ward's method. The findings of Ward's method are presented in Figure B.5 for reference. The clusters were named similarly to the main model for ease of comparison.

The k-means and k-medoids models have very close CH score values. Although these values are lower than those obtained by applying Ward's method, the difference is not very significant. The k-means model shows more stable clusters than the k-medoids model does. Also, in the presence of biased data, the k-medoids model may not fully characterize the clusters, as it bases its conclusions on a single observation. For example, this model characterizes four out of five clusters as business models that do not use offices in their activities at all. Such a conclusion is erroneous, as can be seen from the findings of the k-means model presented below. Therefore, due to this data distortion, we did not base our conclusions on the findings of the k-medoids model. The findings of the k-medoids method are presented in Table A.2 for reference.

Two types of migration ratio were calculated for the Kohonen map. The first is based on the five clusters into which the neurons are grouped. The second is based on all one hundred neurons of the model. As expected, the former is much smaller than the latter. It is interesting to observe that the migration between the clusters when applying the Kohonen map is almost identical to the case with the k-means method.

Given the findings of the quality assessment of the models, we decided to base our conclusions on the findings of the k-means model. Also, since the Kohonen map findings are similar to those of the k-means method, we used the map as a cluster visualization tool.

Next, we built a silhouette graph (Rousseeuw, 1987) for the k-means findings to evaluate the outcome in more detail. The silhouette coefficient is calculated for each observation as:

$$s_i = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (4)$$

where $a(i)$ is the average distance from the sample i to the members of its cluster; and
$b(i)$ is the average distance from the sample i to the members of the nearest neighboring cluster.

The value of the ratio for the model is the average value of the silhouette coefficient of all observations. A silhouette coefficient value of 1 indicates that clusters are clearly distinguished; 0 means that clusters are indifferent; and -1 means that clusters have been assigned wrongly.

The graph of silhouettes (Figure 1) shows that there is only one observation for which members of the neighboring cluster are "closer" on average than members of its own cluster. This observation relates to cluster 1 (Universal "Large" model). It, and those with a silhouette value close to 0, may be located "on the edge" of the cluster. The overall value of the ratio (0.41) indicates a sufficiently high-quality clustering; in addition, the graph shows that the Reinsurance business model (Cluster 4) is the best defined.

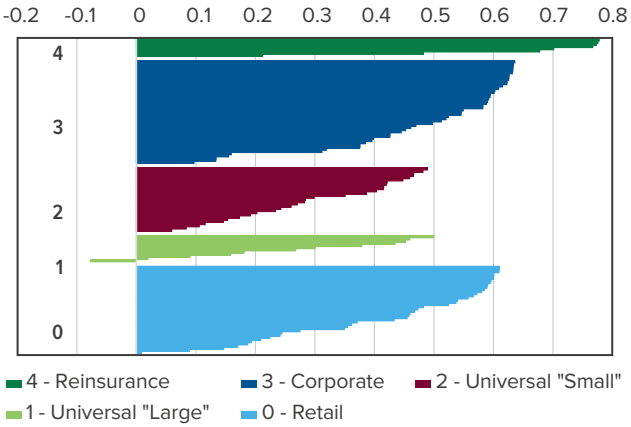


Figure 1. Cluster Silhouettes (the abscissa is a silhouette coefficient, the ordinate is a cluster number)

4. RESEARCH FINDINGS

Description of Business Models Based on Clusters

The model was evaluated on the basis of data for the year 2020 and applied to all years in the sample (2019-2020). The features (the coordinates of the centroids) of the identified clusters are shown in Figure 2. The coordinates in the figure are standardized.

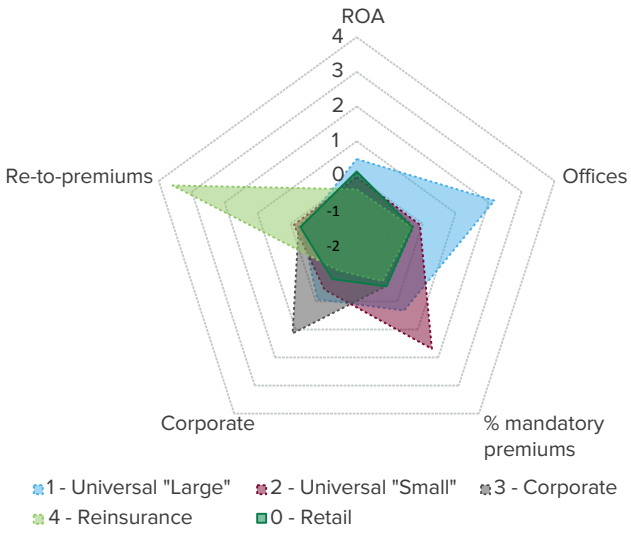


Figure 2. Features of Identified Clusters (standardized)

The identified clusters were numbered and named for convenience. We do not rule out that other homogeneous groups of companies may be identified from the data and may distort the features of the clusters we have identified. However, relying only on the data and the methods described above, we managed to identify those business models that are best separated in terms of quantitative criteria. Then we provide a brief overview of the identified clusters (business models) based on the features used for clustering and on descriptive indicators. Table A.3 summarizes the clustering findings in an unstandardized form.

Business model 0 – Retail, focuses on insuring individuals (who account for 74% of premiums) and has an average level of return on assets (3%). With a small number of offices and a significant ratio of wages to premiums collected, companies with this business model use their own agents as a channel to acquire customers. That is, the companies "hunt" for customers rather than customers themselves coming to their offices. Insurers using this business model offer mostly voluntary types of insurance, their average share of reinsurance in premiums is 1.5%. It is worth noting that this cluster has the highest concentration indicator (76%) on one of the insurance types, which can be a risk factor for the companies in this group. In 2020, the companies of this cluster posted the highest losses, and the share of reinsurance in their insurance reserves was moderate, which indicates the companies' vulnerability to underwriting risk. In 2020, this cluster included 40 companies, which accounted for 19% of the market by premiums.

Business model 1 – Universal "Large" insurers serve both legal entities and individuals and have a distribution between mandatory and voluntary insurance of 28% and 72%, respectively. Their return on assets is the highest among all selected clusters (6.5%). A characteristic feature of this cluster is the wide use of its own offices (they have about 62 offices on average). One sign that companies of this business model actively use both their own offices and agents as sales channels is the high share of wages in premiums collected (6%). Thus, these companies try to diversify their ways of acquiring customers. Companies in this cluster also have the second largest share of reinsured risk (25%), which means lower underwriting risk, as well as the fact that the companies run the risk of counterparty (reinsurer) default. Some 12 companies in 2020 used this business model and had the largest share of gross premiums, estimated at 35.5%. It is worth noting that given the high market share combined with the lowest average premium, these companies tend to sell low-priced policies on a large scale.

Business model 2 – Universal "Small", is characterized by a relatively even (compared to other models) distribution in premiums of mandatory and voluntary insurance and individuals and legal entities (64%/36% and 63%/37%, respectively). However, the share of premiums from mandatory types of insurance in this business model is the largest of all the clusters. Premiums from Motor Third Party Liability (MTPL) insurance account for 71% of the premiums from mandatory insurance types for this cluster. This cluster also has the second lowest rate of return on assets among all groups, and companies own an average of six offices and have the second highest share of reinsurance in premiums (5.6%). In addition to focusing on mandatory insurance, this model differs from Universal "Large" by a significant difference in the average premium, which could be evidence that these companies try to insure more expensive risks. In 2020, this cluster included 29 companies, which together accounted for about 16.5% of the market by premiums.

Business model 3 – Corporate, is characterized by an 89% share of legal entities in premiums, as well as a low rate of return on assets (2.7%) and a small number of offices, while its share of mandatory insurance is close to zero. For companies that do not use a reinsurance business model, their share of inwards (assumed) reinsurance in premiums is significant (27%). With a relatively high level of average premium (UAH 254,000), the companies of this cluster have a fairly low loss ratio compared to other business models (22.5%). A high share of a reinsurer in the insurance reserves is predictable, as such insurers often need to share a corporate client's large exposure. However, this creates the risk of counterparty (reinsurer) default for the companies in this cluster. This cluster encompasses the largest number of companies (47), which, based on the premiums collected in 2020, together account for 19% of the market.

Companies of business model 4 – Reinsurance, have an average share of reinsurance in premiums of about 81%. The return on assets of the companies of this cluster is negative on average, and the average premium is more than UAH 304,000. Companies in this cluster do not use offices as a sales channel at all and have the lowest share of wages in premiums. The share of voluntary insurance in premiums approaches 100%. Reinsurers themselves are weakly reinsured, which may indicate a potential vulnerability to underwriting risks that they do not share (diversify) among themselves. However, the low value of the loss ratio compared to other business models indicates that the underwriting risk may be insignificant. The cluster included eight companies according to 2020 data (10% of the market by premiums).

Histograms with the features of the grouped clusters are shown in Figure B.6.

Neurons on the Kohonen self-organizing map in the process of training become "similar" to the input data in terms of their weights, i.e. they reproduce clusters. The Kohonen self-organizing map provides a convenient tool for visualizing the similarities between different observations and the characteristics of those observations. With the help of the map, one can see the samples that lie on the border of the clusters and how far they are from other elements of the cluster.

The maps of the features in Figure 3 show a map's neuron weights correspond to data features (coordinates are standardized). They should be interpreted as follows: companies for which the neuron with coordinates (1;10) (upper left corner) after training is the BMU (the closest one) have an average market share of premiums from mandatory types of insurance, lots of offices, and almost no reinsurance in premiums.

Neighboring neurons are quite similar due to the mechanism of "cooperation" during learning, so they can be combined into groups. For this, the k-means algorithm with a number of clusters equal to five was used to interpret the results in a similar way as the results of the k-means division. We do not indicate the centroids of these clusters, since grouping by the k-means method was carried out only to mathematically estimate the boundaries of the clusters on the Kohonen map, and such centroids would essentially reproduce the centroids of the previous model.

Figure 4 shows the results of combining neurons into clusters. The dots in the figure indicate companies for which a particular neuron is the BMU after training. As one can see, considering the combination of neurons and feature maps, it is possible to identify business models that are similar to those identified by the k-means method.

However, the map allows us to see the distance (similarity) between the observations. The topographic distance can immediately be seen on the map – neighboring neurons are similar. The Euclidean distance between neurons after training can be seen in Figure B.7. Companies for which the BMU is located on the topographic boundary of a cluster are "weak" representatives of the cluster and may change their cluster over time. It is these companies to which we refer when validating the results of the k-means model.

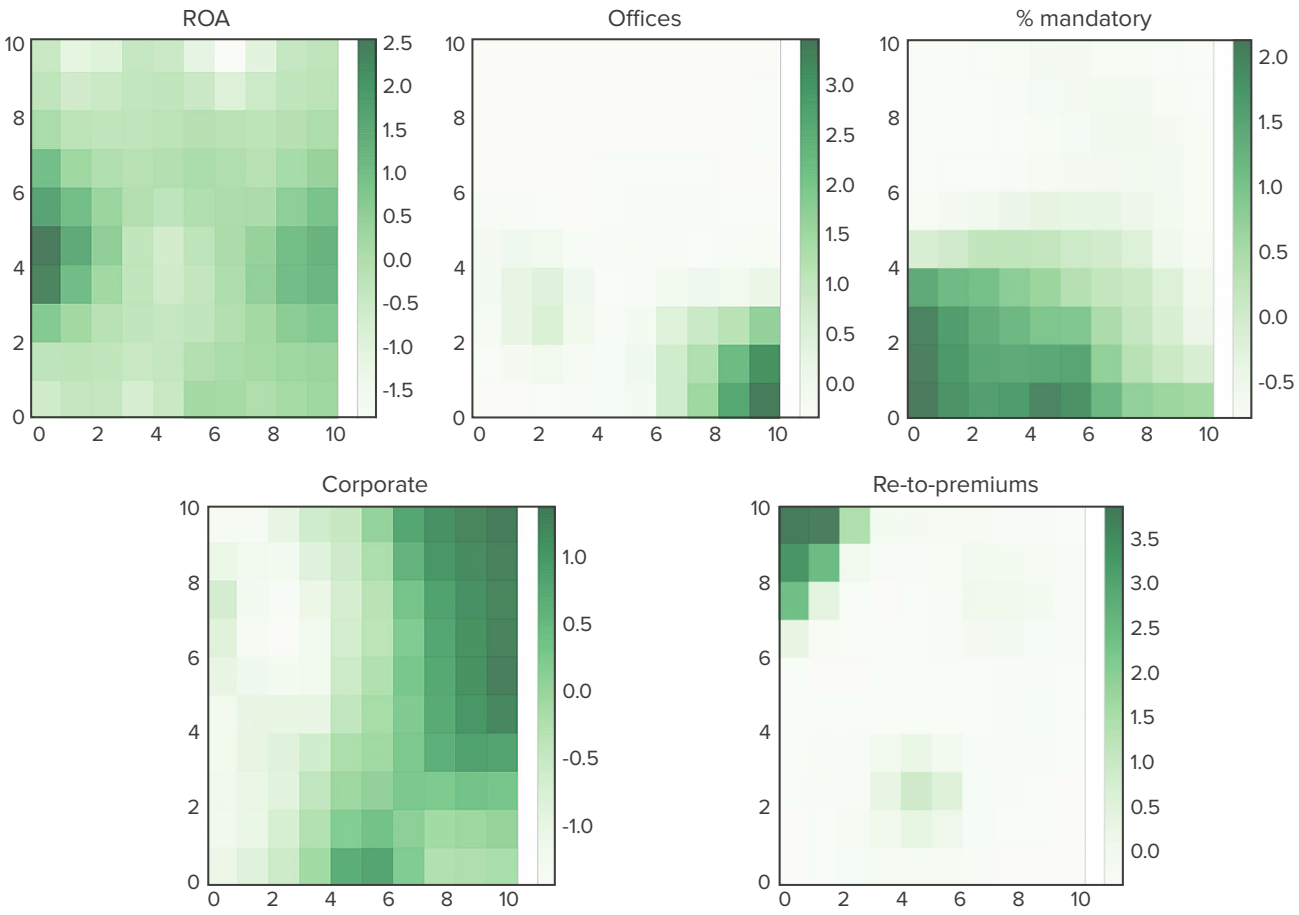


Figure 3. Kohonen Maps of Features

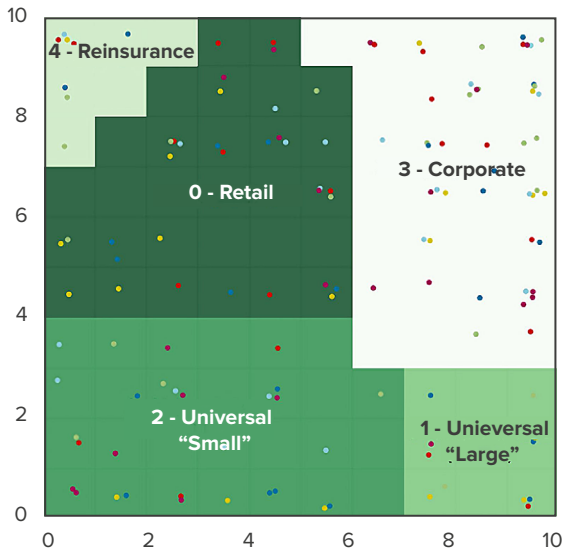


Figure 4. Grouping of Clusters on the Kohonen Map

The migration ratio for the Kohonen map between the five clusters is 16%, the one between 100 neurons is 77%. This indicates that migration within the clusters is greater than between the clusters. However, given this indicator, conclusions regarding the business models are given based on the results of k-means clustering. But it is worth noting that the economic essence of the models determined using the Kohonen map coincides with the results of the K-means method.

Analysis of Migration Between Clusters

Having identified clusters (business models), it is possible to study the dynamics of their constituents throughout the period under review. The migration coefficients from each cluster in the period from 2019 to 2020 were calculated.

As described above, migration between clusters can occur under the influence of two factors – model errors and changes in a company's business model. Knowing that clustering is not an exact method, we considered migration between clusters to be significant if more than 10% of cluster constituents migrated from it. The application of this threshold shows the migrations of at least two companies from a cluster, and most migrations that are greater than the value of the migration coefficient of the model (15%), to be significant.

Since the companies that earned less than UAH 5 million in premiums per year were not included in the k-means algorithm and were assigned to the artificially created Inactive cluster, migration from the selected business models to the Inactive cluster was also observed.

Figure 5 shows significant migrations of companies between the business models. It can be seen that there was a considerable migration in 2020 from the Universal "Large" model to the Universal "Small" model. This is to be expected, as the difference between the clusters and business models is not significant. There are less obvious reasons for the migration of companies from the Reinsurance model to the Corporate model. However, if one looks at the centers of the clusters, it can be seen that the Corporate model is closer to the Reinsurance model than the others, as the companies of the Reinsurance model have a small share of premiums from legal entities, and the companies of the Corporate model have a share of premiums from inwards reinsurance.

Given the migrations to the Inactive group, companies using the Corporate, Retail, and Reinsurance business models have a greater risk of exiting the market, and are therefore seen as less stable. More than half of the companies of the Reinsurance model exited the market in 2020, which may serve as a signal to the regulator that closer supervision is required.

Describing the business models, we noted that insurers in the Corporate model widely use the outward reinsurance of their risks, that is, they depend on the Reinsurers in their operations. Therefore, it is logical that there is a high level of simultaneous exiting from the market among companies of these two clusters.

One can assess the robustness of the conclusions based on migrations to the Inactive group. This group includes the companies whose annual premiums were less than UAH 5 million, so very small companies for which premiums of about UAH 5 million are normal could migrate due to a change in premiums from year to year. A mere 24% of companies that migrated to the Inactive cluster had

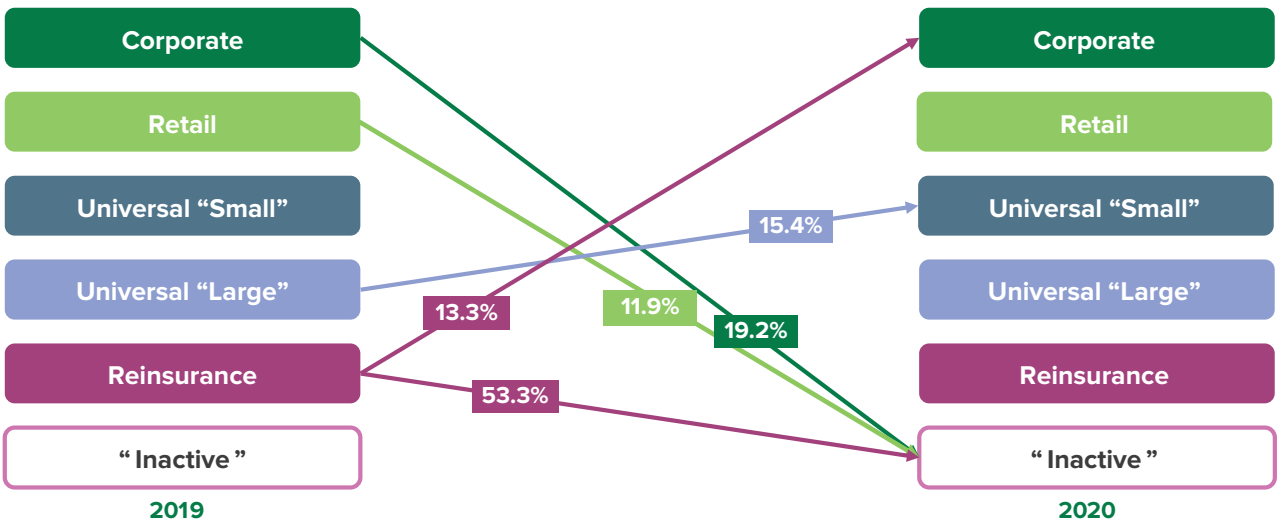


Figure 5. Migrations between the Business Models

less than UAH 10 million in premiums in 2019. The median value of premiums among the companies that migrated to the group is UAH 175 million, the average is UAH 397 million. Since a sharp reduction in the volume of premiums from such values indicates a crisis in a company's activities, the results can be considered robust. A mere 16% of these companies had a nonzero volume of premiums in 2020. That is, the majority completely discontinued insurance activity.

Migrations between the clusters and their causes can be assessed in more detail using the Kohonen map. Figure 6 shows the companies that were included in the respective clusters in either 2019 or 2020. It is worth noting that while Figure 5 shows one-way migrations, the Kohonen map shows two-way migrations (all companies that migrated from cluster to cluster).

This figure demonstrates that migration occurs mostly between neighboring clusters and neurons. The ratio of migration between clusters for the Kohonen map is close to the k-means model, at 15.8%.

For example, it can be seen that a significant migration between the Reinsurance cluster and the Corporate cluster, which was identified by both models, is due to the sharp curtailment by reinsurers of their activities and the start of the servicing of corporate clients. Since this migration does not occur between neighboring neurons and clusters, it can be argued that the companies were not on the edge of the clusters, but significantly changed their business model.

Unlike the k-means model, the Kohonen map does not show a significant migration within the Universal business model. Both clusters of this model demonstrate a slight migration of companies to and from other clusters, which we cannot deem significant. It is interesting that migration for small companies of this business model occurs mostly with neighboring neurons on the edge of the cluster, while migration for large ones happens only far from the edge of the cluster.

By depicting on the Kohonen map the companies that have discontinued providing insurance services, it is possible to highlight those of its zones that are characterized by high risk (Figure 7).

The findings of the Kohonen map are consistent with those of the k-means model; Universal can be considered the safest business model. The companies using the Reinsurance and Corporate business models are empirically the least stable.

The upper-right corner of the map is a particularly risky area of the Corporate model. There are companies whose share of legal entities in premiums is close to 100% and which offer voluntary types of insurance. It is worth noting that unprofitability is hardly the cause of these companies' high risk, as their ROA is close to the average for the market.

The lower part of the cluster on the map is risky for the Reinsurance model. These are companies that provide both reinsurance and direct voluntary insurance services. It can be concluded that more stable reinsurers are engaged solely in reinsurance activity.

It can be seen that the area of the map characterized by the highest return on assets shows absolutely no migration to the Inactive group. That is, profitable operation increases the stability of companies.

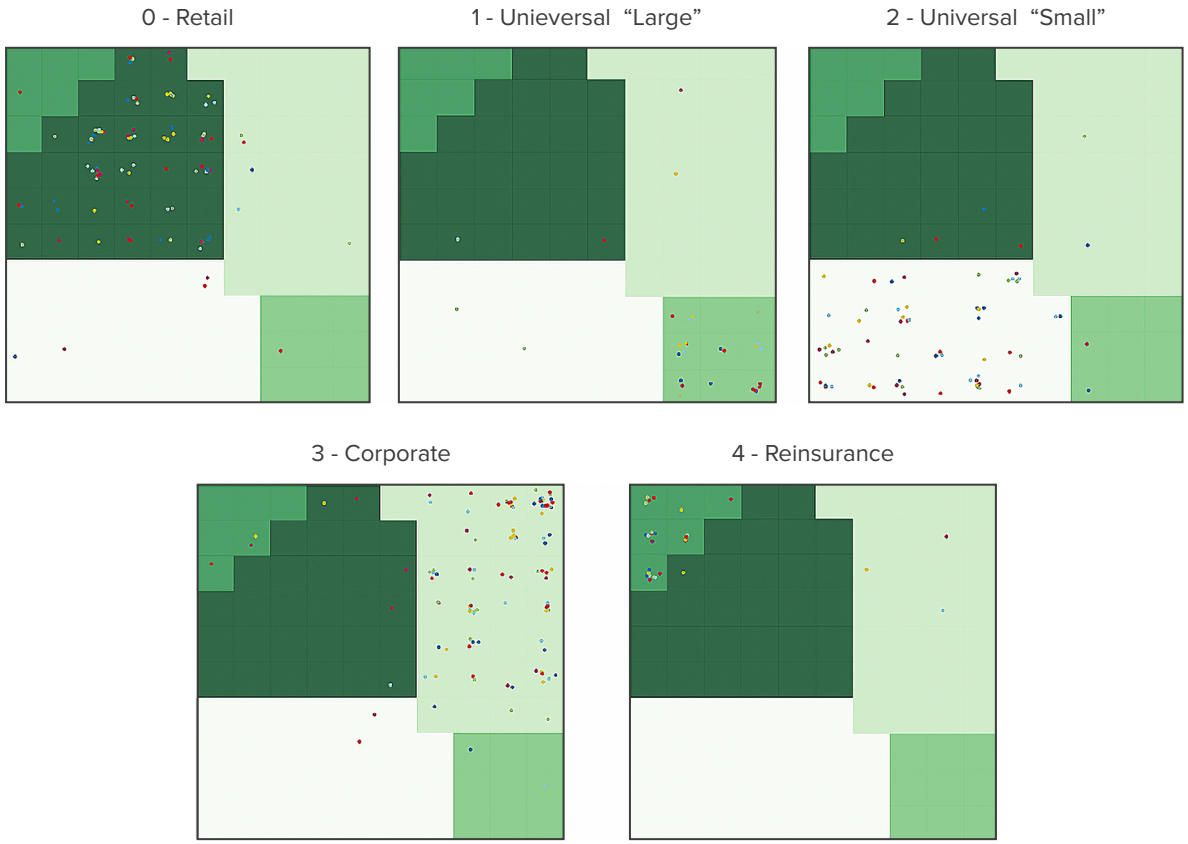


Figure 6. The Number of Neuron Activations on the Kohonen Map

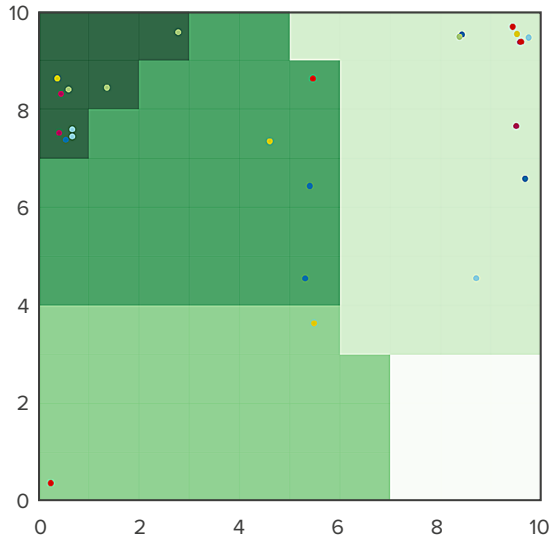


Figure 7. Companies That Migrated to the Inactive Group

5. CONCLUSIONS

In this paper, we study the nonlife insurance market in Ukraine. Particularly, we find persistent and economically reasonable ways of doing business that companies use (business models). The business model highlights not only the operational but also the risk characteristics of a company. Thus, knowledge of business models and the ability to identify which model a specific company uses is of great importance in the supervision process.

First, we decide which quantitative indicators can help to describe the business model of an insurance company. Then, we apply a set of clustering techniques to company-level indicators calculated from the regulatory database and group the companies into clusters. If we know that these groups are stable in time and are formed on the basis of indicators that describe their business model, then the description of a cluster is itself a description of a business model. The use of well-defined algorithms and performance metrics allows us to rule out personal judgment to a great extent. Finally, as we divided companies into clusters, we can study how companies changed clusters over the research period.

We apply a set of clustering algorithms to our data. Specifically, we perform clustering using the hierarchical Ward method, k-means, k-medoids, and Kohonen's SOM. We find that k-means provides the best combination of the quality of clusters' separation and their stability overtime. We also use SOMs as convenient visualization tools for clustering results, as SOM clusters carry the same economic meaning as k-means clusters.

We identify the following four different business models of insurers on the Ukrainian market based on quantitative data: Retail, Corporate, Universal (divided into two clusters, large and small), and Reinsurance. The sixth cluster is formed artificially – it includes insurance companies whose gross premiums for the year amounted to less than UAH 5 million and which were considered inactive for the purposes of this study. The research also describes the mentioned business models on the basis of the key quantitative indicators that characterize them.

Companies whose business model is retail insure individuals and tend to focus on certain types of insurance. This focus, and a low level of outward reinsurance, make them vulnerable to underwriting risk.

Large universal insurers are mostly well-known insurance companies that enjoy the trust of consumers, have many offices, and that have high profitability. They focus on selling a large number of low-priced policies.

Small universal insurers are inclined to provide mandatory types of insurance, in particular MTPL. Thus, the risks of this business model are closely related to the risks of civil liability insurance of vehicles. These companies tend to have low profitability.

Corporate insurers focus on legal entities and insure expensive risks. They make extensive use of outwards reinsurance to reduce underwriting risk. However, this makes them vulnerable to the risk of counterparty default.

We also conclude that reinsurers are the least profitable on the market, reinsuring mostly voluntary types of insurance. We reveal that reinsurers are themselves insufficiently reinsured, which makes them exposed to underwriting risk.

Then, the study shows insurance companies' migration between clusters. According to the model, companies using the Corporate and Reinsurance business models from 2019 to 2020 most often exited the market, which may indicate that such companies need more attention from the supervisor. At the same time, the Retail and Universal business models are the most stable, and therefore may be considered the least risky. Therefore, the proposed combination of methods can be considered effective for market supervision purposes.

This study provides a foundation for further research in two directions. First, we consider the clusters identified in this work to be quite broad, although they correspond to the key areas of the companies' activities. Therefore, identifying more narrowly oriented business models based on the clusters described in this study would be a logical continuation of the development of the topic. Second, in view of the described empirical dependence of an insurer's risk level on the type of business model it uses, it is extremely important to look into the risk factors that affect companies from different clusters. We see the availability of detailed and reliable data on companies on the insurance market in Ukraine as a key factor that would contribute to the further development of this topic.

REFERENCES

Abbas, S. A., Aslam, A., Rehman, A. U., Abbasi, W. A., Arif, S., Kazmi, S. Z. H. (2020). K-Means and K-Medoids: Cluster analysis on birth data collected in city Muzaffarabad, Kashmir. IEEE Access, 8, 151847-151855. https://doi.org/10.1109/ACCESS.2020.3014021

Abolmakarem, S., Abdi, F., Khalili-Damghani, K. (2016). Insurance customer segmentation using clustering approach. International Journal of Knowledge Engineering and Data Mining, 4(1), https://doi.org/10.1504/IJKEDM.2016.082072

Ahmar, A. S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., Azmi, M. (2018). Using K-Means clustering to cluster provinces in Indonesia. Journal of Physics: Conference Series, 1028, 012006. http://doi.org/10.1088/1742-6596/1028/1/012006

Arthur, D., Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (p./pp. 1027--1035), Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Bach, M. P., Vlahović, N., Pivar, J. (2020). Fraud prevention in the leasing industry using the Kohonen self-organising maps. Organizacija, 53(2), 128-145. https://doi.org/10.2478/orga-2020-0009

Caliński, T., Harabasz, J. (1972). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27. https://doi.org/10.1080/03610927408827101

Honkela, T. (1998). Description of Kohonen's self-organizing map. In Honkela, T. Self-Organizing Maps in Natural Language Processing. Helsinki: Helsinki University of Technology. Retrieved from http://www.mlab.uiah.fi/~timo/som/thesis-som.html

Kaufman, L., Rousseeuw, P. J. (1990). Partitioning around medoids (Program PAM). In Kaufman, L., Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster AnalysisFinding Groups in Data, pp. 68-125. John Wiley & Sons. https://doi.org/10.1002/9780470316801.ch2

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59–69. https://doi.org/10.1007/BF00337288

Kramarić, T. P., Bach, M. P., Dumičić, K., Žmuk, B., Žaja, M. M. (2017). Exploratory study of insurance companies in selected post-transition countries: Non-hierarchical cluster analysis. Central European Journal of Operations Research, 26(3), 783–807. https://doi.org/10.1007/s10100-017-0514-7

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Rashkovan, V., Pokidin, D. (2016). Ukrainian banks' business models clustering: Application of Kohonen neural networks. Visnyk of the National Bank of Ukraine, 238, 13-38. https://doi.org/10.26531/vnbu2016.238.013

Rousseew, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America, 99(10), 6567-6572. https://doi.org/10.1073/pnas.082099299

Velykoivanenko, H., Beschastna, G. (2018). Analysis of the stability and rating of Ukrainian insurance companies. Modelling and Information Systems in Economics, 95, 65-81. Retrieved from https://ir.kneu.edu.ua:443/handle/2010/30980

Vettigli, G. (2019). MiniSom: Minimalistic and Numpy-Based Implementation of the Self Organizing Map (release 2.1.5. 2019). Retrieved from https://github.com/JustGlowing/minisom t

Wang X., Keogh E. (2008) A clustering analysis for target group identification by Locality in motor insurance industry. Soft Computing Applications in Business. Studies in Fuzziness and Soft Computing, 230, 113–127. https://doi.org/10.1007/978-3-540-79005-1_7

Zaqueu, J. R. (2019). Customer Clustering in the Health Insurance Industry by Means of Unsupervised Machine Learning: An Internship Report. University of Lisbon, Information Management School. Retrieved from https://run.unl.pt/bitstream/10362/89468/1/TAA0043.pdf

APPENDIX A. TABLES

Table A.1. Descriptive Statistics of Variables in 2019

a) Descriptive Statistics of Variables in the Model

	ROA	Offices	Re-to-premiums	% of mandatory premiums	Corporate
Mean	0.10	6.26	0.17	0.49	0.13
Std. Deviation	0.25	18.44	0.25	0.34	0.27
Min	-0.22	0.00	0.00	0.00	0.00
Q(25%)	0.02	0.00	0.00	0.18	0.00
Q(50%)	0.05	0.00	0.03	0.44	0.01
Q(75%)	0.09	0.00	0.29	0.83	0.08
Max	2.19	115.00	0.90	1.00	1.00

b) Descriptive Statistics of Companies' Additional Characteristics

	Re-to-provisions	Loss ratio	Mean premium	Wages/Premiums	Concentration
Mean	0.23	0.49	243.71	0.02	0.70
Std. Deviation	0.25	0.50	1,199.14	0.02	0.18
Min	0.00	-0.63	0.00	0.00	0.33
Q(25%)	0.03	0.09	0.79	0.00	0.57
Q(50%)	0.14	0.43	2.69	0.01	0.69
Q(75%)	0.37	0.73	26.32	0.02	0.81
Max	1.72	3.33	13,768.22	0.07	1.00

Table A.2. Findings of the k-medoids Method

a) Coordinates of Cluster Centers

	ROA	Offices	% of mandatory premiums	Corporate	Re-to-premiums
0 – Retail	0.002	0.000	0.000	0.242	0.004
1 – Universal "Large"	0.067	44.000	0.242	0.433	0.017
2 – Universal "Small"	0.001	0.000	0.663	0.342	0.068
3 – Corporate	0.011	0.000	0.001	0.913	0.004
4 – Reinsurance	0.001	0.000	0.000	0.038	0.893

b) Number of Companies Included in the Clusters in the Period Under Review

	2019	2020
0 – Retail	44	40
1 – Universal "Large"	14	14
2 – Universal "Small"	28	28
3 – Corporate	52	46
4 – Reinsurance	14	8

Table A.3. Description of the Clusters

a) Coordinates of Cluster Centers

	ROA	Offices	% of mandatory premiums	Corporate	Re-to-premiums
0 – Retail	0.030	0.350	0.050	0.242	0.015
1 – Universal "Large"	0.065	62.330	0.277	0.482	0.008
2 – Universal "Small"	0.014	6.140	0.637	0.368	0.057
3 – Corporate	0.027	1.190	0.047	0.887	0.027
4 – Reinsurance	-0.020	0.000	0.001	0.141	0.810

b) Additional Descriptive Characteristics of the Clusters (2020)

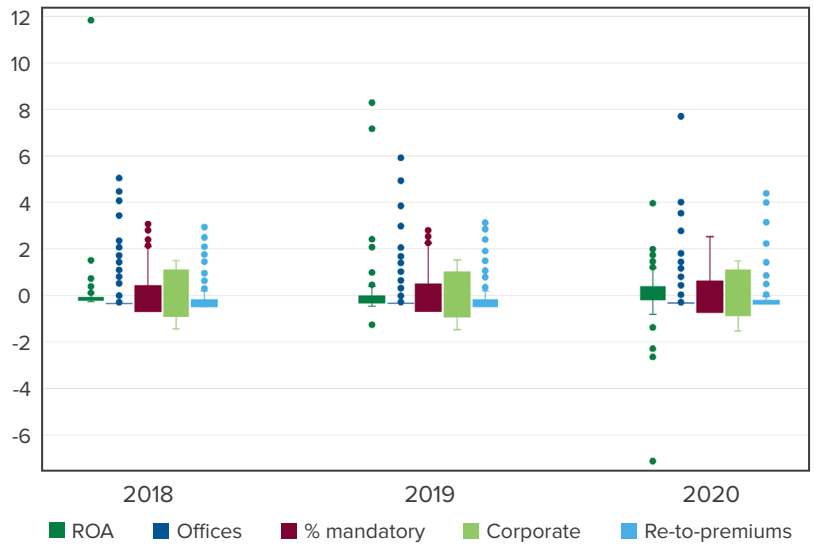
	Re-to-provisions	Loss ratio	Mean premium	Wages/ Premiums	Concen-tration
0 – Retail	0.186	0.457	5.424	0.065	0.759
1 – Universal "Large"	0.250	0.381	1.669	0.060	0.466
2 – Universal "Small"	0.168	0.390	41.900	0.064	0.637
3 – Corporate	0.275	0.225	254.20	0.051	0.701
4 – Reinsurance	0.119	0.049	303.97	0.001	0.679

c) Number of Companies Included in the Clusters in the Period Under Review

	2019	2020
0 – Retail	42	40
1 – Universal "Large"	13	12
2 – Universal "Small"	29	29
3 – Corporate	52	47
4 – Reinsurance	15	8

APPENDIX B. FIGURES

a) before



b) after

Figure B.1. Distribution of Values of Variables Before and After Adjustment of Outliers, value, years.

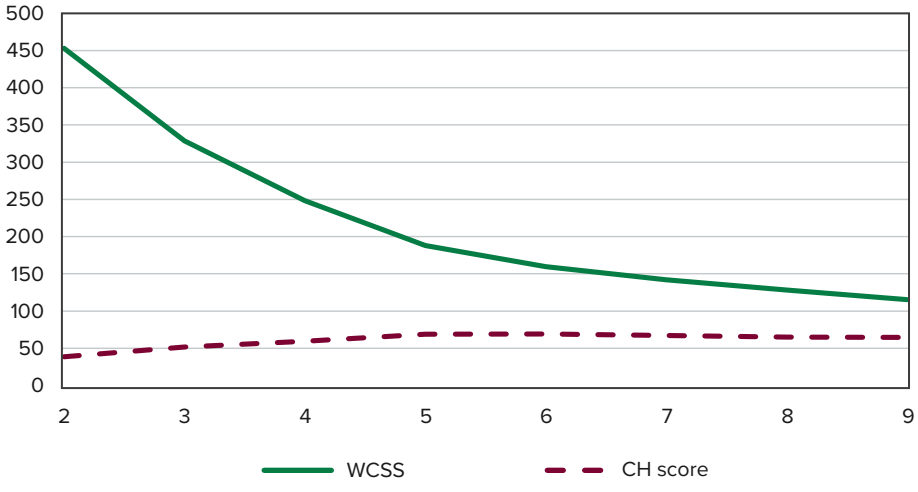


Figure B.2. Criteria for Choosing the Number of Clusters, Elbow method

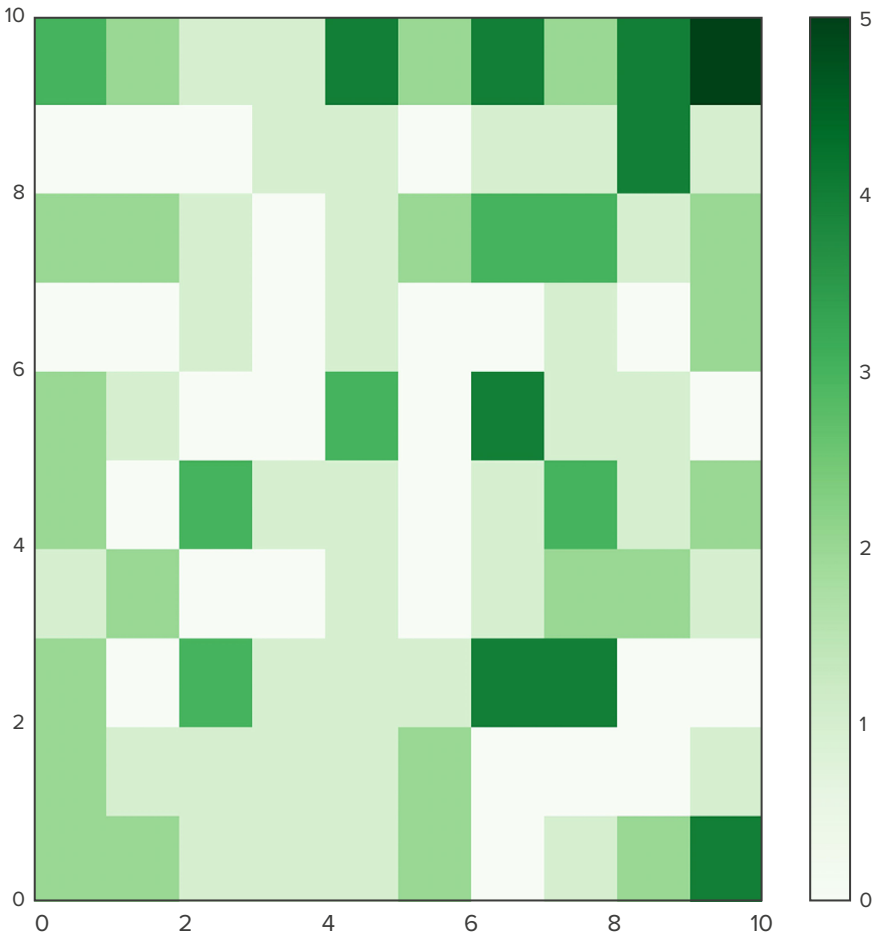


Figure B.3. The Number of Neuron Activations on the Kohonen Map

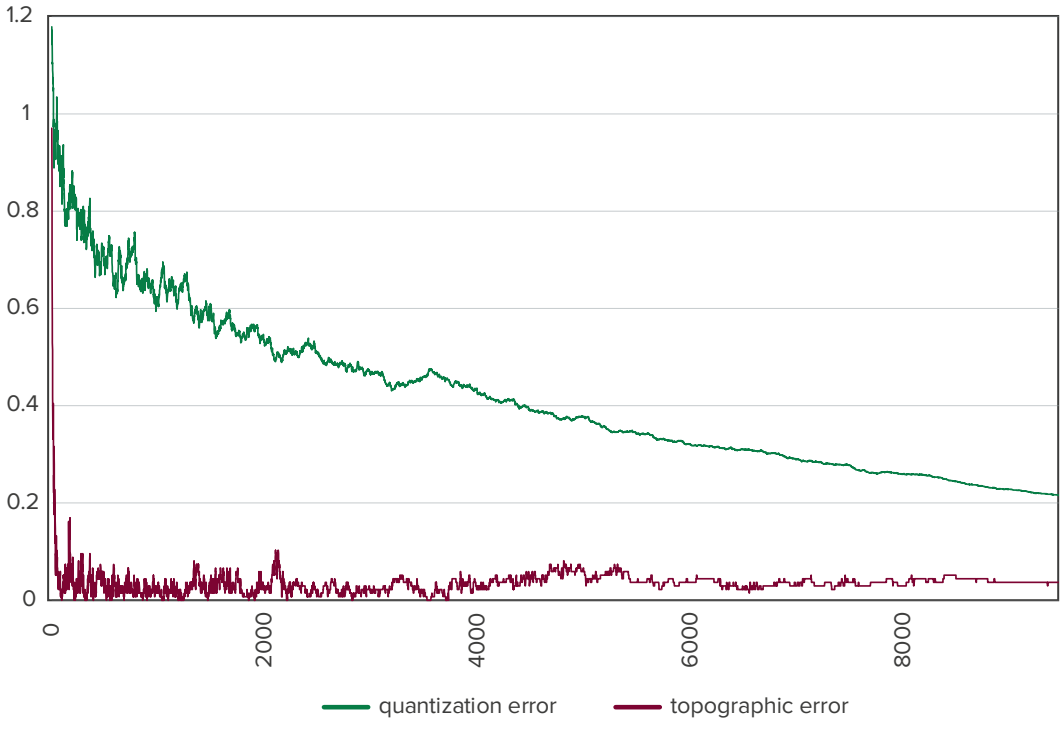
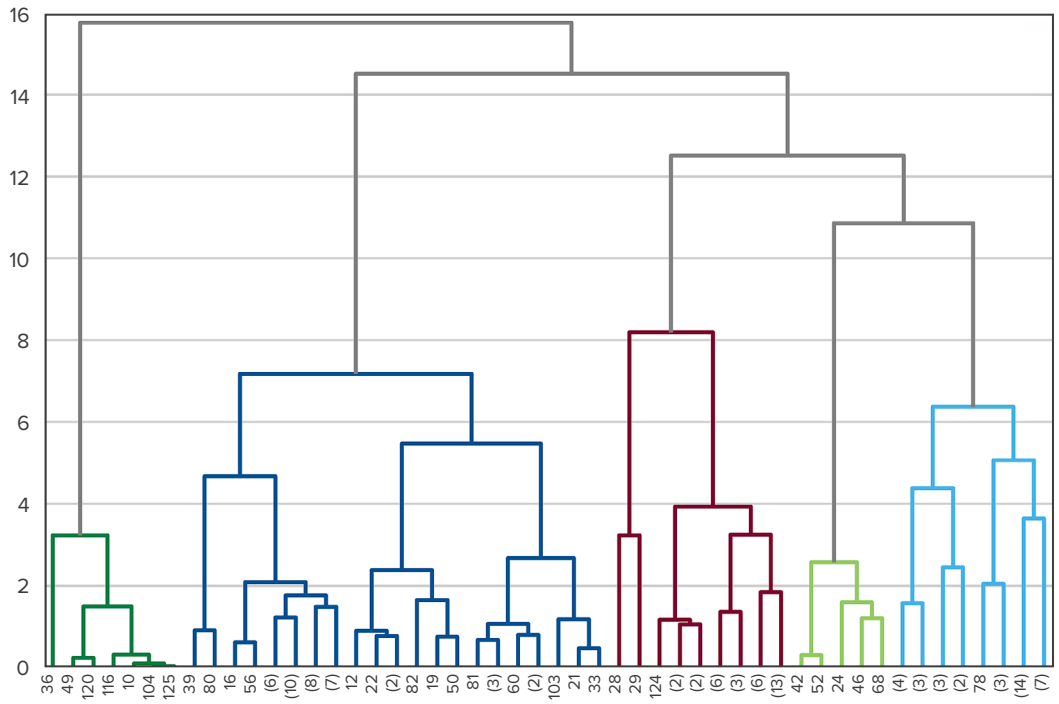


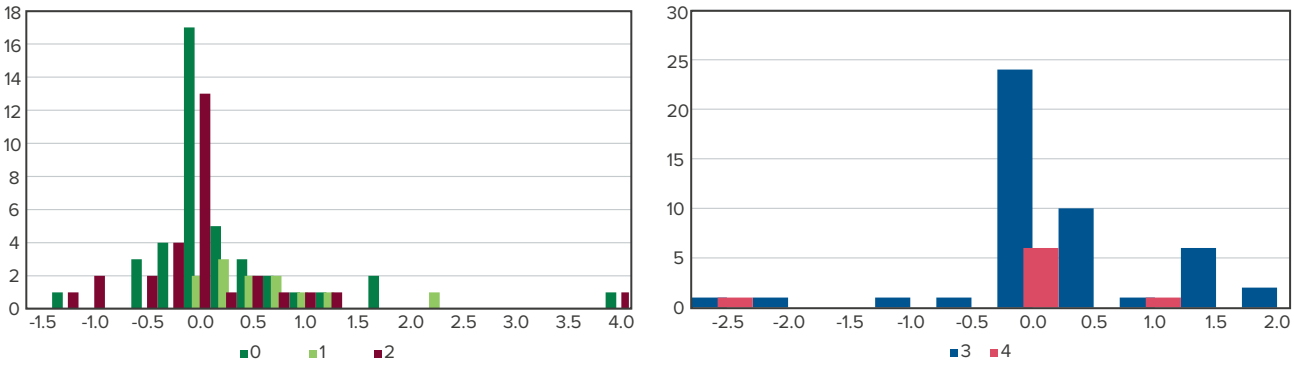
Figure B.4. The Dynamics of Kohonen Network Learning, error, iteration index



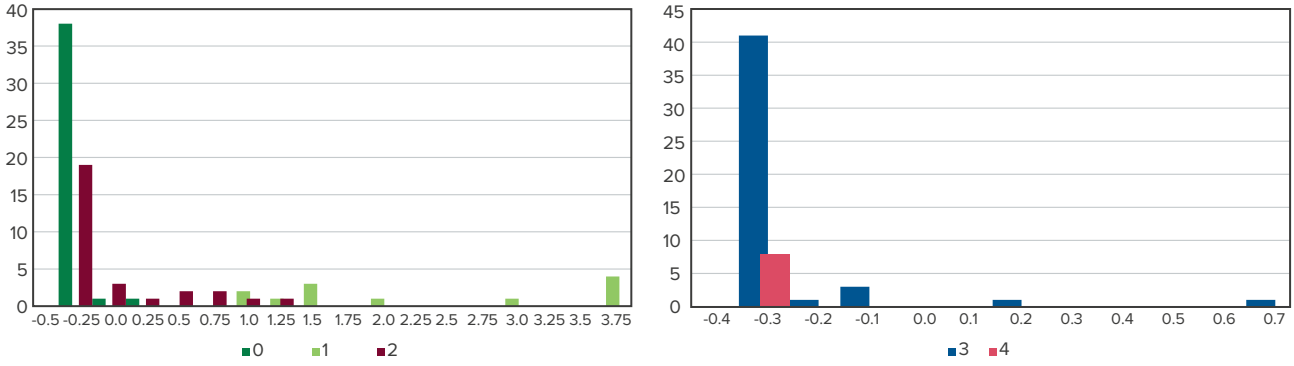
	ROA	Offices	% of mandatory premiums	Corporate	Re-to-premiums	Number of companies (2020)
0 – Retail	0.046	0.143	0.044	0.207	0.009	35
1 – Universal “Large”	0.057	92.800	0.278	0.487	0.001	5
2 – Universal “Small”	0.002	6.838	0.537	0.393	0.066	37
3 – Corporate	0.034	5.269	0.062	0.847	0.026	52
4 – Reinsurance	-0.032	0.000	0.001	0.096	0.849	7

Figure B.5. Findings of Ward’s Method

a) ROA



b) Offices



c) % of mandatory premiums

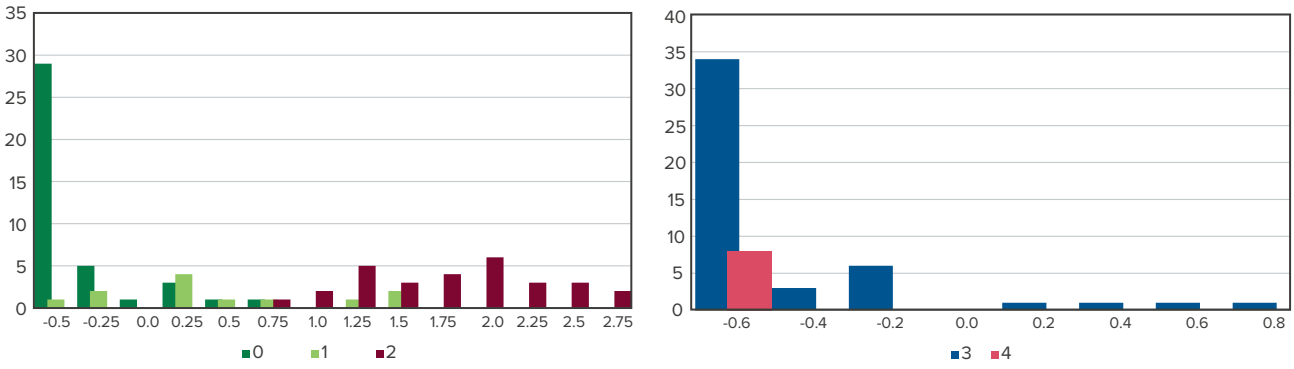
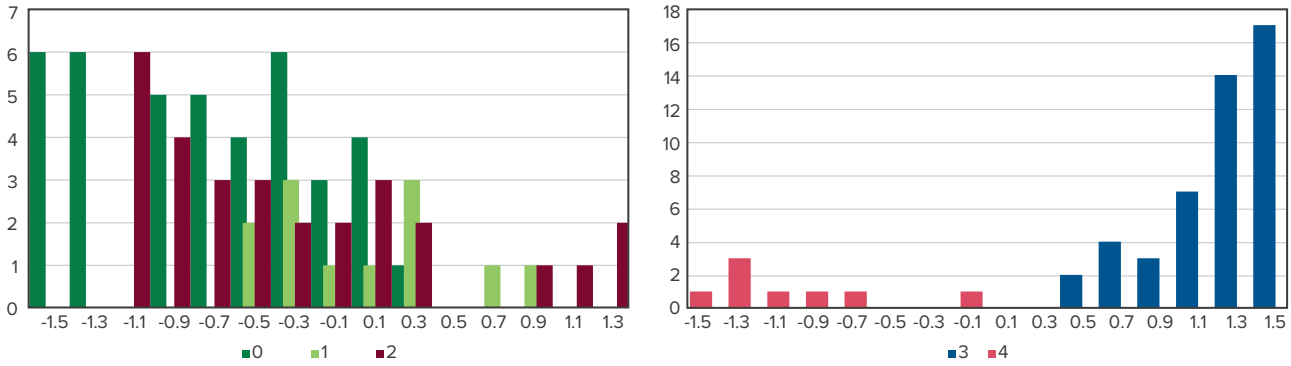


Figure B.6. Histograms of Features of the Identified Clusters (standardized)

d) Corporate



e) Re-to-premiums

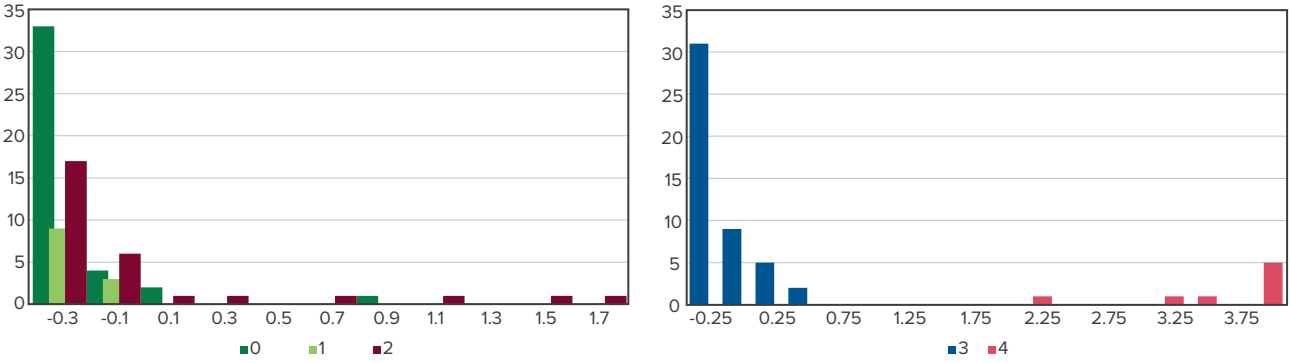


Figure B.6 (continued). Histograms of Features of the Identified Clusters (standardized)

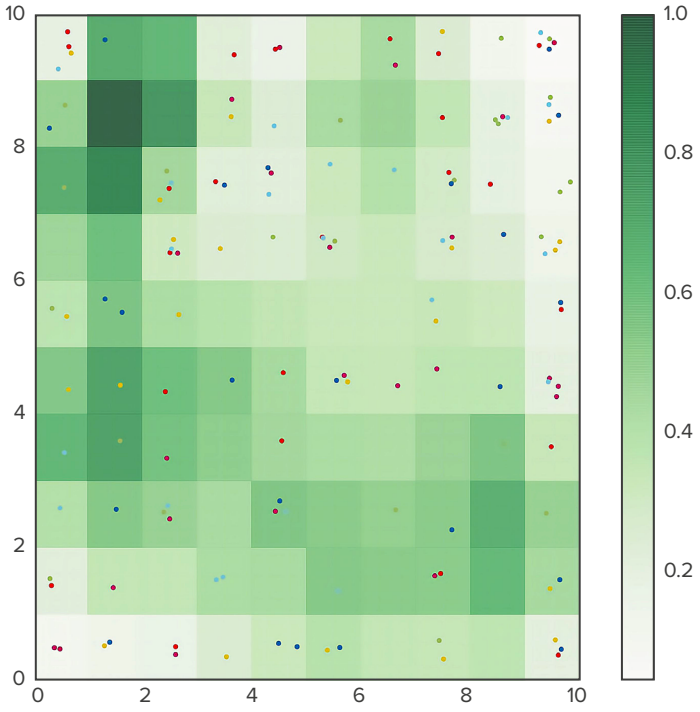


Figure B.7. Euclidean Distance Between Neurons (normalized)