Safe Bank 4(97) 2024

DOI: 10.26354/bb.4A.4.97.2024

Julia Kuchno*
ORCID: 0000-0002-8362-6871
juliakuchno@gmail.com

Clustering overdue receivables in the insurance sector: a mixed data approach

Abstract

This study addresses the issue of overdue receivables from the secondary market. The main objective of the research is to evaluate the application of the Fast K-Prototypes algorithm to the overdue insurance receivables segmentation, considering how selected parameters and data quality influences obtained results. The article also addresses the repayment of receivables from the insurance sector and the assessment of the risks they generate. The research sample includes 2376 recourse claims which arose from motor insurance and have been acquired between 2012–2023 by a Polish debt collection company.

The application of the Fast K-Prototypes method enabled the segmentation of overdue receivables into various credit risk groups, provided that specific parameters were applied, and the input data was of high quality thanks to preliminary analysis and appropriate preparation. The analysis confirms that these assets are associated with a significant level of credit risk. The results indicate that the application of the Fast K-Prototypes method supports the debt recovery process optimization. However, the effectiveness of this method depends on the research sample and suggests the importance of further research in the context of diverse data samples.

Keywords: overdue receivables, insurance sector, recovery claims, credit risk, Fast K-Prototypes

JEL Codes: C38, G22

* Julia Kuchno – PhD candidate, Warsaw School of Economics.

Segmentacja wierzytelności przeterminowanych z sektora ubezpieczeń z wykorzystaniem algorytmów mieszanych danych

Streszczenie

Niniejsze opracowanie podejmuje tematykę segmentacji wierzytelności przeterminowanych z sektora ubezpieczeń, pochodzących z rynku wtórnego. Celem artykułu jest ocena efektywności zastosowania metody Fast K-Prototypes do segmentacji wierzytelności tego typu z uwzględnieniem wpływu parametrów modelu oraz jakości danych wejściowych na jakość uzyskanych wyników. Artykuł podejmuje także tematykę spłacalności wierzytelności z sektora ubezpieczeń i oceny ryzyka przez nie generowanego. Próba badawcza zawiera 2376 roszczeń regresowych z tytułu ubezpieczeń komunikacyjnych, które były nabywane w latach 2012–2023 przez polski podmiot zajmujący się działalnością windykacyjną. Zastosowanie metody Fast K-Prototypes pozwoliło na podział wierzytelności na różne grupy ryzyka kredytowego, pod warunkiem zastosowania określonych parametrów oraz zachowania wysokiej jakości danych wejściowych poprzez odpowiednie przygotowanie i wstępną analizę. Analiza wykazała wysoki poziom ryzyka tego typu wierzytelności i ich niską spłacalność. Wyniki potwierdzają, że metoda Fast K-Prototypes może być skuteczna, ale jej efektywność zależy od jakości danych i wymaga dalszych badań w kontekście różnorodnych prób badawczych.

Słowa kluczowe: wierzytelności przeterminowane, sektor ubezpieczeń, roszczenia regresowe, ryzyko kredytowe, Fast K-Prototypes

Kody JEL: C38, G22

Introduction

One of the methods of obtaining external capital on the financial market is the process of selling overdue receivables. This process can be particularly beneficial for institutions from the insurance sector, as it ensures improved financial liquidity and, consequently, supports the maintenance of their role as public trust institutions (Śliwiński 2011, p. 467). By separating specific assets from their balance sheet, these entities have the opportunity to reduce the risk associated with their operations.

The literature emphasizes that receivables of this type may pose challenges due to the complex nature of the insurer's performance, the variety of potential legal bases and controversies surrounding the mutual duties arising from the insurance contract (Gruszczyński 2018, p. 45). The basis for a claim arising from the legal relationship established by an insurance contract include unpaid insurance premiums, recourse claims or undue benefits. The legal complexity of these claims leads to delays in satisfying the creditor and an increase in credit risk. Therefore, management of these assets, defined as their valuation and servicing in the debt collection process, requires the application of appropriate analytical methods.

The aim of this article is to present the application of the Fast K-Prototypes method to the analysis of overdue receivables from the insurance sector. The research sample consists of 2,376 overdue recourse claims from motor insurance and is described in more detail in the methodology section of the article. The key element

of the presented research lies in categorizing these assets based on their credit risk and rate of return. The Fast K-Prototypes method demonstrates high efficiency and flexibility in the process of segmenting mixed data, which is of great importance in the case of overdue receivables analysis.

The research hypothesis posits that the effectiveness of overdue receivables segmentation using the Fast K-Prototypes method depends on the appropriate selection of model parameters (number of clusters, lambda value) and the quality of the input data.

The article is structured into four parts: literature review, research methodology, analysis of the research results and summary. The literature review includes a presentation of the Fast K- Prototypes method and an overview of the application of cluster analysis in the insurance and banking sectors. The methodological section presents the research sample and the research process. Subsequently, the results of the analyses and their implications for applying the Fast K- Prototypes method for segmenting receivables from the insurance sector are discussed. The article concludes with a summary that emphasizes the key elements of the research and their significance in the context of analyzing overdue receivables.

The results presented in the article provide valuable insights, complementing the existing literature by demonstrating the application of the FAST K-Prototypes method to segment overdue receivables from the secondary market. Moreover, the formulated conclusions provide practical recommendations that can support the process of managing overdue receivables from the insurance sector.

1. Selected Clustering Methods: From Theory to Application

The literature divides clustering methods into hierarchical, non-hierarchical and fuzzy cluster analysis methods (Sala 2017, p. 142). Hierarchical methods focus on creating a hierarchy of clusters in the form of a dendrogram, enabling the analysis of relationships between groups at various levels of detail (Saxena et al. 2017, p. 666). In turn, non-hierarchical methods, such as the k-means method, divide the data set into a predefined number of groups. The assignment of observations to classes is based on their distance to the cluster centers (Sobolewski, Sokołowski 2017, p. 217). The fuzzy cluster analysis method, in contrast, assigns each data point to all clusters with a varying degree of membership instead of definitively assigning it to a single cluster (Saxena et al. 2017, p. 667). Given that the analyzed Fast K-Prototypes method belongs to non-hierarchical methods, the further overview of the database analysis methodology is limited to this domain.

Non-hierarchical methods (*k*-means, *k*-modes) are considered highly efficient and easy to interpret. However, they are criticized for their sensitivity to the initial data distribution and the initial placement of the cluster centroids (Sala 2017, p. 143). The *k*-means algorithm is effective for numerical data, whereas *k-modes*

algorithm focuses on categorical data (Huang 1998, p. 301). The combination of these methods provides the foundation for the K-Prototypes algorithm, which is particularly effective in the analysis of mixed data. The optimized version of Fast K-Prototypes offers enhanced scalability and reduced computation time, making it an effective tool supporting the management of overdue receivables (Kim 2017, p. 3).

1.1. Fast K-Prototypes Model

In the presented research, a modified K-Prototypes method, called Fast K-Prototypes (Kim 2017), was used. Data segmentation allows for a deeper understanding of the specifics of overdue insurance receivables, supporting precise debt collection strategies. High-risk receivables require intensive measures, such as enforcement by bailiffs, whereas low-risk receivables can be addressed through simpler actions, such as automatic reminders. Such a division can also support the process of resource allocation, enabling a focus on segments with the highest potential for debt recovery.

The objective function of the basic K-Prototypes algorithm aims to minimize the total dissimilarity measure between data points and centroids, which represent the centers of clusters in high-dimensional space. It is represented by Equation 1:

Equation 1. Cost Function of Model K – Prototypes

$$F(U,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} u_{i,l}\, d(x_i, q_l)$$

Source: prepared based on Z. Jia, L. Song (2020, p. 2)

where $U = [u_{i,l}]$ is the matrix representing the assignment of points to clusters, Q denotes the set of centroids for each cluster. In turn, the value $d(x_i, q_l)$ represents the difference between the point x_i and the centroid q_l, and the method of estimating this measure is presented in Equation 2:

Equation 2. Measure of dissimilarity

$$d(x_i, q_l) = \gamma \sum_{s=1}^{p} \delta(x_{i,s}, q_{l,s}) + \sum_{s=p+1}^{m} \sqrt{(x_{i,s}^N - q_{l,s}^N)^2}$$

Source: prepared based on Z. Jia, L. Song (2020, p. 2)

where p is the number of categorical variables, $m - p$ is the number of quantitative variables, and a $x_{i,s}$, $q_{l,s}$ – represent the values of the variable s at the point x_i, in the centroid q_l, respectively. The parameter γ (*gamma*) plays a key role in the K-Prototypes algorithm, balancing the impact of categorical and numerical variables on the clustering process. Higher values of γ increase the significance of categorical

variables, whereas lower values enhance the influence of numerical variables. The dissimilarity measure analyzes mixed data by dividing the differences into qualitative (Hamming) and quantitative (Euclidean) components (Sroka 2021, p. 49).

The K-Prototypes algorithm initializes centroids using the mean values of quantitative variables and the modal values of qualitative variables. Each point is assigned to a cluster based on the minimization of the cost function F(U,Q) (Huang 1998, p. 291–292). Subsequently, the value of each centroid is updated. The literature suggests that the K-Prototypes algorithm can be computationally intensive and sensitive to the initial centroid values, which affects the stability of its results (Kim 2017, p. 1).

For this reason, this study employs the Fast K-Prototypes algorithm, which modifies the distance measure equation by estimating partial distances (Kim 2017, p. 2) and introduces the parameter γ (*lambda*), equivalent to the *gamma* parameter from the basic formula of the algorithm (Equation 1). This algorithm minimizes the distance between objects and cluster centroids, reducing the need to calculate distances for all variables (quantitative and qualitative). The maximum difference between centroids in the quantitative variable space is used as a criterion for excluding certain calculations. The key assumptions include dividing data into qualitative and quantitative variables and estimating partial distances. Additionally, centroids are iteratively updated and the cost function, defined as the difference between an observation and its cluster centroid, is minimized. The partial distance estimation considers selected variables, which allows for reducing unnecessary calculations when the differences between centroids are sufficiently large to eliminate the need for further analysis of assigning points to clusters. The Fast K-Prototypes method is sensitive to initial conditions, such as the distribution of centroids, the value of the λ parameter, or the number of clusters. Therefore, the analysis using this method can be computationally more demanding. In this study, the issue of sensitivity to initial conditions was addressed by testing various parameter configurations, which enabled the achievement of stable and reliable results.

1.2. Application of clustering in the insurance and banking sector

Cluster analysis methods have been developed since the 1960s, supporting various fields, such as banking, insurance, and marketing. Although the contexts of these sectors differ, their common denominator is data segmentation, which allows for more efficient customer and risk management, streamlines operational processes and improves financial results.

In the insurance sector, clustering methods enable the tailoring of products to specific customer needs and enhance the efficiency of the management of insurance portfolios (Wen, Gao, Xiao 2021, p. 271). These methods enable the segmentation of customers based on their characteristics without requiring prior class labels. Furthermore, segmentation can be used as a preliminary stage for more advanced

predictive analyses, enhancing their precision by providing homogeneous segments. This flexibility and the ability to analyze nonlinear relationships make cluster analysis a key tool for optimizing risk management strategies and personalizing offers in the insurance sector (Jamotton, Hainaut, Hames 2024, p. 27–28).

The literature on credit scoring also highlights the use of similar methods (Jadwal et al. 2019, 2017). Similarly to insurance, where clustering helps mitigate the risk of fraud, in credit scoring it enhances the assessment of bank customers and identifies their likelihood of default (Caruso et al. 2020, p. 5). The literature highlights the broad applications of clustering, including its use in assessing credit risk and forecasting repayments (Idbenjra, Coussement, De Caigny 2024, p. 2). However, it should be noted that in the context of credit scoring and credit risk management, these methods are most often employed to achieve goals different from those of the typical predictive models (Bijak, Thomas 2012, p. 2434–2435). Cluster analysis methods are most often the initial stage, preparing the database for proper modeling.

Managing overdue receivables requires considering both the specificity of assets and their heterogeneity. Clustering, as a method for identifying hidden patterns, can prove invaluable in this context. In the case of non–performing loans, clustering enables the grouping of receivables based on similar features, such as the age of the debt or the type of collateral (Arutjothi, Senthamarai 2022, p. 88). This allows for a more accurate estimation of losses (LGD) and the calibration of models in compliance with accounting standards, such as IFRS 9 (European Central Bank 2017, p. 68). This, in turn, allows for the adjustment of debt collection strategies, improves the accuracy of repayments forecasting, and enables a more precise valuation of receivables portfolios, which is essential for both primary and secondary creditors.

Based on the literature review, it is worth noting that there is a research gap in the application of clustering methods, such as Fast K-Prototypes, to the analysis of overdue receivables. With the continuous development of machine learning methods, new solutions are being introduced to streamline segmentation processes, including those related to overdue receivables. The application of the Fast K-Prototypes method to receivables from the insurance sector, where customer segmentation is a popular practice, seems particularly interesting.

2. Methodology of the research conducted

2.1. Characteristics of the research set

Segmentation of receivables from motor insurance recourse claims was performed using data from a Polish debt collection company. The research sample comprises of 2,376 receivables acquired in portfolios between 2012 and 2023. The initial nominal value of the portfolios was PLN 7,418,378.64. During the research period, a total of PLN 1,264,319.60 was repaid. As of December 31, 2023, the current balance of receivables amounted to PLN 16,161,073.71.

The data collected included both information available at the time of the receivables acquisition and selected information regarding the receivables servicing process as of December 31, 2023.

The following quantitative variables were selected:

- nominal value at the time of acquisition (*nominal_value*),
- the ratio of costs incurred up to the moment of acquisition to the initial nominal value (*costs_initial_value*),
- costs incurred by the secondary creditor during the recovery of the receivables (*creditor_costs*),
- amount repaid after acquiring the receivable (*amount_repaid*),
- current debt balance as of December 31, 2023 (*current_balance*),
- recovery rate,
- the period of debt collection by the secondary creditor (*portfolio_period*),
- the ratio of costs incurred by the secondary creditor to the value of the amount repaid (*cost_efficiency*),
- number of bailiff repayments (*bailiff_repayments*),
- number of voluntary repayments (*repayment_debtor*),

and categorical variables:

- gender,
- type of debtor (individual, business)
- age of the debtor at the time the enforcement title was obtained by the original creditor in the ranges <18;24), <25;34), <35;44), <45;54), <55;64).

During the initial stage of the analysis, receivables were examined in terms of their repayment capacity. For this purpose, a variable representing the percentage of total repayment of receivables was introduced (Equation 3):

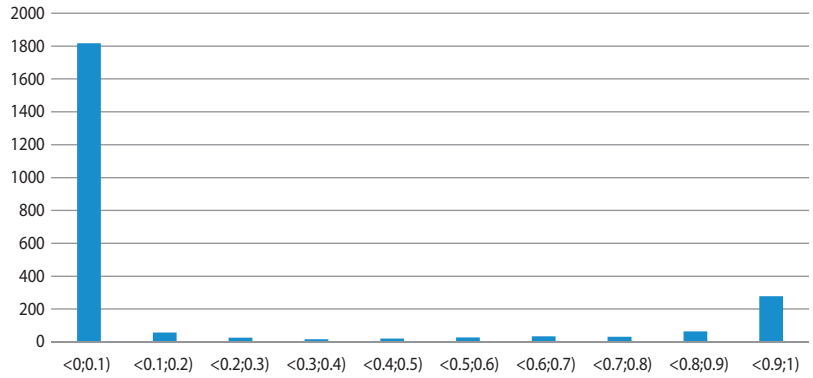
Equation 3. Synthetic Variable Recovery Rate

$$Recovery\ rate = \frac{Amount\ paid\ back}{Initial\ debt + costs\ incurred + accrued\ interests}$$

Source: W. Starosta (2020, p. 196).

The analysis of the *recovery rate* variable (Figure 1) revealed a significant asymmetry in its distribution, which confirms the assumption that overdue insurance receivables are challenging to service and carry high credit risk.

Figure 1. Analysis of the Recovery rate in cases



Source: own study based on data from one of the debt collection companies operating on the Polish market.

The literature defines credit risk as the risk of default by the debtor (Hull 2018, p. 52). Accordingly, the relationship between the recovery rate variable and the credit risk specific to the analyzed receivables can be presented using Equation 4:

Equation 4. Estimation of credit risk specific to receivables

$$Credit\ risk = 1 - Recovery\ rate$$

Source: prepared based on Hull (2018, p. 52).

Further analysis of Figure 1 reveals that, although there are more cases in which repayment has been made (1298) compared to the number of cases without any repayment (1078), a high percentage of non-repayment can still be observed in the dominant part of cases in the range <0;0.1). This highlights a significant number of cases in which the repayment rate was less than 10% of the nominal value of the receivable. Such a distribution may affect the segmentation outcome; therefore, it is expected that the obtained clusters will not be of equal size.

The study sample was analyzed based on selected quantitative variables. The analysis of Table 1 indicates a low nominal value in the cases. However, significant differences between the median and the mean, along with high standard deviation suggest considerable variability.

A similar relationship can be observed for costs incurred prior to the portfolio's sale by the original creditor. The median value suggests that these costs account for approximately 30% of the nominal value of the receivable at the time of the transaction.

Table 1. Basic statistics for the variables studied

Variable name	Mean	Median	Deviation
Initial_name_value	2 867.26	939.76	7,028.49
Initial_costs	1,003.69	322.93	3 797.16
Costs_initial_value	0.36	0.33	0.23
Costs_creditor	1 828.51	630.56	3 758.48
Amount_paid	962.29	22.95	2,668.95
Number_of_payments	3.16	1.00	7.92
Balance_current	8,934.27	1 402.77	25,540.83
Recovery_rate	0.31	0.03	0.41
Portfolio_period	7.53	10.00	4.68
Payment_bailiff	2.23	1.00	5.86
Payment_debtor	0.92	0.00	5.04

Source: own study based on data from one of the debt collection companies operating on the Polish market

Financial data as of 31.12.2023 indicate a significant increase in the nominal value due to accrued interest and debt collection costs. The high efficiency of enforced debt collection is also evident from the significant number of bailiff repayments. Nevertheless, despite the high mean recovery rate in the data sample (31%), the median value for this variable is approx. 3%, which suggests a very high probability of default, given that the median duration of debt collection process (*portfolio_period*) is 10 years. Initial results confirm the assumption that as a result of clustering, one group may significantly outnumber the others.

2.2. Research plan and methodology

This research was designed based on the article presenting the Fast K-Prototypes method (Kim 2017) and a review of the literature on non-hierarchical cluster analysis methods. The research was conducted using the R environment and the Excel spreadsheet. Figure 2 illustrates the course of activities carried out during the research.

The preliminary data analysis was one of the key stages of the study, as it allowed for understanding the specificity of the data, enabling the identification of the key quantitative and categorical (qualitative) variables. During the study, observations with missing data were removed.

Figure 2. Presentation of the research process

Source: own study based on M. Walesiak (2008, p. 45).

Subsequently, using data visualization and principal component analysis (PCA), the database's dimensionality was determined, and outliers were identified based on the Mahalanobis distance (Hubert, Debruyne 2010, p. 38). These observations were removed. In the next step, correlation (Pearson's coefficient) and multicollinearity of quantitative variables [variance inflation factor (Welfe 2018, p. 39, 149)] were analyzed. The correlation analysis was performed using the Pearson coefficient (Welfe 2018, p. 39). In addition, the ANOVA test was used to examine the impact of qualitative variables (Wu, Hu, Zheng 2021, p. 5407) on the *Recovery_rate* variable.

The initial data analysis was completed by normalizing quantitative data using the min-max method (Walesiak 2008, p. 45). Categorical variables were transformed into separate levels, which allowed the algorithm to correctly calculate the measure of dissimilarity between observations (Kim 2017, p. 4).

The next step was to build the Fast K-Prototypes model, as described in Section 1.1. In the Fast K-Prototypes algorithm, the centroids of categorical and numerical variables were randomly selected, and the distances were calculated for both types of variables. The algorithm minimized the objective function, where the lambda parameter (λ) balanced the influence of quantitative and qualitative variables. Observations were assigned to clusters with the smallest distance from the centroids. The centroids were updated in each iteration. To find the optimal model, different configurations of the *k* and *lambda* parameters were tested.

The range for the *k* parameter was set at <2;6> to limit the computational complexity (Kaminskyi, Nehrey 2021; Caruso et al. 2020). In turn, the *lambda* parameter was tested in the range of <0.1;2> (Kim 2017, p. 4). After the segmentation was completed, the quality of the results was evaluated using the cost value of the objective function and the indicators presented in Table 2.

A review of the literature suggests that the distribution of observations across groups may be uneven in the banking or insurance sector (Jadwal et al. 2022; Kaminskyi, Nehrey 2021; Abolmakarem, Abdi, Khalili-Damghani 2016). Taking into account the specificity of overdue receivables from the insurance sector, the number of observations in the sample, as well as the data structure presented in section 2.1, it was decided that segmentation would be considered effective if the average value of the Mean Silhouette Score is higher than zero and the clusters differ in terms of average credit risk levels and rates of return.

Table 2. Presentation of performance evaluation measures for the FAST K- Prototypes algorithm

Indicator name	Equation	Description
Calinski-Harabasz Index (CH)	$CH = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}}$	*SSB* – sum of squared distances between clusters *SSW* – intra-cluster variance. Higher CH index values indicate better clustering quality, due to high separation between clusters (*SSB*) or low intra-cluster variance (*SSW*).
Mean Silhouette Score	$S(i) = \frac{b(i) - a\,(i)}{\max(a(i), b(i))}$	$a(i)$ – average distance of observation *i* to points in the same cluster. $b(i)$ – average distance of observation *i* to points in the nearest neighboring cluster The indicator value lies within the range <-1;1>. The higher the indicator values above zero, the better the point is assigned to the cluster.
Dunn index	$D = \frac{min_{i \neq j} d(C_i, C_j)}{max_k \Delta(C_k)}$	$d(C_i, C_j)$ – distance between clusters *i* and *j*, $\Delta(C_k)$ – maximum distance between two points in a cluster. The higher the value of this indicator, the better the quality of the segmentation.

Source: own study based on M Walesiak M. (2008, p. 8).

3. Analysis of the study results

This part of the article presents the research findings on the effectiveness of using the Fast K-Prototypes algorithm for segmenting receivables from the insurance sector. Particular attention is given to the impact of the selection of algorithm parameters on the segmentation results. The following section presents the segmentation of the research sample under the optimal parameter scenario and highlights key observations from applying this method to the segmentation of overdue receivables.

3.1. Description of the research process and analysis of segmentation quality

During the initial data analysis, outliers (41) were identified and removed. The elimination of these observations improved the quality of further analyses and clustering results.

In the next step, a correlation analysis of categorical variables was performed, the Variance Inflation Factor was calculated and the ANOVA test was performed. Based on the initial analysis, the variables *portfolio_period* and *debtor_type* were removed from the set. The correlation analysis of variables showed strong correlations between the *recovery_rate* variable and the remaining variables related to the

debt collection process (*amount repaid*, *bailiff_payment* and *debtor_payment*). The remaining variables had statistically significant results.

After completing the variable selection, the algorithm was initiated. Using the Fast K-Prototypes algorithm, the results presented in Table 3 were estimated. The table presents the values of segmentation quality indicators for different configurations of the number of clusters (*k*) and the *lambda* parameter.

Table 3. Analysis of segmentation quality using FAST K-Prototypes method

k	lambda	Mean Silhouette	Dunn Index	Cost function	CH Index
2	0.1	0.198	0.00000403	1294.22	15.91
2	0.3	0.196	0.00000403	1505.51	16.21
3	0.1	0.151	0.00000403	1291.62	10.24
3	0.5	0.002	0.00000403	1559.15	6.83
4	0.3	0.026	0.00000403	1458.81	10.71
5	0.7	-0.053	0.00000403	1545.42	6.75
6	0.5	-0.077	0.00000403	1442.03	5.98

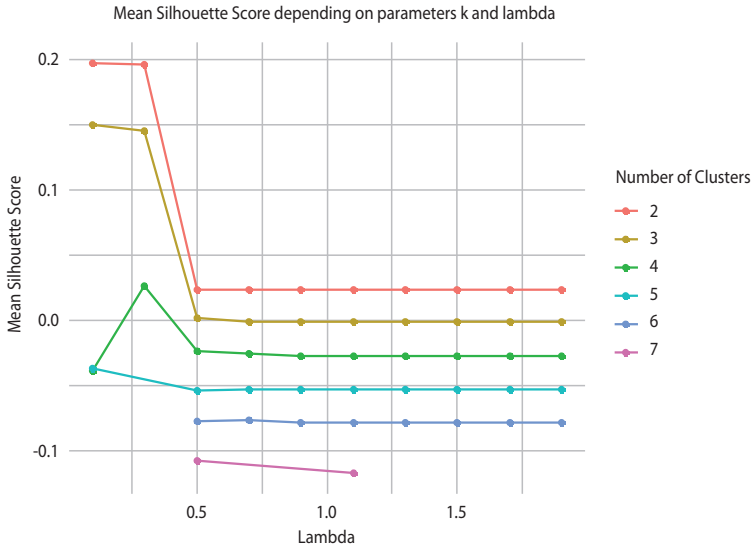
Source: own study based on own research.

The analysis of the clustering efficiency measures indicates that the optimal division was achieved with two clusters and the *lambda* parameter of 0.1. This was confirmed by the highest value of the Mean Silhouette coefficient and the low value of the cost function. This conclusion is further supported by the analysis of Figure 3 presenting the values of the Mean Silhouette Score for various configurations of the number of clusters (*k*) and the *lambda* parameter.

For a larger number of clusters (e.g., 5 or 6) the index values are negative and close to zero, suggesting that cluster boundaries may overlap. This aligns with the observations in the table, where the Mean Silhouette index values were much lower for these configurations.

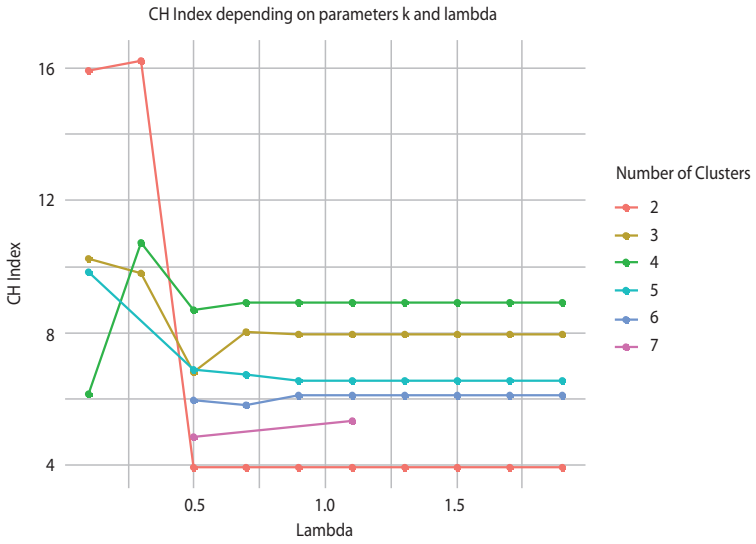
Further analysis of Table 3 reveals that the optimal CH index was achieved for the configuration of 4 clusters and the lambda parameter equal to 0.3. Based on this value, it can be concluded that the highest internal consistency and separation between clusters was achieved for this set of parameters. It can also be observed that the highest CH index value was obtained for the scenario with two clusters and *lambda* parameter set to 0.1. However, as the *lambda* parameter increases, the CH index value decreases, indicating a deterioration of the clustering quality, as shown in Figure 4.

Figure 3. Mean Silhouette Score Value Analysis



Source: own study based on own research in the R environment.

Figure 4. CH Index Analysis



Source: own study based on own research.

A higher number of quantitative variables in the sample favors lower values of the *lambda* parameter, thereby reducing their impact on the cost function. Lambda values also significantly affect the stability of the Dunn index. The analysis results indicate that it takes the same values for all parameter sets, which may also suggest a low level of separation between clusters. The analysis of segmentation quality indices using the Fast K-Prototypes method indicates that the model assigns greater weight to diverse variables; however, further elimination of outliers may be necessary in this regard.

3.2. Segmentation analysis for the optimal scenario

In the subsequent part of the study, results were generated for the scenario of parameters *k* and *lambda*, which was characterized by the best indicator values (2; 0.1). The analysis was carried out for selected variables in relation to the *Recovery_rate*: *balance_current* and *efficiency_cost*. The clustering analysis began with an assessment of the average values of the *Recovery_rate* variable and the corresponding average values of credit risk.

The values of Mean Silhouette indices presented in Table 4 indicate that the first cluster exhibited significantly lower segmentation efficiency compared to the second cluster. This may suggest greater heterogeneity of observations or indicate the presence of outliers in this group. As assumed at the beginning of the study, the clusters differ in terms of the number of assigned observations.

Table 4. Analysis of segmentation results using the Fast K-Prototypes method

Cluster number	1	2
Mean Silhouette	–4%	22%
Number of observations	908	1427
Average recovery rate	21.44%	16.69%
Median value of the recovery rate	0.00%	0.00%
Standard deviation	37.05%	33.46%
Minimum recovery rate	0%	0%
Maximum recovery rate	100%	100%
Medium credit risk	78.56%	83.31%

Source: own study based on own research

The observed average recovery value in the first cluster exceeded the threshold estimated for the second cluster. However, it should be noted that the median value for both clusters is 0%, indicating that most observations in both sets are associated with similar credit risk. A higher level of variability of the rate of return and credit risk is also observed in the first cluster (low Mean Silhouette level, high standard deviation). The preliminary analysis of the performed segmentation suggests that the first cluster contains observations with a higher repayment potential, but more diversified features compared to the second cluster. In turn, the second cluster is characterized by greater homogeneity and stability; however, the average credit risk estimated for this group is correspondingly higher.

In order to formulate conclusions for effective receivables management, the analysis was supplemented by examining the cost efficiency in cases, defined as the ratio of costs incurred by the secondary creditor to the amount repaid (Table 5).

Table 5. Cost Efficiency Index Analysis

	Cluster 1			Cluster 2		
Current balance	**L**	**RR**	**CR**	**L**	**RR**	**CR**
0	478	19%	81%	756	15%	85%
(0; 0.2>	214	21%	79%	321	13%	87%
>0.2	216	27%	73%	348	23%	77%
Grand total	908			1 425		

Source: own study based on own research.

In the case of cost efficiency, the trend of the credit risk observed in cases is similar. As the indicator increases, the credit risk decreases. An increase in the cost efficiency indicator signifies, on the one hand, that the case was repaid, and on the other hand, that the costs incurred by the secondary creditor represented a certain percentage of this repayment. In the first cluster characterized by a higher repayment potential, the higher costs incurred lead to these cases generating a lower average credit risk. In the second cluster, however, despite a similar upward trend in the recovery rate with increasing costs, the profitability of these cases is relatively lower.

In the context of managing overdue receivables, the results from Tables 4 and 5 provide insights into the overdue receivables portfolio from the insurance sector. Since the debt collection process aims to maximize the recovery rate, it can be concluded that the resources allocated to receivables in the first group are utilized more efficiently than those in the second group. This means that enforced debt collection measures should be implemented for these cases. In contrast, the rate of return indicator in the second cluster assumes a relatively higher value for cases where no costs were incurred. In such a situation, it would be worth considering the implementation of amicable debt collection measures for these cases.

Summary

The results of the conducted research confirm the hypothesis that the effectiveness of overdue receivables segmentation using the Fast K-Prototypes method depends on the appropriate selection of parameters and the quality of the input data. The highest segmentation quality indicators were observed for two clusters and a low *lambda* value. This means that the Fast K-Prototypes method requires precise adaptation to the characteristics of the data set. Moreover, understanding the specificity of the data through the principal components analysis (PCA) and the examination of collinearity and correlation between variables enhances the quality of this cluster analysis method. Such conclusions highlight that effective application of the method requires careful preparation of data, parameter optimization, and testing of various parameter scenarios for the data set.

At the same time, applying the Fast K-Prototypes method to segment overdue receivables from the insurance sector enabled the division of the analyzed data set into two groups with differing average rates of return, and consequently different average credit risk indicators. This indicates that despite low cluster separation quality indicators and a relatively homogeneous data set (a large percentage of unpaid cases, the same basis for the claim), the Fast K–Prototypes method divided the analyzed data set in line with initial expectations, into a group with a lower recovery rate potential and a higher level of credit risk, and a group with a lower level of credit risk and a higher profitability. Such division into clusters with varying profitability is also confirmed in the literature (Caruso et al. 2020, p. 5). Based on the segmentation, it was possible to formulate conclusions about the efficiency of the debt collection process in these groups. The group identified during research, with a higher potential rate of return and a lower credit risk indicator, appears to be more profitable in the context of judicial and enforcement debt collection. In contrast, the group with higher credit risk exhibited lower cost efficiency, indicating that the expenditure on debt collection may not yield proportional benefits. Based on the conducted research and analyses, a general conclusion can be drawn that recourse receivables from the insurance sector are characterized by high credit risk. The specificity of overdue receivables from the insurance sector is associated with the high legal complexity, which results in a prolonged debt collection process and a lower rate of return. This is further confirmed by the low percentage of closed (paid) cases in both groups, which suggests challenges in promptly fulfilling the creditor's claims. Based on this, a general observation can be made that for the original creditor, the sale of such claims is effective from the perspective of asset management and risk mitigation, as it helps to reduce costs.

Bibliography

Monographic publications

Hull JC, *Risk Management and Financial Institutions*, John Wiley & Sons, Inc., New Jersey 2018.

Śliwiński A., *Insurer risk and insurance debt collection*, [in:] K. Kreczmańska-Gigol (ed.), *Debt collection, an interdisciplinary approach*, Difin, Warsaw 2011.

Welfe A., *Econometrics*, Polish Economic Publishing House SA, Warsaw 2018.

Articles press and occasional

Abolmakarem S., Abdi F., Khalili-Damghani K., *Insurance customer segmentation using clustering approach*, "International Journal of Knowledge Engineering and Data Mining", vol. 4, no. 1, 2016.

Arutjothi G., Senthamarai C., *Assessment of Probability Defaults Using K-Means Based Multinomial Logistic Regression*, International Journal of Computer Theory and Engineering, vol. 14, no. 2, 2022.

Bijak K., Thomas L.C., *Does segmentation always improve model performance in credit scoring?*, "Expert Systems with Applications", vol. 39, no. 3, 2012.

Caruso G., Gattone S.A., Fortuna F., Di Battista T., *Cluster analysis for mixed data: An application to credit risk evaluation,* "Socio-Economic Planning Sciences", vol.73, no. 100850, 2020.

Gruszczyński A., *Transfer of receivables from a property insurance contract*, „Wiadomości Ubezpieczeniowe", No. 2, 2018.

Hubert M., Debruyne M., *Minimum covariance determinant*, Wiley Interdisciplinary Reviews: Computational Statistics, 2010.

Huang Z., *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, "Data Mining and Knowledge Discovery", 1998.

Idbenjra K., Coussement K., De Caigny A., *Investigating the beneficial impact of segmentation-based modeling for credit scoring*, Decision Support Systems, vol. 179, no. 114170, 2024.

Jadwal P.K., Jain S., Gupta U., Khanna P., *K-Means clustering with neural networks for ATM cash repository prediction*, [in:] S. Satapathy, A. Joshi (eds), *International Conference on Information and Communication Technology for Intelligent Systems,* Springer, Cham 2017.

Jadwal P.K., Jain S., Gupta U., Khanna P., *Clustered support vector machine for ATM cash repository prediction*, [in:] B. Pati, C. Panigrahi, S. Misra, A. Pujari, S. Bakshi (eds), *Progress in Advanced Computing and Intelligent Engineering,* Springer, Singapore, 2019.

Jadwal P.K., Pathak S., Jain S., *Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis*, "Microsystem Technologies", vol. 28, 2022.

Jamotton C., Hainaut D., Hames T., *Insurance Analytics with Clustering Techniques*, "Risks", vol. 12, no. 9, doi :10.3390/risks12090141, 2024.

Kaminskyi A., Nehrey M., *Clustering approach to analysis of the credit risk and profitability for nonbank lenders*, CEUR Workshop Proceedings, Machine Learning Methods and Models, Pre-

dictive Analytics and Applications – 13th Workshop on the International Scientific Practical Conference Modern Problems of Social and Economic Systems Modeling, MPSESM-W, 2021, https://ceur-ws.org/Vol-2927/paper10.pdf [access: 5 December 2024].

Kim B., *A Fast K-prototypes Algorithm Using Partial Distance Computation*, "Symmetry", vol. 9, no. 58, 2017.

Sala K., *Review of data clustering techniques and application areas*, „Society and Education", vol. 25, no. 2, 2017.

Saxena A., Prasad M., Gupta A., Bharill N., Patel O.P., Tiwari A., Joo E.M., Weiping D., Lin C.T., *A review of clustering techniques and developments,* "Neurocomputing", vol. 267, 2017.

Starosta W., *Modeling Recovery Rate for incomplete defaults using time-varying predictors*, Central European Journal of Economic Modeling and Econometrics, no. 12, 2020.

Walesiak M., *Cluster analysis procedure using the ClusterSim computer program and the R environment*, „Scientific Works of the Wrocław University of Economics", No. 7 (1207), 2008.

Wen Ch., Gao K., Xiao Y., *Data-Driven Market Segmentation in Insurance Industry and Other Related Sectors,* Journal of Finance and Accounting, vol. 9, no. 6, 2021.

Wu S., Hu X., Zheng W. et al., *Effects of reservoir water level fluctuations and rainfall on a landslide by two-way ANOVA and K-means clustering,* Bulletin of Engineering Geology and the Environment, vol. 80, 2021.

Zhang X., Yu L., *Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods*, "Expert Systems with Applications", vol. 237 (A), 2024.

Zhou L., Zhang N., *Customer Segmentation and Optimal Insurance Compensation Ratio: Decision-making Analysis in Financial Institutions*, International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 8, 2015.

Jia Z., Song L., *Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient,* "Mathematical Problems in Engineering", 2020.

Internet materials

European Central Bank, *Guidance to banks on non-performing loans*, 2017 https://www.bankingsupervision.europa.eu/ecb/pub/pdf/guidance_on_npl.en.pdf [access: 5 December 2024].