IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47ISM/2026/4728226/
Vol. 2026 (7), Article ID 472826

Valuation of Bulk Debt Portfolios Based on Clustering and Historical Operational Data: Evidence from Poland*

Łukasz JANKOWSKI and Rafał JANKOWSKI

AGH University of Krakow, Poland

Correspondence should be addressed to: Rafał JANKOWSKI, rjankowski@agh.edu.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

Bulk debt portfolios constitute a significant asset class in the financial market, yet their acquisition value is often determined under uncertainty regarding future recoveries and collection costs. Existing valuation approaches rarely integrate debtor-level operational data, segmentation logic, and cost-adjusted recovery estimates into one transparent decision-support model. This study addresses this gap by developing an interpretable valuation model for bulk debt portfolios based on clustering and historical operational data from the Polish debt collection market. The empirical analysis uses 120,000 real debt claims and applies two unsupervised machine learning methods: K-means clustering and Ward's hierarchical clustering. The historical portfolio is divided into homogeneous debtor segments, and each segment is assigned observed financial parameters, including recovery rates and collection costs. The value of a new portfolio is then estimated as the discounted sum of expected recoveries less expected collection costs across the identified segments. The results indicate three stable and economically interpretable portfolio segments. The consistency between K-means and Ward's method is high, with an adjusted Rand index of 0.956, while the silhouette coefficient of 0.306 confirms a moderate but meaningful segmentation structure. The segments differ in debtor characteristics and financial performance, which supports their separate treatment in valuation. The proposed model improves practical portfolio pricing by combining machine learning segmentation with cost-adjusted cash-flow logic, making it suitable for debt collection companies and securitization funds operating in markets with large-scale receivables.

Keywords: valuation of debt portfolios, machine learning, debt management, corporate finance, risk management

Introduction

The market for mass debts is one of the fastest-growing segments of the financial market not only in Poland, but also across Europe. According to data from the Association of Financial Enterprises in Poland, at the end of 2024 entities operating in the debt management sector were servicing a portfolio with a nominal value of PLN 185.9 billion, comprising 19.2 million obligations, with this value having grown at an annual rate of approximately 10% since the beginning of 2020 (ZPF, 2026). Mass debts, typically arising from unpaid bills for telecommunications services, electricity, gas, streaming subscriptions, insurance premiums or small consumer loans, are characterized by low individual value and very large volume, which means that they are grouped into packages and sold to specialized debt collection entities that assume recovery risk in exchange for a purchase price representing a fraction of the portfolio's nominal value (Jankowski and Jankowski, 2025).

The dynamic growth of the market is essentially attributable to two premises. The first is the increasing pressure on assignors, such as telecommunications operators, energy suppliers and financial institutions, which, by selling

Cite this Article as: Łukasz JANKOWSKI and Rafał JANKOWSKI Vol. 2026 (7) " Valuation of Bulk Debt Portfolios Based on Clustering and Historical Operational Data: Evidence from Poland " Communications of International Proceedings, Vol. 2026 (7), Article ID 4728226, https://doi.org/10.5171/2026.4728226

portfolios of overdue debts, seek to rapidly recover the funds committed and improve their financial liquidity. The second is the growing competition among portfolio purchasers, which have increasingly effective tools for conducting the debt collection process for mass debts, particularly at the amicable stage. In this situation, having a highly accurate valuation model becomes one of the key factors determining the financial performance of debt collection entities.

The valuation of a mass debt package consists in estimating the present value of future cash flows generated in the debt collection process, taking into account expected recoveries and the operating costs of conducting the debt collection process (Jankowski et al., 2024). Due to the large number of debts in a package, the individual analysis of each of them is not economically justified; therefore, probabilistic models are used in which the probability of repayment is estimated for groups of debtors with similar characteristics, and the result is aggregated at the level of the entire package. In practice, however, decisions regarding the valuation level of a mass debt package still rely to a considerable extent on expert judgement, which leads to simplifications and the replication of errors resulting from beliefs that are not confirmed by data. This problem has direct managerial implications, since overestimating the value of a portfolio entails the risk of financial loss, whereas underestimating it entails the loss of a transaction to competitors.

Machine learning constitutes a natural response to this gap, and interest in its application in the area of debt claims has been growing steadily for more than a decade. Kajdanowicz and Kazienko (Kajdanowicz and Kazienko, 2009) were among the first to apply debt claim clustering in combination with the forecasting of sequential repayment values, achieving an accuracy of over 93% on real data. Bellotti et al. (Bellotti et al., 2021) showed in a comprehensive comparative study that tree-based algorithms, in particular random forests and boosted trees, outperform traditional regression methods in forecasting the recovery rate from portfolios sold to debt collection entities, and that behavioural variables describing the history of contacts with the debtor significantly improve prediction accuracy. Jankowski and Jankowski (Jankowski and Jankowski, 2025) confirmed the superiority of the random forest method on a dataset of nearly 400,000 real mass debt claims from the Polish market, identifying nominal value, purchase price and debtor age as the strongest predictors of recovery. In the field of portfolio segmentation, Basiura et al. (Basiura et al., 2025) applied K-Modes clustering to a portfolio of more than 850,000 debt claims, assigning each cluster an optimal debt collection strategy on the basis of expert assessment and demonstrating that new portfolios can be rapidly classified into nto existing segments.

Existing studies, however, do not answer the key decision-making question faced by a debt collection entity, namely how much a new debt package is worth. Supervised models forecast the recovery rate for individual debt claims but omit debt collection costs as a component of valuation, while the approach based on segmentation ends with a recommendation of an operational strategy rather than a purchase price. The literature has not yet presented a coherent valuation model integrating clustering and historical operational data concerning both recoveries and costs.

This article fills this gap by proposing a model for the valuation of mass debt portfolios based on clustering and historical operational data. In this model, debt claims from a historical portfolio are grouped into homogeneous clusters using an unsupervised machine learning method, each cluster is assigned historical data on the average recovery rate and the costs of the debt collection process, and the valuation of a new debt claim results from assigning it to the appropriate cluster. The model is interpretable, operationally simple and, unlike previous approaches, explicitly accounts for the cost side as a component of valuation, which corresponds to the actual logic of the decision-making process in a debt collection entity. The article makes three contributions to the literature: it proposes an operationally useful approach to the automatic valuation of mass debt portfolios, integrates clustering with the costs of the debt collection process within a valuation model, and provides empirical results based on real data from the Polish mass debt market, which is poorly represented in the international literature. This situation results from the difficulty of obtaining, for research purposes, real data containing both information from the debt collection process and information collected as a result of activities related to acquiring debt packages.

The Mass Debt Collection Process and Its Significance for Valuation

The debt recovery process can be divided into three basic stages: amicable, judicial and enforcement debt collection (Jankowski and Paliński, 2024) (Kreczmańska-Gigol, 2015). Although this division is presented inconsistently in the literature in terms of the number and names of the stages, it remains the most common and intuitive one and, given the nature of the tools used and the institutions involved, is consistent with the recommendations of the Association of Financial Enterprises in Poland. The sequence of stages is not absolutely mandatory, since the creditor may omit amicable debt collection and refer the case directly to court proceedings; nevertheless, in the practice of servicing mass debts, the amicable stage is the starting point for the clear majority of cases, due to its relatively low unit cost compared with subsequent stages. Given the purpose of this article, the

mass debt collection process will be outlined only briefly in order to illustrate the issue. Readers interested in the details of modelling the debt collection process are encouraged to consult the author's other publications.

Amicable debt collection comprises all actions aimed at recovering amounts due without involving the apparatus of legal coercion, based on negotiations with the debtor and aimed at voluntary settlement of the obligation (Turaliński, 2013). In mass debt collection, this stage is based primarily on automated contact channels, such as SMS messages, e-mail correspondence and reminder letters, supplemented by telephone contact, which constitutes the principal negotiation tool. Voluntary payment of the amount due, either as a lump-sum payment or in the form of an agreed repayment schedule, ends the process at this stage. If amicable actions prove ineffective, the debt claim is referred to judicial proceedings. The decision to proceed to the judicial stage is a managerial decision in which the expected cost of further proceedings is weighed against the probability of recovery, which is directly related to the issue of portfolio valuation.

Judicial debt collection is aimed at obtaining an enforceable title constituting the basis for enforcement by a court enforcement officer. In Poland, in the case of mass debts, the dominant path is electronic writ-of-payment proceedings (EPU), which have operated since 2010 and are conducted exclusively by the District Court in Lublin. EPU enables the automation of preparing, filing and handling batches of statements of claim through the creditor's IT system, with a court fee amounting to a fraction of the basic fee provided for proceedings under the ordinary procedure. After a payment order is issued, the debtor has 14 days to pay the amount due or to file an objection. Filing an objection results in the case being transferred to the court of general jurisdiction (SWO), which significantly extends the proceedings and increases their costs. Once the payment order becomes final, the court appends an enforcement clause ex officio, which opens the way to enforcement by a court enforcement officer.

The enforcement debt collection stage, namely enforcement by a court enforcement officer, is initiated only at the creditor's request, after an enforceable title has first been obtained. The court enforcement officer conducts enforcement against the debtor's assets, with attempts to enforce against the bank account, preceded by checking the debtor in the OGNIVO system, and against remuneration for work or disability and retirement benefits usually being undertaken first in practice, due to their relatively low cost and high effectiveness (Jankowski and Paliński, 2024). Enforcement against real property or movable property requires additional expenditure on the creditor's part and is undertaken at the creditor's express request. The enforcement stage is the most costly and longest of the three stages of the process, and its effectiveness in the case of mass debts may be limited by the debtor's financial situation. Ineffective enforcement does not definitively close the case, since business practice provides for so-called re-enforcement, that is, the repeated initiation of enforcement after the debtor's economic situation has improved, as confirmed in the course of monitoring activities.

Each of the stages described above generates a different level of unit costs, which has a direct impact on the valuation of a debt portfolio. Amicable debt collection, based on automated contact channels, is characterized by the lowest unit cost. Judicial debt collection involves the costs of court fees and legal services, while the enforcement stage involves advances for court enforcement officers and the costs of asset searches. Consequently, the probability that a given debt claim will proceed to judicial or enforcement proceedings is a cost-intensity variable that the valuation model must take into account. This observation constitutes the basis of the approach proposed in this article, in which clusters of debt claims differ not only in expected recovery, but also in the historical debt collection process costs assigned to them. Consequently, at the cost modelling stage, these costs will be assigned at the level of each cluster.

Clustering in the Analysis of Debt Portfolios

Clustering is an unsupervised machine learning method whose objective is to divide a set of objects into homogeneous groups, or clusters, in such a way that objects belonging to the same cluster are as similar to one another as possible, while objects from different clusters are as dissimilar as possible. Unlike classification, which is a supervised method and assumes knowledge of predefined categories, clustering discovers the group structure hidden in the data without any prior knowledge of the number or nature of the groups. It is precisely this feature that makes it particularly useful in situations where rich historical data are available, but no predefined taxonomy of objects exists.

Among the available clustering algorithms, the most widely used is the k-means algorithm, proposed by MacQueen (1967) and algorithmically formalized independently by Lloyd (Lloyd, 1982). The algorithm iteratively assigns each object to the cluster with the nearest centroid and updates the centroids as the means of the objects in the cluster, minimizing the sum of squared within-cluster distances. Its main advantage is simplicity, scalability and low computational cost, which makes it an attractive tool for large operational datasets. It is also frequently used as a baseline algorithm in comparative analyses. Its limitations include the assumption of spherical

clusters, the need to define the number of groups k in advance, and sensitivity to categorical data, which require modifications of the algorithm, for example in the form of K-Modes, operating on modes rather than means.

Alternative approaches include hierarchical clustering methods, which build a dendrogram of object linkages in a bottom-up or top-down manner, and density-based methods, such as DBSCAN, which identify clusters as areas of high object density separated by sparse areas. An overview of clustering methods and their applications in financial data analysis is presented by Jin and Han (2011) in a machine learning publication.

The application of clustering in finance and credit risk management dates back to the 1980s and has been expanding systematically. The fundamental motivation is the observation that populations of debtors or debt claims are heterogeneous and cannot be modelled using a single predictive model without losing information about the group structure of the data. Clustering makes it possible to identify homogeneous segments for which separate, more precise predictive models can be built or differentiated action strategies can be assigned.

In the area of credit risk assessment, Baser, Koc and Selcuk-Kestel (Baser et al., 2023) proposed a credit risk classification method based on fuzzy k-means clustering, demonstrating that the segmentation of borrowers according to behavioural similarity improves the accuracy of predicting the probability of loan default. The authors emphasize that borrowers assigned to the same cluster exhibit similar behavioural patterns, which justifies the construction of separate rules assigned to each segment. A similar approach is analysed by Jadwal et al. (Jadwal et al., 2022), who compare k-means, hierarchical and HK-means algorithms in the segmentation of credit card customers, finding that the combination of clustering methods and predictive models outperforms models trained on an unsegmented dataset.

At a more methodologically advanced level, Carleo and Rocci (Carleo and Rocci, 2024) proposed a modification of the k-means algorithm, called K-RCurves, designed for clustering non-performing loan portfolios according to recovery curves over time. An NPL portfolio is segmented into homogeneous subsets, and a separate recovery curve is estimated for each cluster, taking into account data censoring and exposure weighting. The authors demonstrated that highly heterogeneous portfolios should be divided into clusters before the recovery profile is estimated, since averaged curves for the entire portfolio distort the picture of actual cash flows from individual groups of debt claims.

In the context of mass debt portfolio segmentation, Basiura, Jankowski and Jankowski (Basiura et al., 2025) applied the K-Modes algorithm to a portfolio of more than 850,000 debt claims described by categorical variables, such as the debtor's legal form, geographical region, debt claim value class and contact history. The identification of three clusters made it possible for industry experts to assign differentiated debt collection strategies to each of them, confirming that segmentation by means of clustering produces substantively meaningful and operationally useful groups. Importantly, the authors showed that new debt portfolios can be rapidly classified into existing segments, which enables the immediate implementation of dedicated debt collection procedures.

Previous applications of clustering in the area of debt claims have focused on two objectives: improving the prediction of the recovery rate through segment-based modelling and determining debt collection strategies for individual groups. None of the studies discussed, however, addresses the problem of using clustering directly for the automatic valuation of a mass debt portfolio, that is, for generating a quantitative recommendation of the purchase price based on assigning a new debt claim to the appropriate cluster on the basis of an existing model.

Model Assumptions

The approach proposed in this article fills this gap by combining clustering with historical operational data, including the average recovery rate and the debt collection process costs assigned to each segment. The logic of the model is based on the assumption that debt claims assigned to the same cluster exhibit a similar probability and repayment profile, as well as a similar course of the debt collection process, which makes it possible to extrapolate the cluster's historical financial parameters to new, previously unobserved debt claims. In this way, clustering serves as a link between the analysis of historical data and the purchase decision in a debt collection entity.

The clustering-based model for the valuation of mass debts using historical operational data is based on the following assumptions. The historical debt portfolio is divided into k homogeneous clusters ($k = 1, 2, ..., K$). Each cluster is characterized by its own recovery rate and average debt collection costs, determined on the basis of historical operational data for the debt claims assigned to that cluster. The value of the portfolio is determined as the discounted sum across all clusters. The discount rate r is a parameter adopted by the decision-maker and reflects the cost of capital and the assumed time horizon of the portfolio. Like any model, the proposed model also simplifies reality. In this case, a single common r is assumed, which accounts for the average horizon of the entire portfolio and is set by the decision-maker as a model parameter. Alternatively, discounting could be omitted

on the assumption that the process is sufficiently short for the effect of discounting to be negligible. In practice, the currently assumed horizon for conducting the debt collection process is 3 to 5 years under Polish conditions.

For each cluster k, two financial parameters are determined on the basis of historical operational data. The first is the recovery rate of cluster k, defined as the ratio of the total amount recovered to the total nominal value of the debt claims assigned to cluster k. This relationship is expressed by equation (1).

$$R_k = \frac{\sum_{i \in k} recovery_i}{\sum_{i \in k} FV_i} \tag{1}$$

Where: $recovery_i$ denotes the amount actually recovered from debt claim i, and FV_i denotes the nominal value of debt claim i.

The debt collection costs of cluster k, expressed by equation (2), constitute the ratio of total debt collection costs to the total nominal value of the debt claims assigned to cluster k.

$$C_k = \frac{\sum_{i \in k} cost_i}{\sum_{i \in k} FV_i} \tag{2}$$

Where: $cost_i$ denotes the total cost of the debt collection process incurred for debt claim i, including costs resulting from actions undertaken within the amicable, judicial and enforcement stages, in proportion to the actions actually undertaken.

Clusters in which amicable debt collection predominates will be characterized by a relatively low C_k. With regard to actions conducted within stages involving legal coercion, namely the judicial stage and the enforcement stage, it should be expected that these costs will be higher than the costs of the amicable stage. In business practice, however, the average cost of conducting debt collection activities is adopted without taking into account the stage at which repayment of the obligation was achieved.

Consequently, the value of the cash flows generated by the debt package will be expressed as in equation (3), as a discounted sum across all clusters.

$$P = \sum_{k=1}^{K} \frac{(R_k - C_k)}{1 + r} \tag{3}$$

Where: r is the discount rate adopted by the decision-maker, with one rate applied to the entire debt collection process.

The result P obtained from equation (3) represents the net present value of the cash flows generated by the debt portfolio and expresses the maximum purchase price at which the transaction remains profitable for the purchaser, assuming that the costs include all components. This value is indicative in nature and is expressed in relation to the unit nominal value of the portfolio, which means that, in order to obtain the purchase offer amount, P must be multiplied by the total nominal value of the debt claims included in the valued debt package. P therefore serves as an analytical reference point in the process of negotiating the purchase price; the decision-maker may adjust it by their own profit margin or by additional risk factors not captured by the model, such as uncertainty regarding data quality or changes in macroeconomic conditions.

The model assumes a uniform discounting horizon, which is an exogenous parameter set by the decision-maker on the basis of the debt collection entity's historical experience regarding the average recovery time for a given type of portfolio, for example 3 to 5 years. The result P, in turn, represents the objective value of the portfolio from the perspective of cash flows.

Data and Research Methodology

Source and Scope of the Data

The data used in the study come from a capital group that has operated continuously on the Polish debt market since 2001 and included four debt collection entities conducting debt collection at all stages of the process, covering both amicable collection and compulsory collection in the form of judicial and enforcement proceedings. The dataset covers a period of nearly two decades and consists of real operational data obtained as part of the activities of these entities. For the purposes of this study, the dataset was narrowed down to debt claims serviced within one of the managed securitization funds, comprising exclusively mass debts originating from telecommunications operators.

The original combined dataset from all entities in the group contained more than 1 million records. As a result of preprocessing, which included the removal of incomplete data, debt claims not having the character of mass debts, and records outside the scope of the managed securitization funds, the dataset was reduced to nearly 880,000 records. Data organization and processing were carried out using the MySQL database system. A total of 120,256 debt claims after the final stage of data cleaning were used for clustering.

It should be emphasized that the use of real operational data constitutes a significant added value of this study. Data generated in the debt collection process are business-sensitive, and entities operating on the debt market are reluctant to make them available for research purposes, fearing the disclosure of the operational strategies used, portfolio structures and achieved financial results. Obtaining such an extensive and homogeneous dataset of real data was possible only through direct cooperation with the entities of the group. Unlike synthetic datasets or public benchmark datasets commonly used in the literature, although such datasets are also difficult to obtain in the case of mass debt collection, real data reflect the full complexity and heterogeneity of a mass debt portfolio under the conditions of the Polish debt collection market, which translates directly into the practical usefulness and credibility of the results obtained.

Preparation of Variables for Clustering

The dataset prepared for clustering contained 29 variables, after excluding the *case_id* identifier, derived from six original attributes describing the debt claims.

Two variables were continuous and were normalized before clustering. The first was the nominal value of the debt claim, which was subjected to a logarithmic transformation in order to reduce the skewness of the distribution resulting from the presence of both small amounts due and high-value debt claims (*nominal_claim_amount_log_scaled*). The second was the debtor's age at the time of portfolio purchase, rescaled to a comparable range using the min-max method (*age_scaled*).

The remaining four attributes were categorical and were transformed using one-hot encoding, which consists in replacing each category with a separate binary variable taking the value of 1 for a given category and 0 otherwise. The use of this method is necessary because distance-based algorithms such as k-means operate on the Euclidean metric and require numerical variables, while directly encoding categories as integers would suggest a false ordering between categories, for example that a joint-stock company is "larger" than a limited liability company.

The debtor entity type was encoded using three binary variables: company, natural person and missing data. The debtor's gender was encoded analogously using three variables: female, male and missing data. The legal form of the entity (*legal_form*) was encoded using ten binary variables covering an association, civil law partnership, registered partnership, joint-stock company, limited liability company, limited partnership, natural person, sole proprietorship, other forms and missing data. The geographical region was encoded using eleven binary variables corresponding to ten regional areas of Poland and the missing data category.

The missing data category (*MissingData*) was included as a separate value in all categorical variables instead of removing these observations, which made it possible to preserve the full sample size of 120,256 debt claims and, at the same time, to identify clusters composed of debt claims with incomplete documentation, which is of practical significance in the valuation of a mass debt package.

Clustering Methods Applied

Two clustering methods were applied in the study: k-means and Ward's hierarchical method. Both methods were tested for the number of clusters k ranging from 2 to 10, and k = 3 was ultimately adopted. The k-means algorithm, proposed by MacQueen (1967) and formalized by Lloyd (Lloyd, 1982), consists in iteratively assigning each observation to the cluster with the nearest centroid and then updating the centroids as the means of the observations in a given cluster. The optimization criterion is the minimization of the sum of squared distances of observations from the centroids of their clusters, referred to as inertia or WCSS (*Within-Cluster Sum of Squares*). The algorithm was run with a random seed (47) ensuring the replicability of the results. The study used the k-means implementation from the scikit-learn library.

Ward's hierarchical method (Ward, 1963) builds a dendrogram by successively merging pairs of clusters in such a way as to minimize the increase in the sum of squared within-cluster deviations after each merger. This method does not require the number of clusters to be determined a priori; clustering proceeds from n single-element clusters to one cluster containing all observations, and the division into k clusters is obtained by cutting the dendrogram at the appropriate level. The study used the Euclidean metric, which is the standard choice for Ward's method.

Results of the Empirical Study

The values of k, that is, the number of clusters, from 2 to 10 were tested. For each k, inertia and the silhouette coefficient were calculated. The results for each of the considered numbers of clusters are presented in Table 1.

Table 1. Results of the assessment of k-means clustering quality for different numbers of clusters

k	Inertia (WCSS)	silhouette index
2	375510.486921	0.257585
3	**275668.939507**	**0.306535**
4	255937.096713	0.276615
5	227483.294788	0.233793
6	215436.246601	0.227952
7	203303.811432	0.216028
8	195903.180562	0.181268
9	189706.975366	0.175681
10	184182.131074	0.179475

Inertia decreased monotonically as *k* increased, while the elbow method analysis presented in Figure 1 indicated k = 3 as the point of inflection of the curve. The silhouette coefficient reached its maximum for k = 3 and then decreased systematically. The development of the silhouette index in relation to the number of clusters is presented in Fig 2.

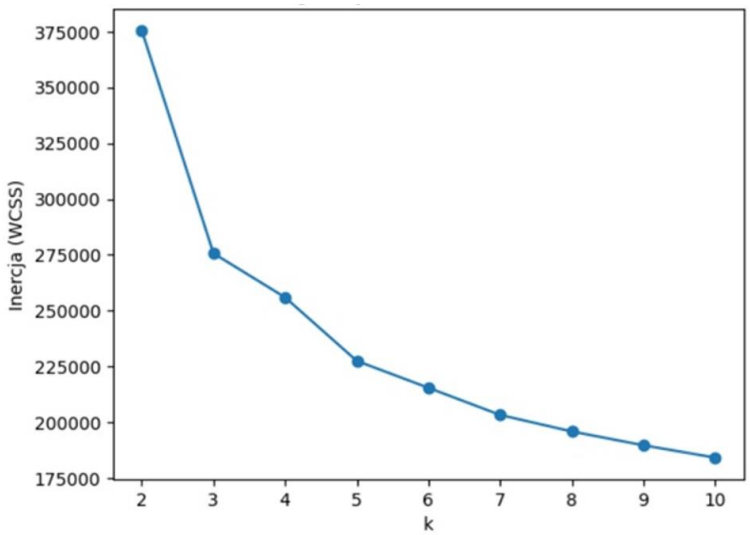

Fig 1. Determining the optimal number of clusters using the elbow method

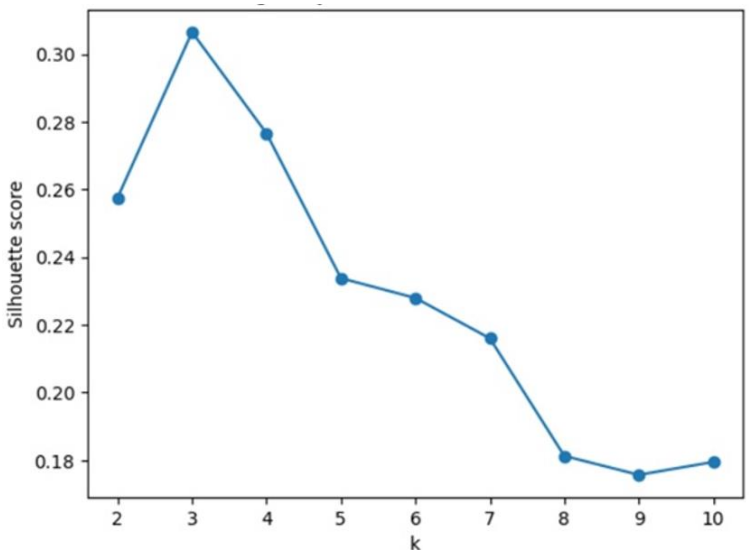

Fig 2. Analysis of the silhouette coefficient as a function of the number of clusters

Fig 3 presents the dendrogram obtained as a result of clustering using Ward's hierarchical method.

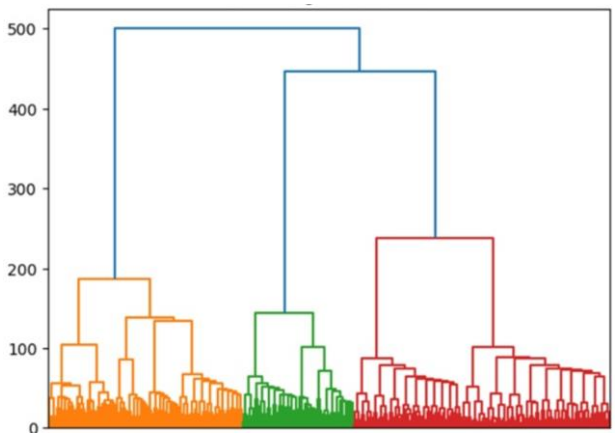

Figure 3. Dendrogram of hierarchical clustering using Ward's method

The dendrogram in Figure 3 was obtained by hierarchical clustering using Ward's method. The horizontal axis represents the analysed observations, while the vertical axis represents the linkage distance, reflecting the level of dissimilarity at which individual clusters are merged. 6.2. Comparison of the Results of K-Means and Ward's Method

The clustering analysis demonstrated that the mass debt portfolio is not homogeneous, but contains a clearly identifiable segmentation structure. Solutions comprising from two to ten clusters were tested using, among other methods, the elbow method and the silhouette coefficient. The highest value of the silhouette coefficient was obtained for the solution comprising three clusters ($k = 3$), for which the value of the indicator was 0.3065. At the same time, the analysis of WCSS values indicated a clear decrease in inertia between the $k = 2$ and $k = 3$ solutions, after which further increasing the number of clusters led to smaller cognitive gains. Consequently, the three-cluster solution was adopted as the best compromise between statistical quality, model simplicity and the possibility of economic interpretation, while preserving business relevance.

Table 2. Comparison of clustering quality using k-means and Ward's methods

Metryka	K-means	Ward
Silhouette	0.3065	0.3065
Davies-Bouldin	1.3068	1.3068
Calinski-Harabasz	49032.9299	49032.9299

The quality of the final segmentation was assessed using three internal measures of clustering quality: the silhouette coefficient, the Davies–Bouldin index and the Calinski–Harabasz index. For the k = 3 solution, the silhouette coefficient was 0.3065, the Davies–Bouldin index was 1.3068 and the Calinski–Harabasz index was 49k. These values indicate moderate but clear cluster separation. This is not ideal separation, which is typical of real economic data, particularly data covering mass debt portfolios characterized by high heterogeneity. At the same time, the measures obtained confirm that the division is not random and may serve as a basis for further segment-level interpretation.

The comparison of the results obtained using k-means and Ward's hierarchical method indicated very high agreement in the assignment of observations to clusters. The value of ARI = 0.956 means that both methods lead to an almost identical segmentation structure. This result is of significant methodological importance, as it confirms the stability of the obtained division regardless of the algorithm applied. This means that the identified cluster structure is not an artifact of a single computational method, but reflects actual similarities between debt claims in the analysed dataset. Table 3 presents the size distribution of the resulting clusters.

Table 3. Distribution of the number of observations across clusters.

Cluster	Number of debt claims
1	41717
2	23771
3	54768

The analysis of cluster characteristics indicates that the segmentation obtained differentiates the portfolio primarily by debtor type, legal form, debtor age and nominal value of the debt claim.

Cluster 1 comprised 41,717 observations, accounting for 34.7% of the sample. This segment was characterized by an above-average share of business entities (*entity_type_Company*, difference from the mean: 0.653) and sole proprietorships (*legal_form_SoleProprietorship*, 0.531). At the same time, natural persons (*entity_type_Person*, -0.653) and the legal form of a natural person (*legal_form_NaturalPerson*, -0.653) were below the mean. The cluster was also distinguished by a higher-than-average value of the variable *nominal_claim_amount_log_scaled* (0.513), which indicates relatively higher nominal values of debt claims after logarithmic transformation and scaling. This segment may be interpreted as a group of debt claims against entrepreneurs, in particular those operating sole proprietorships, with a relatively higher nominal value of debt claims.

Cluster 2 comprised 23,771 observations, that is, 19.8% of the sample, and was the smallest of the identified segments. Its most distinctive feature was a very high level of the variable *age_scaled* (1.740), together with a lower-than-average value of *nominal_claim_amount_log_scaled* (-0.456). Natural persons (*entity_type_Person*, 0.346) and the legal form of a natural person (*legal_form_NaturalPerson*, 0.346) occurred in this cluster more frequently than on average, whereas the share of business entities was below the mean (*entity_type_Company*, -0.346). This segment may be interpreted as a group of natural persons characterized by a higher level of the age variable and a relatively lower nominal value of debt claims.

Cluster 3 was the largest segment and comprised 54,768 observations, that is, 45.5% of the sample. It was distinguished primarily by a lower-than-average level of the variable *age_scaled* (-0.697). As in Cluster 2, natural persons (*entity_type_Person*, 0.346) and the legal form of a natural person (*legal_form_NaturalPerson*, 0.346) occurred in it more frequently than on average. At the same time, business entities (*entity_type_Company*, -0.346) and sole proprietorships (*legal_form_SoleProprietorship*, -0.282) were below the mean. This cluster may be

interpreted as the largest portfolio segment, comprising natural persons with a lower level of the age variable relative to the mean for the analysed debt claims portfolio.

Overall, the results indicate that clustering for k = 3 does not create random groups, but identifies segments that differ in economically and operationally significant characteristics. The most important variables differentiating the clusters are *entity_type_Company*, *entity_type_Person*, *legal_form_NaturalPerson*, *legal_form_SoleProprietorship*, *age_scaled* and *nominal_claim_amount_log_scaled*. Such a division is consistent with the objective of the valuation model, since debtor type, legal form, age and nominal value of the debt claim may affect both expected recovery and the costs of conducting the debt collection process at each stage.

Ultimately, although both k-means and Ward's method led to very similar results, k-means may be adopted as the final method. The justification is not statistical superiority, but primarily the operational usefulness of this method. K-means makes it possible to easily assign new debt claims to existing clusters on the basis of their distance from the centroids, which is crucial in the practical application of the model to the valuation of new portfolios. Thus, this method better corresponds to the applied objective of the study, which is not merely to describe the structure of the historical dataset, but to create a tool supporting the ongoing valuation of mass debts.

From a practical perspective, the results obtained indicate that applying a uniform approach to the entire debt portfolio may lead to excessive simplifications. The identification of three segments makes it possible to treat the portfolio as a structure composed of groups of debt claims with different feature profiles, which may support more precise assignment of expected recoveries and debt collection process costs. In the context of the proposed valuation model, this means that financial parameters, such as the recovery rate and debt collection costs, may be estimated at the cluster level rather than for the valued portfolio as a whole. This approach increases the interpretability of the model and better reflects the actual decision-making logic of debt collection entities or securitization funds acquiring mass debt portfolios.

For practical purposes, the average level of recoveries was assigned to and then calculated for each of the resulting clusters. The same procedure should be applied to costs separately in each cluster. For the purposes of the developed model, a simplification was adopted, consisting in the assumption that, in the further analysis, costs amounted to 20% relative to the nominal value. The obtained values of the average level of recoveries for each cluster are presented in Table 4.

Table 4. Comparison of the average level of recoveries across clusters.

Cluster	Average level of recoveries
1	21,94%
2	45,56%
3	10,34%

Discussion, Managerial Implications and Limitations

The clustering results obtained and the financial parameters of the clusters determined on their basis make it possible to formulate a number of conclusions of both methodological and practical significance from the perspective of a debt collection entity or a securitization fund.

The proposed model makes it possible to determine the present value of cash flows from a debt portfolio on the basis of historical data, without the need to build separate predictive models for individual debt claims, which, as a rule, not only does not prove effective, but is also not widely practised with respect to mass debts. This approach is particularly important in the operational conditions of debt collection entities, where purchase decisions must be made quickly and analytical resources are limited. Assigning a new debt claim to the appropriate cluster on the basis of its distance from the centroid requires knowledge of only a few descriptive variables concerning the debtor, which makes the model easy to implement in existing IT systems.

From a managerial perspective, a key feature of the model is the possibility of directly assessing the operating profitability of individual portfolio segments. The difference between R_k and C_k determined for each cluster informs the decision-maker, among other things, which groups of debt claims generate a surplus over the debt collection costs incurred and which are operationally unprofitable even at a zero purchase price. This information has a direct impact on the entity's purchase strategy. Portfolios dominated by clusters with a negative difference between R_k and C_k should be rejected, provided that the assignor allows the purchase of only part of the debt

package, or valued with an appropriately higher discount, whereas portfolios concentrating debt claims from high-recovery clusters constitute an acquisition target of increased attractiveness.

Explicitly accounting for debt collection process costs C_k at the level of each cluster constitutes a key advantage of the model over approaches based solely on forecasting the recovery rate. As shown in the section devoted to debt collection processes, unit costs differ significantly depending on the stage at which repayment of the obligation occurs; amicable debt collection is many times less costly than judicial or enforcement proceedings. Clusters comprising debt claims with a low propensity of debtors to repay voluntarily will therefore be characterized by a higher C_k, which, with a comparable R_k, significantly reduces their value for the purchaser. Omitting this relationship, which is typical of models forecasting only the recovery rate, leads to a systematic overestimation of the valued debt package.

The result P obtained from equation (3) constitutes the present value coefficient of net cash flows in relation to the unit nominal value of the portfolio. To obtain the absolute amount of the purchase offer, P should be multiplied by the total nominal value of the debt claims included in the valued package. The amount determined in this way constitutes the upper limit of the purchase price at which the transaction remains profitable; offering a higher price means incurring a loss already at the moment of concluding the transaction. The decision-maker may adjust this value by their own profit margin, risk aversion or qualitative factors not captured by the model. It should be emphasized, however, that discounting in equation (3) refers to the R_k and C_k coefficients, and not to actual cash flows distributed over time, which means that this operation is a formal simplification. The model is adaptive, with the parameters R_k and C_k updated as the historical database expands, which makes it possible to account for changes in the effectiveness of the debt collection process and in market conditions. As the debt collection entity accumulates data on new portfolios, cluster centroids may be periodically recalculated and financial parameters updated, without the need to rebuild the model architecture. This constitutes a significant advantage over approaches requiring a supervised model to be retrained each time on new labelled data.

The approach presented is not without limitations. First, the adopted uniform discount rate r does not differentiate the time horizon of individual clusters, although in practice debt claims from different segments may be characterized by different recovery time profiles. Second, the model assumes the stability of the cluster structure over time, which may not hold in the case of significant changes in the composition of acquired portfolios or in the macroeconomic environment. Third, valuation quality is directly dependent on the representativeness of the historical database. Portfolios with a profile that differs significantly from the training data may be valued with lower accuracy.

Conclusions

This article presents a model for the valuation of mass debt portfolios based on clustering and historical operational data, constituting a response to the gap identified in the existing literature on the subject. Previous approaches have focused either on forecasting the recovery rate for individual debt claims while omitting debt collection costs, or on portfolio segmentation for operational purposes without generating a quantitative recommendation of the purchase price. The proposed model integrates both elements, determining the value of the portfolio as the discounted sum of the differences between the historical recovery rate and debt collection costs for each identified cluster.

The empirical study conducted on a dataset of more than 120,000 real mass debt claims from the Polish market showed that the portfolio is not homogeneous, but contains a clear segmentation structure that can be represented using clustering methods. The three-cluster solution, selected on the basis of the silhouette coefficient and elbow analysis, achieved high agreement between the k-means method and Ward's hierarchical method (ARI = 0.956), confirming the stability and reproducibility of the obtained segmentation. The identified clusters differ in the economic and demographic profile of debtors, which translates into different financial parameters of the segments and justifies differentiating valuation at the cluster level rather than for the entire portfolio jointly.

The model is characterized by three properties that are particularly important from a practical perspective: interpretability, operational simplicity and explicit inclusion of debt collection process costs as a valuation component. The model result, including the coefficient P, constitutes the upper limit of the profitable purchase price of the portfolio expressed in relation to its total nominal value, which makes it a directly useful tool supporting purchase decisions of debt collection entities and securitization funds.

A separate scientific value of this study lies in the fact that the analyses are based on real operational data from the Polish mass debt market. Data generated in the debt collection process are business-sensitive, and entities operating in this market are reluctant to make them available for research purposes, fearing the disclosure of operational strategies used, the structure of managed portfolios and achieved financial results. The results of this

article are therefore rooted in the operational realities of the Polish debt collection market, which translates directly into their practical usefulness for debt collection entities and securitization funds considering the implementation of automated tools supporting the valuation of mass debt portfolios as part of the decision-making process.

The results obtained open several important directions for further research, the pursuit of which would allow the proposed approach to be significantly enriched and refined.

The most important extension of the model would be to replace the average recovery rate with recovery curves determined in successive periods and clusters, for example in 30-day intervals. Such an approach would make it possible to capture the dynamics of repayments over time. In the practice of managing a debt collection entity, recovery does not occur once, but is distributed unevenly over the entire horizon of the process, with the largest part usually concentrated in the first months after the purchase, or commencement of the debt collection process, of the debt package, and gradually decreasing in subsequent periods. Determining recovery curves separately for each cluster would also make it possible to replace simplified coefficient discounting with proper discounting of actual cash flows in successive time intervals, thereby giving the present value analysis a fully-fledged character. An analogous approach may be applied on the cost side by determining cost profiles over time for each cluster, which would make it possible to account for the fact that the costs of the judicial and enforcement stages are incurred significantly later than the costs of amicable debt collection.

Another direction for further research is the dynamic updating of the model in response to changes in the structure of the debt market. The current version of the model assumes the stability of clusters and of the financial parameters assigned to them over time, which may not hold under conditions of significant changes in the macroeconomic or regulatory environment, or in the composition of acquired portfolios. Developing a mechanism for periodic recalculation of centroids and updating of recovery and cost curves would increase the model's robustness to such changes.

Acknowledgment

Funding: The APC was funded under subvention funds for the Faculty of Management and by the program "Excellence Initiative–Research University" for the AGH University of Krakow.

References

- Baser, F., Koc, O., Selcuk-Kestel, A.S., 2023. Credit risk evaluation using clustering based fuzzy classification method. Expert Syst. Appl. 223, 119882. https://doi.org/10.1016/j.eswa.2023.119882
- Basiura, B., Jankowski, Ł., Jankowski, R., 2025. Unsupervised Clustering of Debt Portfolios Followed by Expert Strategy Allocation: An Exploratory Data Analysis. Commun. Int. Proc. https://doi.org/10.5171/2025.4540525
- Bellotti, A., Brigo, D., Gambetti, P., Vrins, F., 2021. Forecasting recovery rates on non-performing loans with machine learning. Int. J. Forecast. 37, 428–444. https://doi.org/10.1016/j.ijforecast.2020.06.009
- Carleo, A., Rocci, R., 2024. Functional clustering of NPLs recovery curves. Socioecon. Plann. Sci. 95, 102018. https://doi.org/10.1016/j.seps.2024.102018
- Jadwal, P.K., Pathak, S., Jain, S., 2022. Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis. Microsyst. Technol. 28, 2715–2721. https://doi.org/10.1007/s00542-022-05310-y
- Jankowski, Ł., Jankowski, R., 2025. Modelling Recovery Rates in Mass Debt Portfolios Using Ma-chine Learning Algorithms. Commun. Int. Proc. https://doi.org/10.5171/2025.4648525
- Jankowski, R., Paliński, A., 2024. Zarządzanie procesem windykacji wierzytelności masowych: integracja uczenia maszynowego z wiedzą ekspercką. AGH University Press. https://doi.org/10.7494/978-83-68219-29-6
- Jankowski, R., Paliński, A., Jankowski, R., Paliński, A., 2024. Debt Collection Model for Mass Receivables Based on Decision Rules - A Path to Efficiency and Sustainability. Sustainability 16. https://doi.org/10.3390/su16145885
- Jin, X., Han, J., 2011. K-Means Clustering, in: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Ma-chine Learning. Springer US, Boston, MA, pp. 563–564. https://doi.org/10.1007/978-0-387-30164-8_425

- Kajdanowicz, T., Kazienko, P., 2009. Prediction of Sequential Values for Debt Recovery, in: Bayro-Corrochano, E., Eklundh, J.-O. (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 337–344. https://doi.org/10.1007/978-3-642-10268-4_40
- Kreczmańska-Gigol, K., 2015. Windykacja polubowna i przymusowa. Proces, rynek, wycena wierzytelności. Difin SA.
- Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489
- Turaliński, K., 2013. Windykacja gospodarcza. e-bookowo.
- Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845
- ZPF, 2026. Wielkość polskiego rynku wierzytelności - Raport ZPF. ZPF - Związek Przedsiębiorstw Finans. URL https://zpf.pl/wielkosc-polskiego-rynku-wierzytelnosci/ (accessed 2026-04-11).