*stats*

MDPI

*Communication*

# Beyond Expectations: Anomalies in Financial Statements and Their Application in Modelling

**Roman Blazek * and Lucia Duricova**

Department of Economics, Faculty of Operation and Economics of Transport and Communications, University of Zilina, Univerzitna 1, 010 26 Zilina, Slovakia; lucia.duricova@uniza.sk
* Correspondence: roman.blazek@uniza.sk

**Abstract**

The increasing complexity of financial reporting has enabled the implementation of innovative accounting practices that often obscure a company's actual performance. This project seeks to uncover manipulative behaviours by constructing an anomaly detection model that utilises unsupervised machine learning techniques. We examined a dataset of 149,566 Slovak firms from 2016 to 2023, which included 12 financial parameters. Utilising TwoSteps and K-means clustering in IBM SPSS, we discerned patterns of normative financial activity and computed an abnormality index for each firm. Entities with the most significant deviation from cluster centroids were identified as suspicious. The model attained a silhouette score of 1.0, signifying outstanding clustering quality. We discovered a total of 231 anomalous firms, predominantly concentrated in sectors C (32.47%), G (13.42%), and L (7.36%). Our research indicates that anomaly-based models can markedly enhance the precision of fraud detection, especially in scenarios with scarce labelled data. The model integrates intricate data processing and delivers an exhaustive study of the regional and sectoral distribution of anomalies, thereby increasing its relevance in practical applications.

**Keywords:** anomaly index; creative accounting; centroids; unsupervised learning

check for updates

Academic Editor: Wei Zhu

Received: 13 May 2025
Revised: 9 July 2025
Accepted: 11 July 2025
Published: 15 July 2025

**Citation:** Blazek, R.; Duricova, L. Beyond Expectations: Anomalies in Financial Statements and Their Application in Modelling. *Stats* **2025**, *8*, 63. https://doi.org/10.3390/stats8030063

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

The main aim of this project is to create and evaluate an unsupervised anomaly detection model that can identify companies exhibiting unusual financial behaviour using only quantitative metrics. Most current research on financial fraud and creative accounting depends on labelled datasets and supervised learning methods; however, such datasets are few or incomplete in practical scenarios, especially with small and medium-sized firms (SMEs).

This paper fills the research gap by providing a scalable, interpretable model that does not necessitate pre-labelled fraudulent cases. The primary hypothesis posits that financial anomalies are not randomly allocated but rather tend to cluster among particular industries, regions, and legal structures, indicating inherent structural tendencies in financial reporting methodologies. This research is novel in its integration of robust clustering approaches with eight years of longitudinal data, facilitating a more nuanced, context-aware detection method that can enhance audit planning and policy formulation.

### *Inspiration and Previous Investigations on Fraud Detection*

The detection of fraudulent financial conduct has advanced considerably in recent decades. Initial methodologies predominantly depended on rule-based models, exemplified

by the Beneish M-Score [1], which employed established financial measures to identify probable manipulation. These models provided openness and interpretability, although they were constrained by fixed thresholds and vulnerability to evasion.

Subsequent innovations employed statistical and machine learning techniques. One study employed decision trees and logistic regression within the Romanian context [2], while another assessed the effectiveness of Beneish models across various Slovak sectors [3]. A subsequent study advanced the discourse by analysing novel accounting methodologies within the Spanish Professional Football League, emphasising industry-specific challenges, and demonstrating how regulatory deficiencies facilitate earnings manipulation in sports financial reporting [4].

Traditional models, while beneficial, face limitations due to their reliance on labelled data and their restricted ability to detect new fraud patterns. Unsupervised learning methods, such as clustering, are increasingly popular because of their adaptability and capacity to detect undiscovered anomalies.

A recent study highlighted the increasing significance of internal and external auditors, along with stakeholders, in identifying creative accounting practices [5]. Their examination of UAE enterprises brought home the importance of forensic audits, stakeholder involvement, and structural audit autonomy as integral components to algorithmic detection methodologies.

Recent advancements highlight the imperative for model interpretability. A study proposed a way to predict creative accounting using easy-to-understand machine learning models like Random Forest and XGBoost, along with SHAP values to make the results clearer [6]. Research showed that adding explainability to predictive models helps auditors and analysts understand why certain issues are found, making the results of algorithms more useful in corporate governance. Although their research relies on supervised learning and labelled instances, our study advances this field by exploring unsupervised detection, hence enabling wider use in data contexts where fraud labels are either absent or incorrect.

Our research advances this developing domain by employing unsupervised machine learning techniques (TwoStep and K-means clustering) on an extensive financial dataset from Slovak firms. We intend to identify suspected organisations by computing an anomaly index derived from deviations within behavioural clusters without relying on direct fraud labelling. This strategy fills a methodological void in the fraud detection literature, especially in data contexts marked by opacity and restricted regulatory transparency. This study aims to determine if financial anomalies systematically cluster across sectors and regions and if clustering methods can function as scalable tools for anomaly detection without fraud labels.

## 2. Literature Review

At the birth of accounting, a tool that helped businessmen become victims as the beginnings of manipulating economic records began to emerge. We can imagine the beginnings of manipulation when various records were impaired due to spilt ink in the accounting books. Believe that these manipulative origins are the source of the term camouflage [7]. This trend stretches back centuries to the present day, when we no longer encounter spilt ink but much more sophisticated techniques for devaluing a company's economic information. A new term, 'creativity' or 'creative accounting', has replaced the concept of camouflage. Accounting is a fragile tool that can be misused very quickly for its own good. The constant changes in the business environment, the impact of economic crises, the emergence of new types of property, forms of foreign resources, legislation, the impact of fiscal policy, and the constant expansion of the grey and black economy have transformed accounting into a vulnerable apparatus that relies on the intervention of a higher power. Creative

accounting is linked to creative activity that generates original and atypical ideas. An article critiquing the US GAAP (United States Generally Accepted Accounting Principles) was published, contending that the accounting framework facilitated companies in manipulating financial records to retain shareholders [8]. Research subsequently corroborated this perspective, describing such manipulation as a technique to align reported profits with owner preferences [9]. The phrase "creative accounting" emerged more prominently in the 1970s, marked by an increase in pertinent publications by economists, accountants, and financial analysts [10]. A thorough study further developed the concept, presenting synonymous terms such as earnings management, income smoothing, big bath accounting, and window dressing [11]. In various regions, creative accounting practices were characterised by metaphorical phrases: "the art of faking a balance sheet" [12], "calculating the benefits" [13], "presenting a balance sheet" [14], and "saving money" [15].

Recent scholarly research has enhanced the comprehension of creative accounting by examining its definitions, mechanisms, and motivating factors. Certain studies emphasise the ambiguous interpretation of creative accounting, frequently perceived as manipulation but sometimes linked to innovation [16]. Additional studies define the particular accounting tactics employed to manipulate financial statements, emphasising both individual and market-driven incentives for these behaviours [17]. Subsequent analyses differentiate creative accounting from blatant fraud, indicating that although it may adhere to accounting standards, it can still undermine the objective of accurately portraying a company's financial status [18]. Frequently referenced reasons include tax optimisation, earnings stabilisation, and alignment with investor expectations [19]. Research underscores the significance of psychological and organisational factors influencing the use of creative techniques by highlighting leadership styles, ethical ambiguity, and internal performance pressure as pivotal drivers of such behaviour.

Creative accounting can be defined through various conceptual frameworks. Certain studies clarify this term from an accounting perspective, emphasising its representation of various methodologies for addressing conflicts between competing financial results intended for presentation and the underlying transactions [20]. A key feature of creative accounting is that it intentionally strays from normal accounting rules, avoiding established regulations to reach a desired reporting outcome [21]. More research shows that these practices often happen when companies try to change accounting from the legal format to one that fits their management goals [22]. A dual-level interpretation has been suggested [23], with the first level pertaining to initiatives designed to address new economic occurrences that remain unregulated by current accounting standards. The second level generally defines this term as activities that lead to the manipulation of financial statements. The work Creative Financial Accounting: Its Nature and Use articulates an academic perspective wherein creative accounting is characterised by the alteration of financial data from its original recorded state to a representation that aligns with the desired image of company proprietors—accomplished either through the manipulation of permissible policies or by selectively disregarding specific regulations [24]. Another contribution to the discourse offers a critical evaluation, indicating that while creative accounting may enable firms to influence financial results, such manipulation does not inherently lead to economic advantages. Conversely, over a prolonged timeframe, such practices may negatively impact the company's performance and sustainability [25]. The motivations for creative accounting are varied. Alongside traditional motivations like tax minimisation and shaping investor perceptions, prior research [17] has revealed systematic behavioural patterns among financial professionals, affirming a common rationale that includes income smoothing and achieving established performance targets. Additional research [26,27] further links creative accounting to individual motivations, inadequate internal controls,

and ambiguities within the legal framework. One study contends that creative accounting should not invariably be viewed in a strictly negative light, consistent with the viewpoint articulated in [25]. The primary conclusion is that organisations can improve their reported performances as long as they adhere to legal accounting standards. In such instances, the actions, while innovative, remain lawful. Another contribution [26] outlines two different views on creative accounting: in the U.S., it is often seen as related to fraud, while in the U.K., it is viewed as a legal way to use rules flexibly—unless it crosses into dishonest practices [28]. Jones's simplified definition is even more stringent when defining creative accounting, asserting that it applies to any business that does not adhere to the fundamental principles of accounting, thereby failing to maintain a faithful and authentic image [29]. Consequently, numerous nations regard creative accounting as straddling legal and regulatory ambiguities. A study [30] indicates that the likelihood of accounting manipulation escalates with the size of the company. In contrast, the influencing elements were later categorised into two primary groups [31]: external factors, encompassing government revenues and regulations, and internal factors related to the firms themselves. Another study [32] contests this conclusion, demonstrating no significant correlation between firm size and manipulation. The aforementioned study identifies cultural, social, and legal contexts as more pivotal factors influencing the decision to engage in manipulation.

The first group consists of factors that influence the enterprise through state revenues, and the second group consists of factors of the enterprises themselves. Nevertheless, focusing solely on firm size, legal structure, or country of operation encompasses only a limited range of potential factors influencing accounting manipulation. To mitigate this limitation, a study [22] performed an industry-specific analysis, concentrating on firms within a particular sector. The researchers sought to ascertain if the gender of a company's CEO influences the propensity for manipulation, considering that strategic decisions related to financial reporting generally emanate from senior management roles. Research has demonstrated that women exhibit a conservative approach to manipulation, whereas men exhibit a high propensity for manipulation. This research also revealed that mixed leadership falls into the category of potential manipulators, resulting in ambiguous and unclear outcomes.

Creative accounting is not limited to private enterprises; governments also partake in such practices. Numerous studies have examined the phenomenon of creative accounting at the governmental level [32–37]. However, understanding the results about how states manipulate their finances is challenging because different hidden factors—such as political pressures, economic changes, or off-balance-sheet activities —can create confusing signals or hide what is really happening. Bibliometric analyses [38] indicate that research on creative accounting has markedly increased recently, garnering heightened interest across multiple academic fields. Notwithstanding this expansion, a significant gap remains in country-specific research, especially in Eastern Europe and developing economies—a shortcoming this paper seeks to rectify. The latest bibliometric analysis [39] has tracked the growth of creative accounting research in academic databases, highlighting an increasing focus on different fields, especially in agricultural and rural financial management. Their findings emphasise the significance of creative accounting in corporate settings and public and sectoral reporting frameworks. Slovak authors deal, in most cases, with a simpler but much more important part of the group that could manipulate accounting, as it is the main driver of the GDP of each state. Most Slovak research on creative accounting focuses on the enterprise level due to its importance as a primary contributor to national GDP. A study examined the phenomenon of creative accounting in the transport sector among the Visegrad Four nations [40]. A separate analysis investigated the aggregation of firms within the Slovak agribusiness sector, which has traditionally

demonstrated a significant prevalence of accounting manipulation [41]. The same authors further expanded their analysis to encompass the construction industry [41]. The authors of [3] further examined the issue of financial disclosure in the transport sector, highlighting the risks associated with underreporting and data distortion. The COVID-19 pandemic, as noted in [42], caused significant disruption and instability for businesses of all sizes, prompting the adoption of smart sensors for real-time financial performance monitoring. A related study [43] examined publicly traded companies in the Visegrad Four, focusing on their utilisation of earnings management within the framework of shared corporate responsibility and behavioural tendencies in financial reporting. A recent study focused on identifying how corporations manage their earnings using Kasznik's model [44]. In addition to this approach, the authors employed correspondence analysis, a visual method well-suited for illustrating patterns in financial data. Benford's law was also applied in their analysis to detect earnings dilution based on the frequency and order of numerical digits [44]. According to [45], the Beneish model remains one of the most reliable tools for uncovering creative accounting practices. Firms often attempt to present manipulated results as quickly as possible in order to attract investors or fulfil reporting benchmarks imposed by regulatory or international frameworks.

In another study, researchers examined the extent to which earnings management facilitates the rapid achievement of strategic business goals [46]. This issue was further analysed with an emphasis on corporate indebtedness within Slovak firms [47]. Their results revealed a notable relationship between debt management and earnings manipulation. When a firm alters the true values of its indicators to align with targeted outcomes, as observed in the work of [48], the distortion becomes evident in the firm's overall financial condition.

Researchers have formulated numerous models to identify companies engaged in creative accounting practices. Fundamental contributions to this domain encompass the work of [1,49,50], who introduced models for identifying manipulation in financial reporting. The Jones model has emerged as one of the most prevalent linear regression-based instruments for detecting earnings management. It employs variations in sales and long-term assets as explanatory variables and introduces the notion of accruals, which are categorised into discretionary and non-discretionary components. By accounting for variations in receivables, the model subsequently evolved into the Modified Jones model [50]. The Beneish M-Score is another significant method that uses eight financial ratios to evaluate the probability of financial manipulation. Various empirical contexts have validated its robustness. For instance, reference [51] highlighted the extensive economic ramifications of financial manipulation, proposing that an increase in fraudulent reporting could lead to macroeconomic instability, including a decline in GDP. Consequently, there is a distinct necessity to create and enhance instruments that can promptly detect manipulative practices and safeguard both markets and investors. As indicated in [52], linear regression continues to be the primary method employed in fraud detection models, forming the basis of numerous existing approaches. The Modified Jones model is considered one of the most effective instruments for identifying financial irregularities, as noted in references [45,53]. This assertion is corroborated by findings from [54], which validated its superior performance across various contexts. Nevertheless, certain scholars have advocated for the adoption of more adaptable, non-linear methodologies. Research by [55,56] highlights the inadequacies of linear models in representing complex connections and recommends models that incorporate non-linearity. For example, ref. [57] used discriminant analysis to improve the accuracy of the Beneish model. Expanding upon the Beneish framework, ref. [45] further enhanced the model by integrating indicator values across three successive years to augment temporal precision. Beyond traditional models, novel approaches have also emerged. Ref. [58] introduced the CFEBT method, which aims to detect creative accounting

used to postpone bankruptcy. According to [59], any effective financial reporting system must prioritise accuracy and truthfulness, especially when financial statements are used for managerial, regulatory, and strategic decision-making purposes. For most interest groups, accounting information remains a primary source of information about the company. Therefore, it is logical that each interest group will expect different values to meet their own goals. Consequently, it is important to determine the limits in accounting legislation by considering the use of different accounting techniques and procedures that accounting entities use to express their objectives. The approach to evaluating the company typically provides the motivation for these goals and the necessary values. In this direction, business evaluation is most conveniently expressed through the improvement of mathematical apparatuses, the decomposition of indicators, and subsequent applications based on selective data that may not always yield the desired exact result. Contend that creative accounting transcends mere technical manipulation, frequently arising from ingrained organisational compromises and decision-making cultures [60]. Their qualitative findings demonstrate that accounting outcomes mirror internal power dynamics, necessitating a more profound contextual comprehension beyond just quantitative study.

## 3. Materials and Methods

### *3.1. Data Used in the Study*

The dataset used in this study was sourced from the publicly available Finstat database, which aggregates financial statements and registration information for Slovak and Czech firms. The database facilitates the systematic extraction of firm-level data and is commonly employed in academic research and financial analysis. The database's accessibility ensures the replicability of the analysis, allowing other researchers to obtain all data using the same filters and timeframes. This source has been utilised in previous empirical investigations concerning Slovak SMEs and financial performance [45,48,61,62], confirming its pertinence and dependability for scholarly research.

The first dataset comprised 257,234 Slovak enterprises. The final sample, including data cleaning and filtering, consisted of 149,566 firms with consistent reporting from 2016 to 2023. The selection process adhered to the following criteria:

- Activity sectors: Companies were classified according to the NACE Rev. 2 taxonomy. We incorporated all private business enterprises throughout sectors A to S, excluding public administration (O), extraterritorial organisations (U), and unclassifiable or state-run entities.
- The sample encompasses micro, small, medium, and large firms, categorised according to the European Commission's SME classification, which is based on turnover and staff count.
- Exclusion criteria: Firms without complete or available essential financial statistics (e.g., total assets, equity, revenues), exhibiting zero turnover throughout the whole period, or possessing duplicate/inactive entries were excluded.
- Continuity filter: Only firms with valid annual financial data for the entirety of eight years (2016–2023) were included, hence confirming the stability of observed trends.

Each firm in the final dataset is characterised by 12 financial metrics, chosen for their theoretical significance in creative accounting and financial manipulation. These encompass metrics of profitability, liquidity, leverage, and efficiency.

To achieve more granularity, we categorised enterprises by region (NUTS 3 level), facilitating a regional analysis of aberrant conduct. The Bratislava region represented the greatest proportion (32.87%), succeeded by Košice (10.71%) and Žilina (10.65%). This stratification facilitates territorial comparisons and the identification of regional concentration effects.

Figure 1 illustrates a three-dimensional depiction of the input dataset used for model building. It depicts the distribution of enterprises based on their legal framework, economic sector (NACE classification), and geographic association (NUTS 3 level). This visual summary contextualises the composition of the sample, guaranteeing that the employed clustering algorithm is based on a sufficiently varied range of elements. The graph was generated using IBM SPSS Statistics, which enables the visualisation of categorical data in multidimensional spaces.

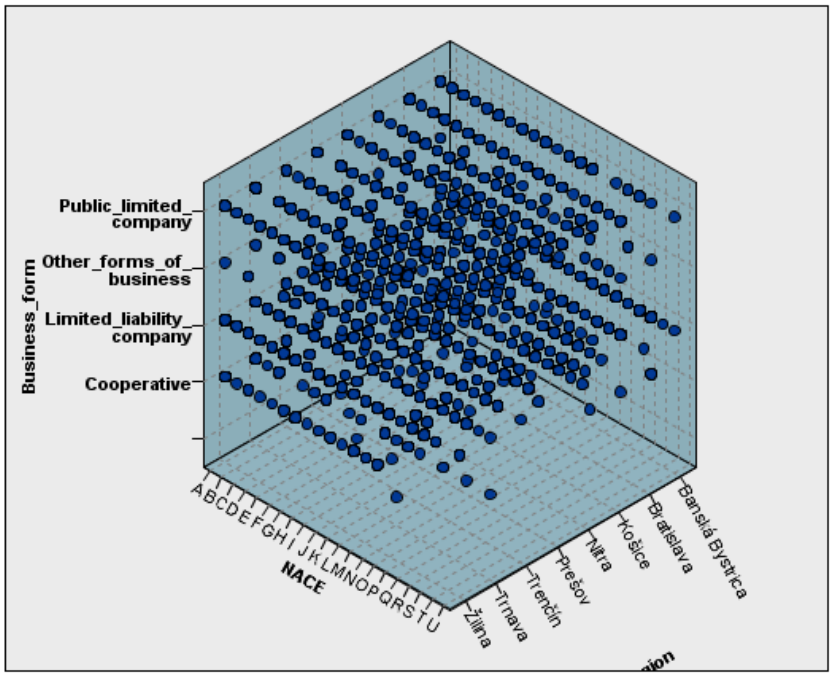


**Figure 1.** Structure of the input dataset by business form, NACE sector, and region, illustrating the diversity of included SMEs. Based on data from Finstat (2016–2023); own processing.

The comprehensive inclusion of enterprises across all regions provides a robust foundation for examining the spatial distribution and economic dynamics of Slovak businesses. The methodology ensured that no relevant business entities were excluded, thereby enhancing the accuracy and generalisability of the study's findings.

According to data from the Slovak Republic's Statistical Office, small and medium-sized enterprises (SMEs) with various legal forms account for up to 99.95% of the country's business environment. In this study, the most prevalent legal form among Slovak enterprises is the limited liability company (LLC), which accounts for 143,506 entities or 95.95% of the analysed sample. This reflects the dominant role of this legal structure in the Slovak business landscape, particularly due to its flexibility and accessibility for smaller enterprises.

The second most numerous legal form is the public limited company (PLC), comprising 4146 entities or 2.77% of the dataset. This form is typically associated with larger corporations operating on a broader, often national or international scale.

The third most common are cooperatives, with 1110 entities (0.74%), which play a niche but stable role, especially in sectors such as agriculture and community-based activities.

Lastly, other business forms account for 804 entities (0.54%), covering a wide range of less common legal structures tailored to specific operational or regulatory needs.

Figure 2 illustrates the distribution of firms categorised by legal forms in the input dataset. This visual representation was generated in Microsoft Excel 365 and enhances the numerical summary by distinctly illustrating the predominant presence of LLCs and the

general composition of the examined population. The dataset was sourced from Finstat and encompasses the years 2016 to 2023.

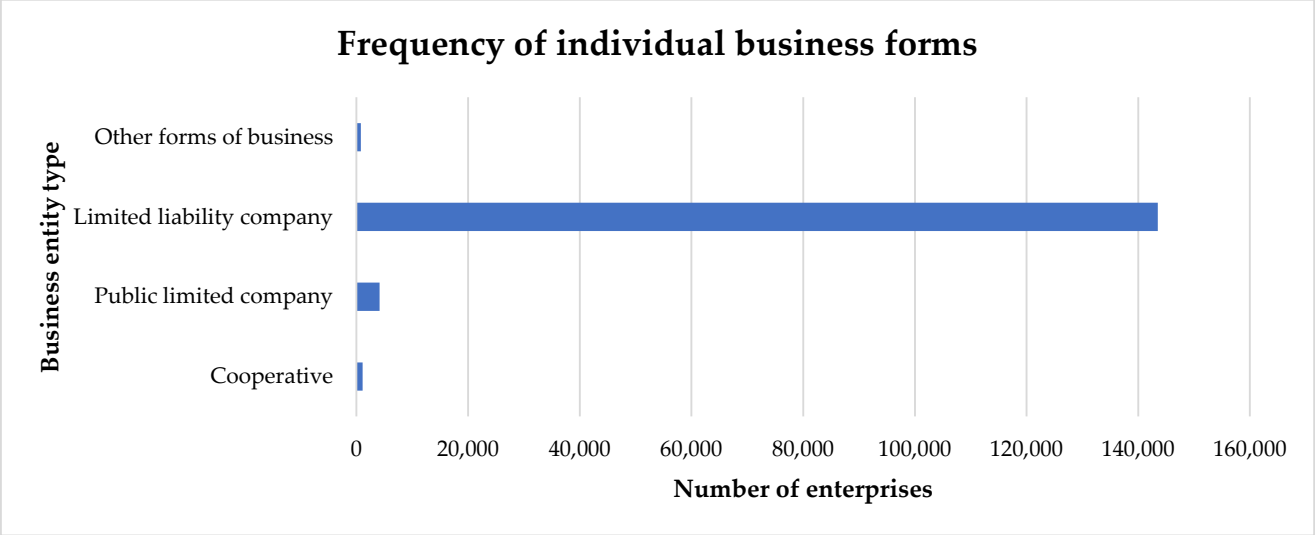


**Figure 2.** Distribution of enterprises by legal form of business, highlighting the most common legal structures among Slovak companies. Based on Finstat data (2016–2023); own processing.

This distribution underscores the predominance of limited liability companies within Slovakia's business environment, aligning with global trends in the SME sector. The significant share of PLCs and cooperatives further illustrates the diversity of legal forms supporting various economic activities across the country.

To develop a comprehensive model of creative accounting, publicly available data from companies' financial statements were utilised. A total of 12 financial ratios were chosen for model construction. Table 1 presents the median values of all selected variables during the reviewed period. Median values were preferred over means, as they provide a more accurate representation of the descriptive characteristics, given that financial ratios often exhibit a strongly right-skewed distribution. Therefore, mean values would not effectively reflect the central tendency of the variables.

**Table 1.** Overview of the median values of individual items.

| | **2016** | **2017** | **2018** | **2019** | **2020** | **2021** | **2022** | **2023** |
|---|---|---|---|---|---|---|---|---|
| Cash flow | 3422 | 4263 | 4113 | 4566 | 4148 | 4616 | 4778 | 5103 |
| Total assets | 46,627 | 55,707 | 62,501 | 67,753 | 70,940 | 75,944 | 81,032 | 84,667 |
| Tangible fixed assets | 4148 | 6187 | 7339 | 7784 | 7135 | 6960 | 7043 | 7421 |
| Current assets | 28,476 | 33,373 | 37,443 | 40,631 | 43,251 | 46,537 | 49,492 | 51,143 |
| Liabilities | 6639 | 7883 | 8597 | 9033 | 9165 | 9500 | 10,655 | 10,768 |
| Current liabilities | 17,674 | 20,268 | 21,173 | 21,779 | 21,618 | 23,425 | 24,210 | 24,668 |
| Non-current liabilities | 11 | 31 | 51 | 70 | 80 | 92 | 0 | 114 |
| Depreciation | 782 | 1443 | 22,015 | 2387 | 2400 | 2258 | 2083 | 1999 |
| Receivables | 6639 | 7883 | 8597 | 9033 | 9165 | 9500 | 10,655 | 10,768 |
| Operating profit | 960 | 960 | 1244 | 240 | 141 | 185 | 212 | 221 |
| Value added | 10,502 | 12,778 | 13,868 | 14,639 | 12,793 | 13,074 | 14,370 | 15,307 |
| Sales | 25,548 | 30,147 | 32,200 | 33,254 | 29,338 | 30,377 | 33,716 | 35,126 |

These variables were carefully selected to capture the essential components of financial reporting, ensuring consistency and reliability for the modelling process. They provide a comprehensive basis for analysing patterns and behaviours associated with creative accounting practices.

### 3.2. Methodology of Model Creation

The choice to use clustering-based anomaly detection was motivated by the lack of ground-truth labels in financial data related to fraudulent or irregular accounting activities. Supervised approaches, such as decision trees or neural networks, necessitate labelled datasets, which are rarely accessible or dependable in the context of small and medium-sized enterprises (SMEs). Clustering, specifically the integration of TwoStep and K-means, facilitates the segmentation of companies based on multivariate similarity and the detection of statistical outliers without preconceived assumptions. This methodology is particularly appropriate for extensive datasets, provides clear cluster interpretation, and facilitates replication in empirical research with the same data limitations.

The aim of the project was to develop an unsupervised model for identifying aberrant financial behaviour that could suggest applying creative accounting techniques. Consequently, we employed a two-phase clustering methodology utilising statistical and machine learning approaches executed in IBM SPSS Statistics and IBM SPSS Modeller. These tools were chosen for their reliability in processing large-scale tabular data, their strong implementation of TwoStep and K-means clustering algorithms, and their visual interpretability, which facilitates intuitive comprehension of clustering results. The visualisation and post-processing of the chosen outputs were conducted via Microsoft Excel 365, incorporating Power Pivot and Power Query features.

The modelling process included the subsequent steps:

- Variable selection: We employed 12 financial indicators that denote profitability, liquidity, leverage, and operational efficiency. These were selected because of their significance in previous fraud detection research [2,6].
- Standardization: Before clustering, all continuous variables were standardised via z-scores to ensure comparability and mitigate scale effects.
- Initial clustering (TwoStep): We initially employed the TwoStep technique to determine the optimal number of clusters. This approach uses a blend of distance metrics and model-based criteria, including the Bayesian Information Criterion, to ascertain the inherent groups within the dataset.
- The final clustering (K-means) used the best number of clusters found by TwoStep to help group the companies more effectively. This amalgamation improves both precision and resilience [63].
- Anomaly detection: We computed the anomaly index for each firm as the Euclidean distance from the corresponding cluster centroid. We identified the companies with the most significant deviation as probable anomalies. We utilised the upper quantile of the distance distribution (Q4) to ascertain the conclusive group of dubious firms.

The system is structured to function without the necessity of prior labelling for fraudulent or non-fraudulent instances. This renders it appropriate for data contexts where financial fraud is underreported, obscured, or only indirectly detectable. It enables regulators and analysts to examine extensive populations and pinpoint cases warranting more forensic scrutiny.

We used the silhouette coefficient to assess the suitability of the cluster structure, measuring both intra-cluster cohesiveness and inter-cluster separation. The finalised model achieved a silhouette score of 1.0, signifying outstanding compactness and differentiation among clusters. This result matches earlier studies showing that performance-related

measures for grouping data—like the silhouette coefficient and new validation metrics—are closely linked to how well anomalies are detected, often exceeding Pearson correlation coefficients of 0.95 [64]. Their findings corroborate our methodological approach and validate the efficacy of employing internal validation criteria to optimise cluster-based anomaly models in unsupervised contexts.

Fraud detection is one of the most common applications of data mining in the field of unsupervised learning. A typical task in this field often involves identifying fraudulent grant applications, insurance claims, tax evasion, or financial statement manipulation. Unlike supervised learning, a key characteristic of unsupervised learning is the absence of an outcome (dependent or target) variable. This absence may arise either from a type of solved problem, the lack of available data or, most typically, from the challenge of measuring the target variable. In our study, this situation arose from the inability to determine from historical records who committed fraud and who did not, except for a few publicly known cases, which are, however, insufficient for a comprehensive evaluation. As a result, traditional supervised machine learning models cannot be effectively utilised to make predictions of fraudulent behaviour based on historical data.

Unsupervised models, as fraudulent financial reporting indeed is, require a different approach and methodology for constructing prediction models. This approach involves identifying typical behavioural patterns within the population of companies and then detecting cases that significantly deviate from these patterns. Such cases are then classified as anomalies and are considered "suspicious" (unusual or outlier) instances in the data that must be further investigated. This procedure is based on the idea that companies with normal behaviour form well-defined clusters, while fraudulent activities appear as outliers or in less cohesive clusters. By analysing cluster characteristics and their deviations, fraudulent behaviour can be effectively detected.

In statistics, anomalies are defined as observations that markedly diverge from the anticipated pattern or trend of the dataset [65]. Such deviations may occur due to multiple factors, including measurement inaccuracies, inconsistencies in data collection methods, or infrequent occurrences. Anomalies can show up as outliers—values that are far away from the average or other typical values—or as unusual patterns that disrupt the basic organisation of the data [66].

Figure 3 offers a conceptual representation of the clustering-based anomaly detection method, illustrating how abnormalities are identified by their deviation from standard behavioural patterns. The concentric circles depict escalating divergence from the normative centre, which signifies "anticipated" financial behaviour resulting from clustering. Each dot signifies an observation (i.e., enterprise). Dark blue dots in the centre indicate standard or anticipated financial profiles. Conversely, the yellow, orange, and red dots are progressively further away from the centre, indicating greater levels of abnormality. The colour gradient represents the degree of deviation: red indicates the most anomalous entities, whereas yellow shows minor abnormalities. This picture, while not based on actual data, exemplifies the core premise of our analysis: the farther a data point is from the behavioural cluster centroid, the higher the probability that it is deemed anomalous.

The prediction model for identifying potential fraud behaviour is based on the above-mentioned principle of finding the typical behavioural patterns within the data cases. This step is conducted through cluster analysis, which aims to find clusters of cases that exhibit similar behaviour within the dataset. This similar behaviour is typically assessed by measuring the minimal distance between cases based on input variable values. Consequently, the fraud detection model is based on the identification of anomalies by finding the companies with the highest distance from the "normal" patterns.

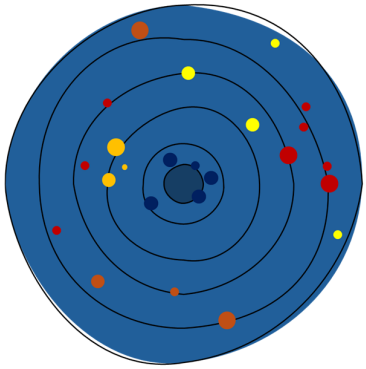

**Figure 3.** Conceptual illustration of anomaly detection based on distance from behavioural norm; own illustration created for explanatory purposes.

It must be stated that anomalies do not directly mean fraudulent behaviour. Instead, the anomalous case found is potentially fraudulent and needs to be further investigated using appropriate mechanisms. On the other hand, detecting anomalies identifies potential cases that must be subjected to this control instead of complete or random case screening.

Based on the described principle, creating a detection model for fraudulent behaviours involves two main steps. Firstly, the cluster analysis is used to find similar cases in the data. Secondly, the anomalous cases are detected by calculating the measure of anomaly based on the distance of the cases in created clusters from a typical mean representative of the particular cluster (cluster centroid).

In the following section, we will briefly describe the main principles of both these steps.

#### 3.2.1. Cluster Analysis

Cluster analysis is an unsupervised machine learning technique designed to identify groups of similar observations within a dataset. The method guarantees increased intra-cluster similarity while optimising inter-cluster dissimilarity [67].

The importance of clustering lies in its ability to identify and separate different data groups based on similarities and differences, which later serve as a foundation for constructing a reference model for anomaly detection. Without clustering, detecting anomalies in the data would be highly complex and computationally demanding, as it would require an exceptionally detailed understanding of all possible structures that anomalies might take. Clustering simplifies anomaly detection by identifying underlying structures in the data that may contain anomalies. The more effective the clustering process is, the more accurately anomalies can be detected, ultimately improving the performance of the final model.

Clustering relies on predefined similarity or distance metrics, computed from the values of the selected input variables. The Euclidean distance is favoured for continuous variables, whereas the Manhattan distance is generally employed for categorical or ordinal data [68]. In practical applications, the squared Euclidean distance is frequently employed to streamline calculations. An additional option is the Mahalanobis distance, which considers correlations among variables. This study employed Euclidean distance as a similarity metric for clustering.

Clustering can be performed utilising various algorithms, including hierarchical clustering and K-means clustering. Hierarchical clustering establishes groups by progressively merging or dividing clusters according to their pairwise distances. This method does not necessitate a predetermined number of clusters; however, it becomes computationally impractical for extensive datasets [69].

The second clustering algorithm mentioned, K-means clustering, is widely used in financial fraud detection due to its computational efficiency, scalability to large datasets, and ability to group transactions into well-defined clusters. This method defines clusters using centroids and assigns instances based on proximity to the nearest centroid, minimising intra-cluster variances. It is easy to implement and provides intuitive cluster groupings, facilitating the identification of patterns of normal and suspicious behaviour. However, its effectiveness relies on selecting an appropriate number of clusters, k, and it is sensitive to outliers, which can distort centroid positions and lead to the misclassification of fraudulent transactions. Therefore, it is important to note that instances marked as anomalies may not be considered fraud automatically but must be further investigated.

We used TwoStep Cluster Analysis, a scalable, model-based clustering technique included in IBM SPSS Statistics, to categorise the data into internally coherent and externally different groupings. This approach is especially appropriate for extensive datasets and facilitates the use of both continuous and categorical factors in the clustering procedure. The TwoStep technique employs a log-likelihood distance metric, predicated on the premise that continuous variables adhere to normal distributions and categorical variables conform to multinomial distributions, with variables exhibiting mutual independence. Log-likelihood distance is applicable to both continuous and categorical variables. The distance between two clusters is inversely related to the reduction of the natural logarithm of the probability function when they are amalgamated into a single cluster. Log-likelihood distance calculation assumes that continuous variables follow normal distributions, categorical variables adhere to multinomial distributions, and that the variables are independent of one another [70]. The distance between clusters $i$ and $j$ is defined as follows:

$$d(i,j) = \xi_i + \xi_j - \xi_{<i,j>} \tag{1}$$

where

$$\xi_c = -N_s\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2\right) + \sum_{k=1}^{K^B}\hat{E}_{sk}\right) \tag{2}$$

And in Equation (2), we obtain the following:

$$\hat{E}_{sk} = -\sum_{l=1}^{L_k}\frac{N_{skl}}{N_s}\log\frac{N_{skl}}{N_s} \tag{3}$$

with the following notations:

$d\ (i,j)$ is the distance between clusters $i$ and $j$;

$< i,j >$ index that represents the cluster formed by combining clusters $i$ and $j$;

$K^A$ is the totatl number of continuous variables;

$K^B$ is total number of categorical variables;

$L_k$ is the number of categories for the $k$-th categorical variable;

$N_s$ is the total number of data records in cluster $s$;

$N_{skl}$ is the number of records in cluster $s$ whose categorical variable $k$ takes $l$ category;

$N_{kl}$ is the number of records in categorical variable $k$ that take the $l$ category;

$\hat{\sigma}_k^2$ is the estimated variance (dispersion) of the continuous variable $k$, for the entire dataset;

$\hat{\sigma}_{sk}^2$ is the estimated variance of the continuous variable $k$, in cluster $j$.

The method employs a dual-phase approach to autonomously ascertain the quantities of clusters. First, for each possible number of clusters in the specified range, the BIC or AIC (Akaike Information Criterion) is calculated; this value is then used to make an initial guess

about how many clusters there are. This study utilised the BIC principle for J clusters, as defined by the following formula:

$$\mathrm{BIC(J)} = -2\sum_{\mathrm{j=1}}^{\mathrm{J}} \xi_{\mathrm{j}} + \mathrm{m_j log(N)} \tag{4}$$

where

$$\mathrm{m_j} = \mathrm{J}\left\{ 2\mathrm{K^A} + \sum_{\mathrm{k=1}}^{\mathrm{K^B}} (\mathrm{L_k} - 1) \right\} \tag{5}$$

The technique determines the ideal number of clusters by employing the Bayesian Information Criterion, which reconciles model fit with complexity. The TwoStep method's primary advantage is its automated clustering, which begins with the creation of a Cluster Feature Tree and advances to a probabilistic categorisation of instances, allocating each example to a cluster according to the calculated probability of cluster affiliation. The resultant clusters signify collections of enterprises exhibiting analogous financial profiles, which are thereafter used as a benchmark for anomaly detection.

This study employed K-means clustering with k designated as 5 clusters, which proved to be a suitable quantity, corroborated by the same value obtained from the TwoStep clustering method.

#### 3.2.2. Anomalies Detection

During the clustering process, observations are assigned to different clusters, and for each cluster, a centroid is calculated as the mean of the attribute values within that cluster. This centroid serves as a typical representative of the cluster. In mathematics, a centroid represents the central point of a geometric shape, symmetrically positioned relative to all other points within that shape. In data analysis, centroids are often used to represent the central points of clusters, calculated as the mean, acting as a typical cluster representative.

Furthermore, centroids are frequently used to compute distances between data points in data analysis. The distance between two data points is commonly assessed based on the distance between their respective centroids. This concept is particularly relevant in anomaly detection, as anomalies are identified by their significant deviation from the centroid of a given cluster. Thus, clustering not only helps structure data but also enhances the efficiency and accuracy of anomaly detection within datasets.

In this study, the anomaly index serves as a quantitative metric for deviations from standard financial behaviour. It gives a number to each observation that shows how much it differs from the usual pattern, making it easier to spot possible fraud or unusual cases. The anomaly index is generally computed as the distance between a data point and the centroid of its designated cluster, with elevated values signifying a higher probability of abnormal behaviour [71].

Mathematically, the anomaly index for a company $x_i$ can be defined as the ratio of the distance of $x_i$ from the centroid to the average distance of all cases in the cluster from its centroid, written as follows:

$$\text{Anomaly index} = \frac{\mathrm{d(x_i, C_{x_i})}}{\frac{1}{\mathrm{n}}\sum_{\mathrm{j}} \mathrm{d\left(x_j, C_{x_i}\right)}} \tag{6}$$

where $d\left(x_j,\, C_{x_i}\right)$ is an Euclidean distance between company $x_i$ and its corresponding cluster centroid $C_{x_i}$ and the denominator is the mean of the distances of all companies $x_j,\ j = 1,\ \ldots,\ n$ from the cluster centroid $C_{x_i}$.

Integrating the anomaly index in fraud detection models enables financial institutions to systematically prioritise high-risk transactions for further examination. A predetermined

anomaly threshold is set to detect potentially fraudulent behaviour by flagging instances that exceed the critical anomaly index value [72]. Should the distance from the centroid exceed this threshold, the respective observation is deemed anomalous. Another option should be to set a proportion of companies which would be considered anomalies, for example, 5 or 10% of companies, and consider them candidates for further investigation.

All calculations in this study were conducted using IBM SPSS Statistics (version 29) and the IBM SPSS Modeller (version 18.3), which provide functionalities for creating the model for detecting anomalies. To construct such a model, it is essential that this exploratory method, designed to detect unusual or outlier observations, is integrated with both nominal and numerical variables. This ensures that after thorough data preparation, the final model is successfully developed.

### *3.3. Model Evaluation*

The quality of clustering is contingent upon the judicious selection of input variables, the selection of an effective clustering algorithm, and the identification of the ideal number of groups. The TwoStep clustering approach estimates this number using the BIC, which reconciles model fit with complexity.

We used the silhouette coefficient, a recognised internal validation metric, to assess the internal structure of the resultant clusters. This metric evaluates intra-cluster cohesiveness and inter-cluster separation, making it especially appropriate for unsupervised learning tasks that lack labelled data for model validation.

Substantiate the utilisation of internal validation methodologies, highlighting the strong correlation between clustering quality and the efficacy of anomaly detection models [64]. Their research indicates that internal metrics such as the silhouette coefficient can function as reliable indicators of model validity when ground-truth labels are unavailable.

The silhouette is calculated as a measure of similarity between individual cases and the clusters to which they have been assigned. Specifically, in clustering with $k$ clusters and $n$ cases, let $a_i$ represent the average distance of case $i$ from all other cases within the cluster to which it was assigned, and let $b_i$ be the smallest average distance of case $i$ from all cases in any other cluster. Simply said, the silhouette compares the distances from the closest and the second closest cluster representatives. Then, the silhouette can be calculated as follows:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{7}$$

The silhouette score for the entire model is computed as the average silhouette score across all instances. The silhouette value ranges from $-1$ to 1, where the definitions are as follows:

- A value of 1 indicates that the instances are well separated within their clusters and exhibit strong similarity to their assigned clusters.
- A value of $-1$ signifies that the instances have been misclassified, implying a faulty clustering process.
- A value of 0 means that the instances are equally close to their assigned clusters and other clusters they were not supposed to belong to.

Subsequently, we calculated an anomaly index for each organisation based on its Euclidean distance from the cluster centroid. The larger the distance, the more unconventional the firm's financial profile is compared to its designated cluster. This unsupervised, distance-based methodology facilitates the detection of possible anomalies without dependence on established fraud cases.

Purity is calculated as the ratio of correctly assigned data cases to the total number of data cases. Specifically, in clustering with $k$ clusters and $n$ instances, let $p(i, j)$ represent

the number of cases in the true cluster *i* assigned to cluster *j* during the clustering process. Then, purity can be computed as follows:

$$\text{purity}(C) = \frac{1}{n} \sum_{p=1}^{k} \max_{p^+=1,\ldots,k^+} \left| C_p \bigcap C_{p^+}^{+} \right| \quad (8)$$

Purity is a value ranging from 0 to 1, with 1 indicating a perfect match between the true clusters and the clustering model, while 0 signifies a complete mismatch. It is essential to note that purity may become an ineffective metric in situations where samples across multiple clusters are highly similar or when clustering is not entirely distinct. In such cases, additional metrics such as the silhouette score or the Calinski–Harabasz index may be more suitable for evaluating clustering quality.

## 4. Results

The implementation of the integrated TwoStep and K-means clustering algorithms yielded five distinct clusters, each embodying a specific financial behaviour pattern across the study timeframe. The computed anomaly index, which is defined as the Euclidean distance from the cluster centroid, classified 231 companies (roughly 0.15% of the dataset) as anomalous. These organisations significantly diverged from their cluster's standard financial profile and were consequently identified for additional analysis.

Cluster analysis was conducted using the two-step clustering algorithm included in IBM SPSS Statistics. This approach autonomously identified the ideal cluster count via the BIC and categorised the companies according to their financial behaviours.

A total of 96 input variables, resulting from 12 principal financial indicators, were utilised and picked according to the variables outlined in the Beneish models, including the eight consecutive years under examination (2016–2023). This temporal granularity enables the model to identify structural and behavioural irregularities throughout time.

The clustering process yielded five discrete clusters, as illustrated in Figure 4. The upper panel of the graphic displays the model configuration, whereas the lower bar depicts the silhouette measure of cohesiveness and separation, which is categorised as "Good". The result indicates that the clusters are clearly delineated and internally coherent, hence enhancing the reliability of later anomaly identification.

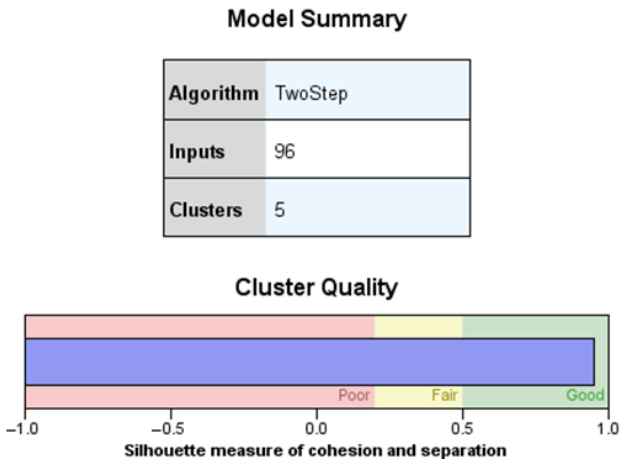


**Figure 4.** Model summary and silhouette-based cluster quality resulting from TwoStep clustering, based on own processing in IBM SPSS Modeller.

This internal validation technique adheres to the principles established by Foorthuis [64], which enhances the methodological rigour of the model.

The K-means technique used in IBM SPSS Modeller indicates a total of five clusters. The operating profit from 2016 to 2023 was used as the input sample variable for demonstration purposes.

The quality of the clustering solution was assessed using the silhouette coefficient, which evaluates both cohesion (within-cluster similarity) and separation (between-cluster dissimilarity). The resulting average silhouette value of 1.00 indicates excellent cluster quality, meaning that the individual cases are strongly associated with their assigned cluster and clearly distinct from cases in other clusters. According to the methodology employed in SPSS Modeller, this reflects a high degree of internal homogeneity and external heterogeneity among the clusters.

Figure 5 illustrates a comparative chart of clusters showing the annual distribution of operating profit across five clusters from 2016 to 2023. This visualisation was created using the IBM SPSS Modeller for post-clustering analysis. The figure facilitates an intuitive comparison of profitability patterns both within and among clusters over time.

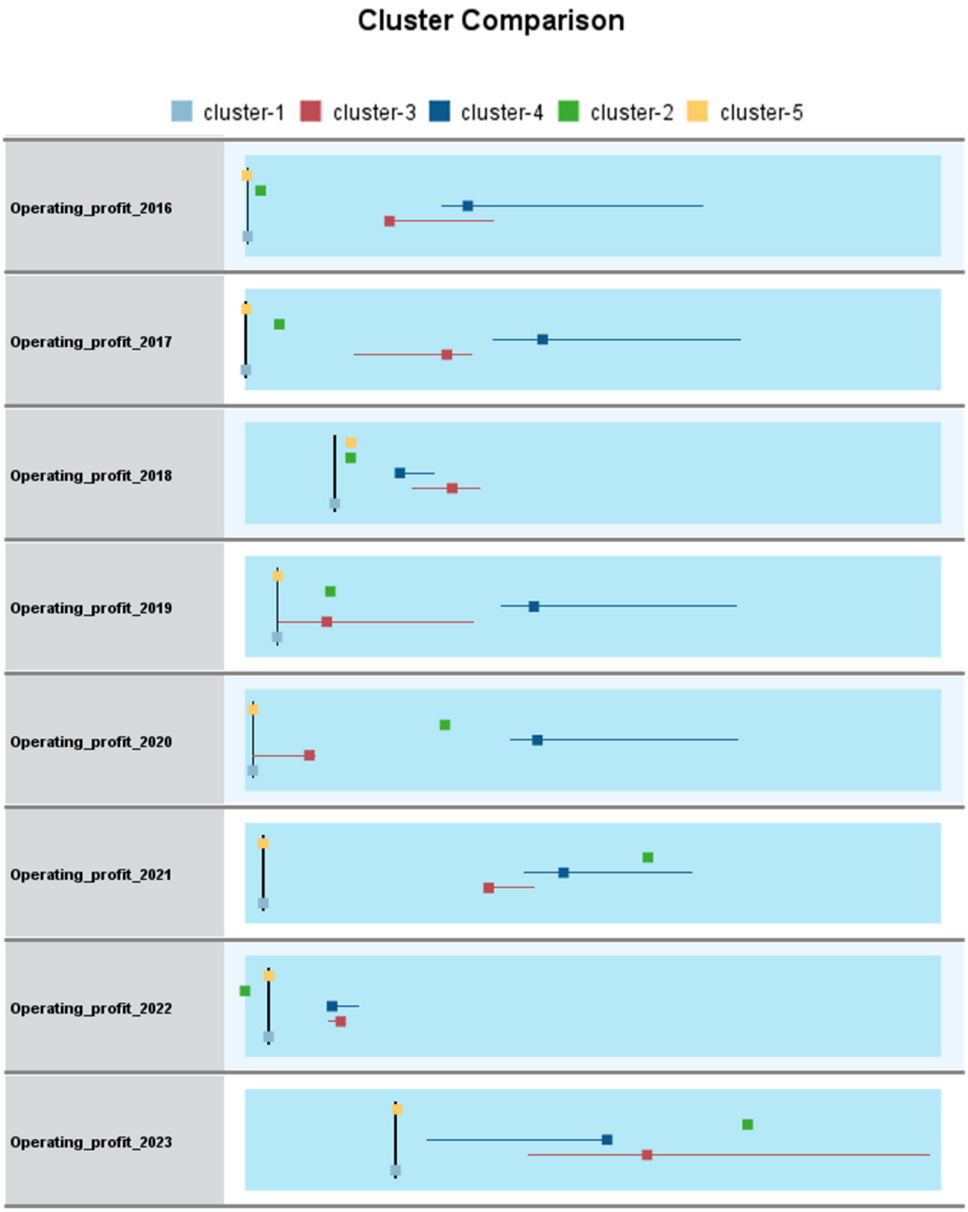


**Figure 5.** Cluster comparison chart for operating profit (2016–2023), based on own processing in IBM SPSS Modeller.

Each coloured sign represents one of the five clusters, while the horizontal placement reflects the distribution of operating earnings for a specific year. The image illustrates the historical dynamics of financial performance, facilitating the discovery of systemic disparities between groups.

Visually, distinct patterns emerge: Cluster 4 consistently exhibits elevated operational profit values, while Cluster 5 comprises firms with diminished or negative profitability. The inter-cluster disparities underscore the varied characteristics of financial conduct within the sample. Certain clusters demonstrate increased variability throughout the years, indicating disparities in risk exposure, operational stability, or sectoral sensitivity.

The chart illustrates the efficacy of the clustering method, particularly the application of the K-means algorithm subsequent to the TwoStep calculation of the ideal number of clusters. The model attained a high silhouette coefficient, reinforcing the statistical validity of the segmentation. This output establishes a basis for subsequent targeted analysis, such as identifying cluster-specific financial paths, conducting strategic benchmarking, or detecting anomalies based on temporal aberrations.

Figure 6 encapsulates the outcomes of the final clustering operation, which employed the K-means algorithm with 96 input variables across the period from 2016 to 2023. The ideal number of clusters (k = 5) was previously determined by two-step clustering using the BIC.

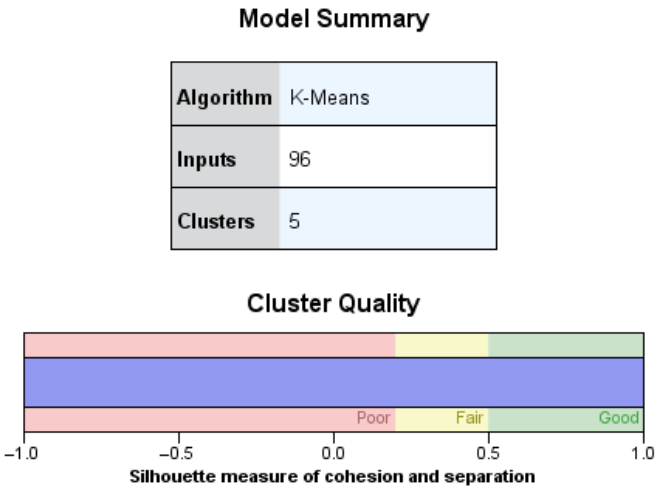


**Figure 6.** Silhouette-based cluster quality summary from K-means clustering, based on own processing in IBM SPSS Modeller.

The silhouette coefficient, shown in the lower panel of the figure, attained a value of 1.0, regarded as outstanding in clustering analysis. This score signifies that each observation is strongly cohesive inside its designated cluster and distinctly differentiated from other clusters.

This outcome offers robust statistical evidence for the validity and reliability of the created clusters, enhancing their appropriateness for further interpretation and anomaly identification. The figure was produced using IBM SPSS Modeller, which combines cluster summary tables with graphical validation tools.

This internal validation substantiates the efficacy of the clustering methodology, as advocated in the literature [65], wherein the silhouette index is acknowledged as a highly reliable metric of clustering performance in unsupervised learning contexts.

We computed the anomaly index for each company by measuring its Euclidean distance from the designated cluster centroid to identify those exhibiting possibly fraudulent

financial conduct. This distance-based index measures the divergence of each observation from the standard cluster configuration.

Figure 7 illustrates a histogram of anomaly index values for all enterprises in the sample. The majority of firms congregate around low index values, signifying adherence to anticipated financial conduct. The distribution displays an elongated right-hand tail, indicating firms with progressively unusual financial profiles.

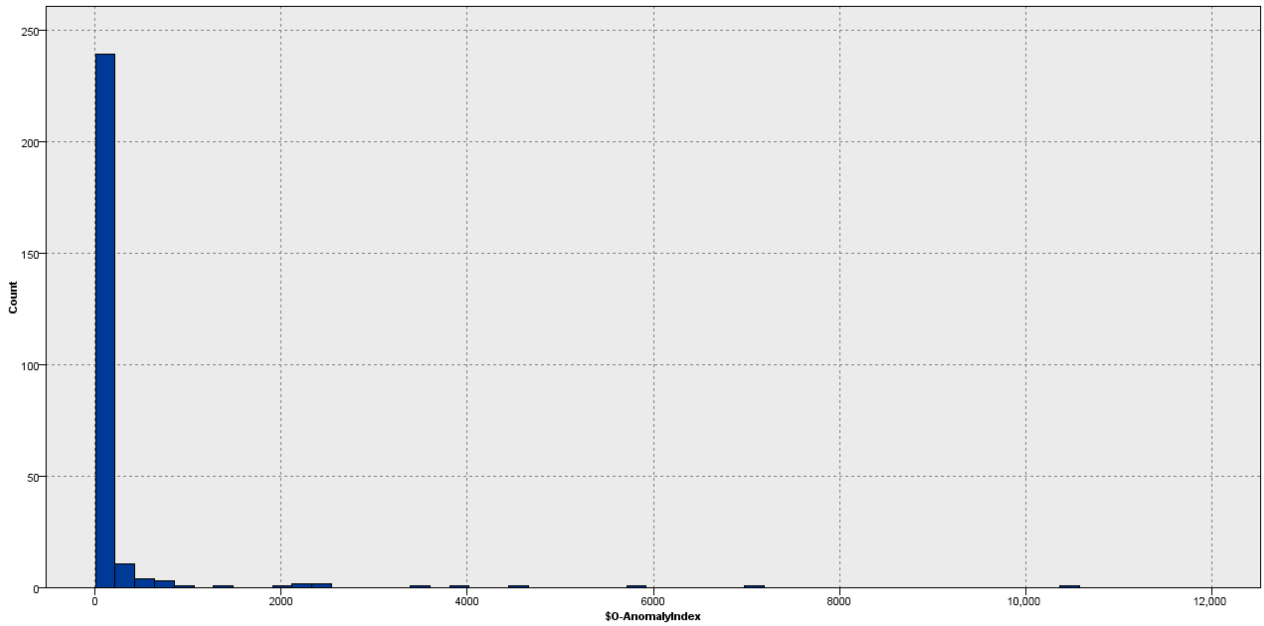


**Figure 7.** Histogram of anomaly index values across the sample, based on own processing using IBM SPSS Modeller.

This graphical form facilitates visual analysis of the distribution shape and aids in establishing an appropriate threshold for anomaly categorisation. In later stages, a cut-off value is determined to isolate the most severe outliers from the sample for comprehensive examination. The histogram was generated in IBM SPSS Modeller, facilitating the smooth integration of cluster results with exploratory visualisations.

To mitigate the pronounced skewness in the anomaly index distribution (refer to Figure 7), we implemented a logarithmic modification (base 10) on the index data. This adjustment enhances data interpretability by broadening the range of lower values and condensing outliers into a more comprehensible scale.

Figure 8 displays the histogram of the log-transformed anomaly index, facilitating improved visual differentiation between standard and anomalous instances. The vertical red line indicates the designated cut-off point for identifying outliers. This strategy improves the detection of the most significant anomalies and enables a clearer selection of companies for further analysis.

The histogram was produced with IBM SPSS Modeller, utilising the log-transformed anomaly index as a numerical input for visual analysis.

This picture provides a more elucidative representation of the anomaly index distribution. The use of a logarithmic transformation elucidates the points at which values begin to deviate significantly, revealing distinct gaps among the most extreme data. These gaps indicate a distinct behavioural divergence, and values in this area may be regarded as significantly deviating from the standard.

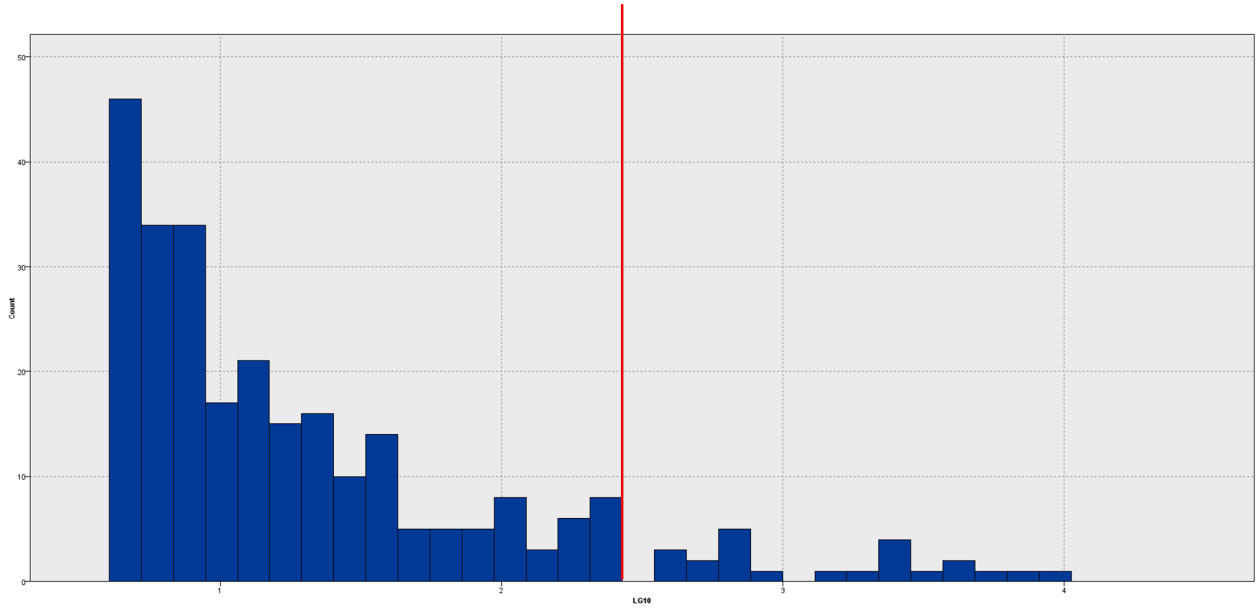


**Figure 8.** Histogram of log-transformed anomaly index values with marked anomaly threshold, based on own processing using IBM SPSS Modeller.

We implemented a quantile-based thresholding technique to identify organisations that displayed possibly fraudulent conduct. We specifically identified the top 10% of firms exhibiting the highest anomaly index values, which correlate to the greatest distances from their respective cluster centroids. These organisations' financial trends significantly diverge from those of their peers, indicating the potential for creative accounting procedures or unconventional reporting frameworks.

In the absence of labelled data, quantile-based threshold algorithms are commonly employed to identify anomalous cases. Prior studies by West and Bhattacharya [71] and Goldstein and Uchida [72] have employed criteria ranging from 1% to 10% of the highest anomaly scores, thereby creating a standard methodology for unsupervised anomaly detection.

In accordance with this methodological framework, we established a 10% threshold to identify the most severe observations based on their distance from cluster centroids. This level was determined by expert assessment and represents a practical equilibrium between sensitivity and specificity. The selection of 10% guarantees an adequate number of instances are identified for examination while preventing an excess of false positives in the study. This threshold is adaptable and may be modified based on the analytical goal or the necessary risk tolerance in actual applications.

*The Final Comprehensive Model of Creative Accounting*

Figure 9 illustrates the process flow of the anomaly detection model executed in IBM SPSS Modeller. The graphic delineates the sequence of components utilised, including data importation, variable selection, clustering, anomaly scoring, transformation, and result filtration.

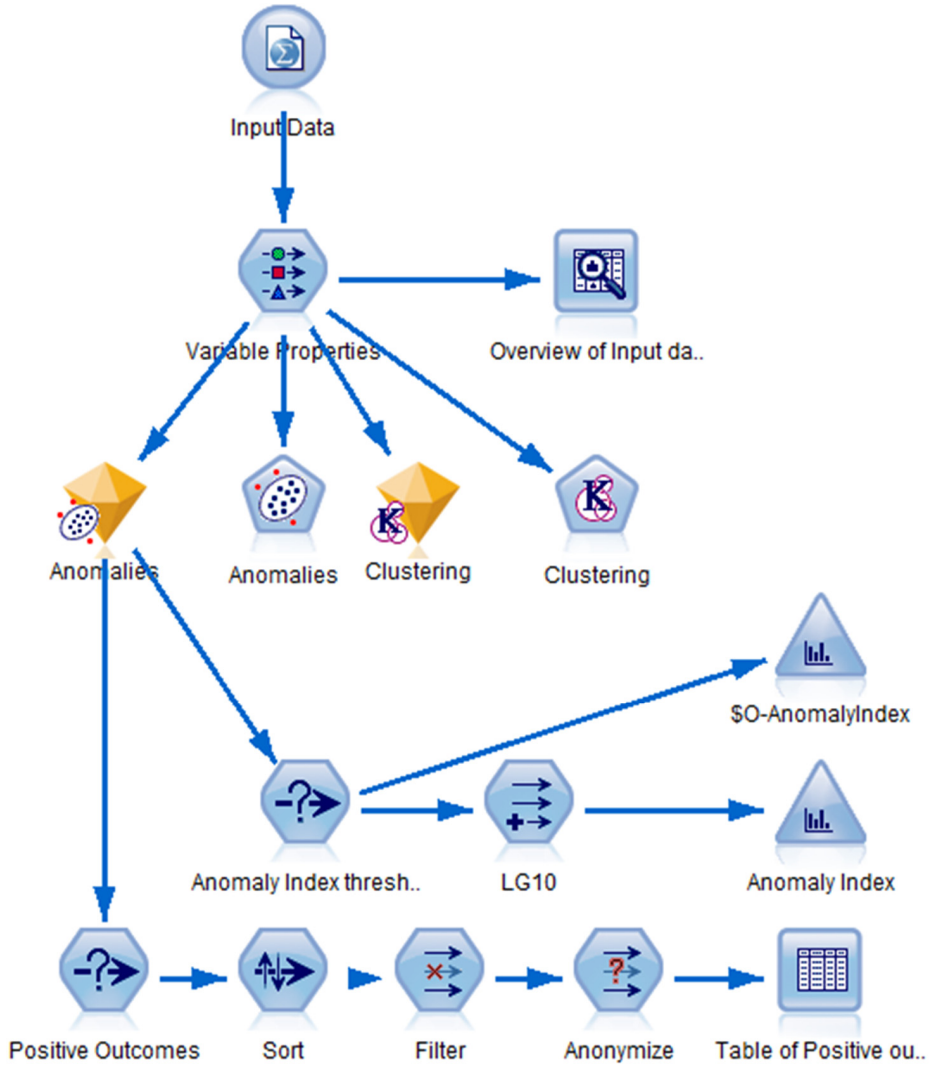


**Figure 9.** The final model for anomaly detection in business entities operating in the conditions of the Slovak Republic.

The model used the "number of anomalous cases" approach in the anomaly settings for the final selection of anomalous entities. The output was specifically designed to produce the top 10% of companies with the largest anomaly index values, representing the enterprises that most significantly diverge from the conventional financial behaviour of their designated cluster.

This parameter can be dynamically modified in SPSS Modeller, allowing users to select among the following three methodologies:

- Minimum anomaly index—establishing a definitive threshold value determined by the distance from the centroid;
- Proportion of cases—identifying a specified percentage (e.g., top 5%, 10%) of the most anomalous enterprises;
- The research determines the quantity of cases by isolating a precise count of the most extreme outliers.

Figure 10 provides a three-dimensional surface map showing the distribution of anomaly index values across NACE sectors (x-axis) and the estimated anomaly index, based on the Euclidean distance from the assigned cluster centroid as indicated in the approach. y-axis. The z-axis represents the legal structure of the enterprises.

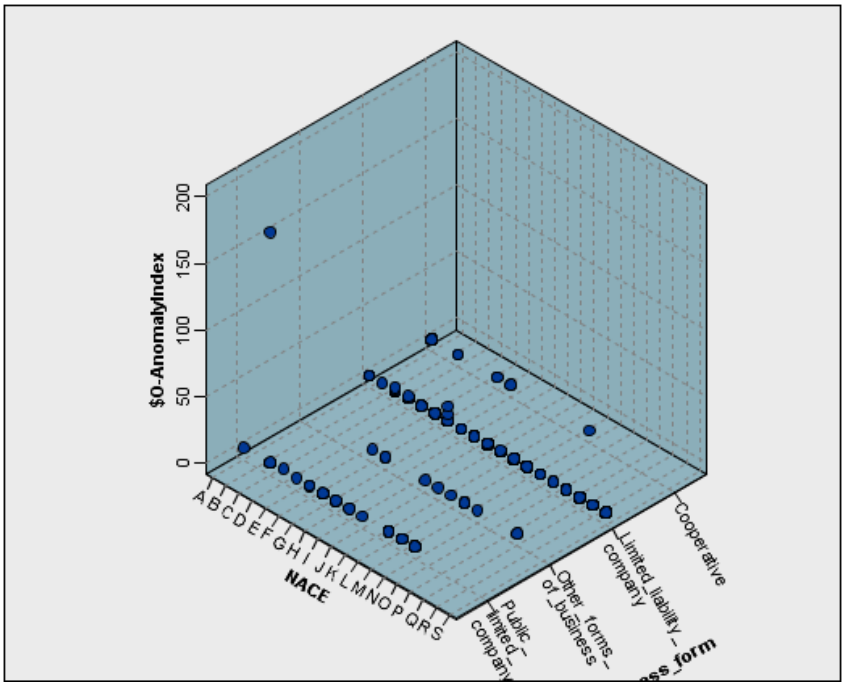


**Figure 10.** Three-dimensional distribution of anomaly index values by NACE sector and legal form, based on own processing in IBM SPSS Statistics.

The visualisation indicates that limited liability firms (LLCs) in certain sectors have markedly elevated anomaly ratings. These elevated ratings may indicate financial irregularities particular to the sector or abnormalities in structural reporting. Conversely, joint-stock corporations (PLCs) and cooperatives exhibit lower index values, signifying more uniform financial patterns within their respective categories.

The plot was generated using IBM SPSS Statistics. The 3D format offers an integrated perspective of sector and legal structure; however, it must be approached with caution due to potential visual distortion and overlapping surfaces. The presence of elevated index values in specific sector–form combinations suggests the importance of sector-sensitive methodologies in anomaly detection and risk assessment.

Figure 11 illustrates the fluctuations in average anomaly index values across Slovak firms, categorised concurrently by economic sector (NACE), NUTS3 area, and legal company structure. The anomaly index indicates the Euclidean distance between each firm and the centroid of its designated cluster, acting as a surrogate for the extent of divergence from standard financial conduct.

The chart indicates significant variability in anomaly intensity. Specific NACE categories, those designated H, J, and M, have elevated anomaly values across various locations and legal structures. Enterprises located in the Bratislava and Prešov regions, particularly limited liability corporations, consistently demonstrate high anomaly indices across many sectors. This pattern may indicate regional economic traits, sector-specific reporting methodologies, or differences in firm scale and regulatory scrutiny.

A high anomaly index does not inherently signify fraudulent behaviour; nonetheless, it denotes statistically unusual financial behaviour that warrants further scrutiny. The identified multi-dimensional discrepancies highlight the necessity of context-sensitive risk evaluations. These findings may assist auditors and regulators in focusing their analytical endeavours on sectoral, legal, and geographical risk considerations.

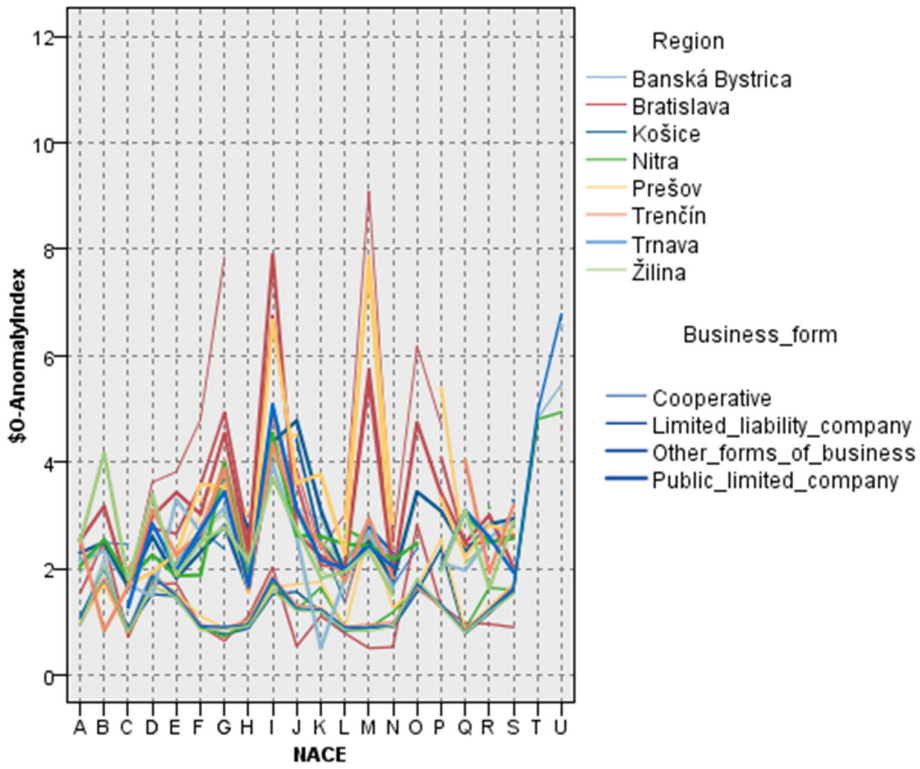


**Figure 11.** Average anomaly index by economic sector, region, and legal form, based on own processing in IBM SPSS Modeller.

Figure 12 illustrates a three-dimensional scatterplot depicting the distribution of average anomaly index values across various combinations of regions (NUTS 3), NACE sectors, and legal types of enterprises. The dimensions and hues of each marker reflect the average anomaly index for each particular combination, providing a comprehensive perspective on how structural attributes affect financial behaviour.

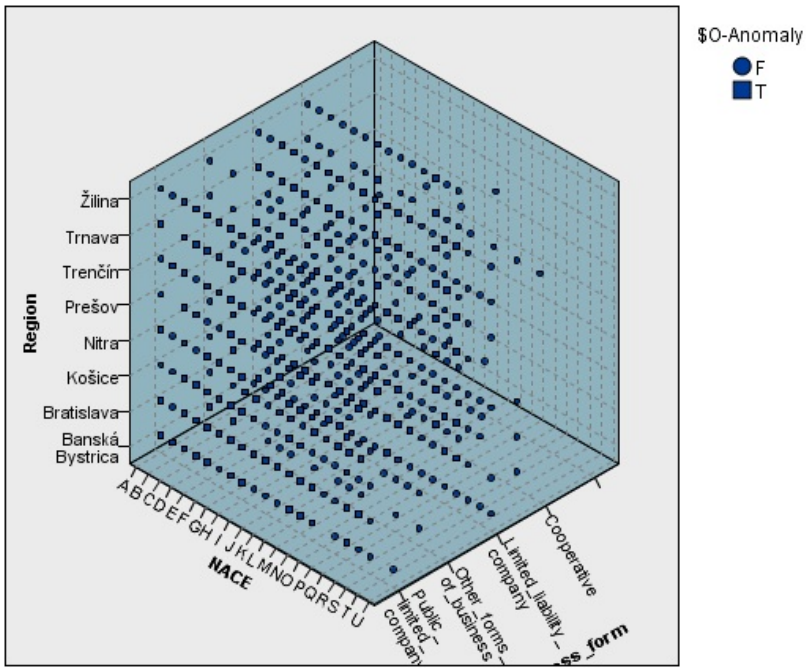


**Figure 12.** Three-dimensional distribution of average anomaly index by region, legal form, and NACE sector, based on own processing in IBM SPSS Statistics.

The data indicates that limited liability corporations (LLCs) predominate in all areas and industries, aligning with their widespread presence in the Slovak economic environment. Nonetheless, there exist nuanced differences in the composition of legal forms, particularly in areas like Nitra and Prešov, where the prevalence of LLCs and the lack of specific formal structures may affect anomaly index results.

In contrast, cooperatives and public limited corporations (PLCs) typically exhibit more regionally concentrated distributions, frequently linked to particular sectors. The disparities in organisational structure and sectoral focus may partially explain the regional variations in anomaly scores and highlight the necessity for contextual analysis when assessing anomalous activity.

## 5. Discussion

This study concentrated on identifying abnormal financial behaviour in Slovak firms through the application of an unsupervised machine learning model using clustering and distance-based scoring. Our findings, based on a dataset of 149,566 entities from 2016 to 2023, identified 231 enterprises as anomalous, determined by their Euclidean distance from cluster centroids. This methodological technique allowed us to systematically identify organisations that exhibit financial trends that considerably deviate from their counterparts.

### *5.1. Comparison with the Existing Literature*

The identification of financial anomalies and irregular accounting procedures has garnered heightened academic interest. Numerous recent studies have suggested diverse methodologies for detecting dubious financial patterns. The research undertaken by Urdaneta-Camacho et al. [4] examined the function of internal auditors and their stakeholder impacts in alleviating creative accounting, emphasising the deficiencies of conventional detection methods in intricate organisational environments. In 2015, Abubakr and his colleagues [5] conducted a study analysing creative accounting procedures in Spanish professional football clubs, emphasising the absence of effective automated screening systems capable of swiftly detecting inconsistencies.

Our findings align with the extensive literature acknowledging that specific economic sectors demonstrate increased susceptibility to creative accounting. Our study specifically indicated that sectors C (Manufacturing), G (Trade), and L (Real estate activities) had the largest proportion of anomalous enterprises, indicating a sectoral pattern of deviation. This corresponds with the conclusions of Blue et al. [6], who proposed a probabilistic approach for identifying creative accounting and observed a heightened risk in capital-intensive and transaction-heavy sectors.

Our methodology, utilising two steps and K-means clustering [73], revealed remarkable cohesiveness and separation in groupings, as shown by a silhouette score of 1.0. This is consistent with Foorthuis [64], who demonstrated that internal clustering validation measures, especially silhouette coefficients, exhibit a substantial correlation with the efficacy of unsupervised anomaly detection models. Although several research studies have explored intricate algorithms, such as the hyperparameter-tuned anomaly detection model introduced by Bakumenko and Elragal [74], which identifies nuanced, embedded aberrations, our methodology prioritises interpretability and computational efficiency. Furthermore, Crépey et al. [75] examined anomaly identification from a transactional viewpoint by scrutinising abnormalities in general ledger journal entries, which offers an alternate yet complementary method to our entity-level, ratio-based research. Although several studies have explored more intricate models (e.g., neural networks, PCA-based methodologies by Aros et al. [76]), our application of interpretable clustering techniques ensures transparency and replicability in scenarios devoid of labelled data.

### 5.2. Interpretation of Key Findings

The identified abnormalities were not randomly distributed; rather, they displayed discernible trends across sectors, legal structures, and geographical locations. The prevalence of limited liability corporations (LLCs) among atypical organisations indicates their superiority in the Slovak SME sector and potentially enhances flexibility in financial reporting. In eastern Slovakia, particularly in Prešov and Nitra, elevated average anomaly indices were noted, likely due to structural variables including industry makeup, company size, and local regulatory practices.

These findings suggest that financial irregularities are likely to aggregate within certain institutional and regional contexts, underscoring the necessity of contextualised analysis [77]. Unsupervised approaches demonstrated significant efficacy in recognising clusters without prior labels, making them appropriate for large-scale screening tasks in contexts where fraud is underreported or obscured.

The term "anomalies" in this study specifically denotes statistically significant deviations from anticipated financial conduct within each cluster. These discrepancies do not inherently indicate fraudulent behaviour; instead, they underscore instances that necessitate further scrutiny. The anomaly detection model should be regarded as a screening instrument that assists auditors, analysts, or regulatory bodies in prioritising situations for comprehensive examination, rather than as a conclusive fraud detection mechanism.

### 5.3. Research Gap and Original Contribution

Although prior studies have thoroughly examined fraud detection through supervised models, a notable deficiency persists in scalable, unsupervised methodologies that function without ground-truth data. Our research addresses this deficiency by introducing a transparent, reproducible, and interpreted anomaly detection algorithm specifically designed for financial data from SMEs in a Central European context.

Additionally, we present a quantile-based thresholding method, corroborated by the existing literature from West and Bhattacharya [71] and Cai et al. [73], to systematically identify the most extreme cases. The incorporation of eight years of temporal financial data, comprising 12 variables for a total of 96 variables, significantly augments the model's uniqueness and robustness.

This study addresses the conceptual and methodological requirements identified by [78], who underscored the necessity for a solid theoretical foundation in anomaly detection research. Our research provides both an application model and a conceptual framework for analysing aberrant financial arrangements in a practical context.

### 5.4. Practical Implications and Future Research

The model introduced in this paper possesses significant practical importance for regulatory authorities, auditors, and internal control frameworks. Anomaly detection aids executives in identifying unusual transactions that deviate from expected trends, as noted by Thudumu et al. [78]. Our results validate that clustering-based models can efficiently function as screening instruments to identify high-risk items for further scrutiny, even without labelled data.

Notwithstanding the model's encouraging outcomes, various limitations must be noted. The silhouette score of 1.0, while indicative of robust internal cohesion and separation, may imply overfitting or be an artefact of the two-step method's automatic preclustering. This result requires careful interpretation and necessitates future benchmarking through external validation methods or alternative clustering validity indices. The existing model employs a fixed threshold (top 10%) for anomaly detection, which may not be the most effective criterion across diverse industries with varying data distributions. Using

adaptive thresholding methods like z-scores, interquartile ranges (IQRs), or percentile-based criteria could make the model better at detecting differences in various industries and help lower the number of false alarms. Additionally, the anomaly index, which only uses Euclidean distance, might not fully capture complex relationships between multiple variables, particularly when financial variables are related. Future research may investigate the use of Mahalanobis distance, which considers variance and covariance structures in the data and has proven effective in recent anomaly detection applications (Savran et al. [79]). Subsequent research may explore ensemble methods or hybrid strategies that combine clustering with classification techniques, as suggested by Bakumenko et al. [80] and Lokanan et al. [81].

Additionally, incorporating macroeconomic elements into the model, as proposed by Yildiz et al. [82] and Majumder [83], may improve its sensitivity to contextual risks. Ultimately, frameworks for cost–benefit analysis could be established to assess the possible financial and reputational repercussions of unreported anomalies.

To improve the external validity of the model, subsequent research should aim to compare identified anomalies with actual audit results, public enforcement actions, or documented instances of financial malfeasance. The process may encompass expert interviews, coordination with auditing professionals, or triangulation with external risk databases. This validation would ascertain whether the detected abnormalities correspond with actual instances of creative accounting or financial misrepresentation.

### *5.5. Theoretical Contributions*

This research enhances the body of knowledge on anomaly detection and financial irregularities in three principal aspects. Initially, it constructs an interpretable, unsupervised model that functions efficiently in data environments devoid of labelled fraudulent instances—a scenario frequently encountered in SME datasets. Secondly, the findings empirically confirm the hypothesis that anomalies are concentrated both sectorally and regionally, thereby reinforcing the structural aspect of financial misreporting. The incorporation of multi-year financial indicators offers a dynamic view of firm behaviour, surpassing the static financial snapshots commonly employed in earlier models.

## 6. Conclusions

This research employed unsupervised machine learning methods to identify unusual financial conduct among Slovak firms. Employing clustering-based methods on a dataset of 149,566 enterprises and 96 financial indicators spanning eight years (2016–2023), we found 231 entities with anomalous financial trends. This work enhances the methodology of anomaly detection by integrating a multi-year view, allowing for the identification of structural aberrations that may be overlooked in short-term analysis. This represents one of the initial extensive implementations of unsupervised clustering-based anomaly detection in Central European SMEs using public registry data. The greatest concentration of abnormalities was identified in sectors C (Manufacturing), G (Trade), and L (Real Estate). The findings indicate that certain industries may be more vulnerable to accounting irregularities, which should be prioritised for regulatory oversight and audit intervention.

From a theoretical standpoint, our findings enhance the expanding body of work on anomaly detection by illustrating that unsupervised models may effectively recognise systematic discrepancies in financial data without the necessity of tagging fraud instances. This facilitates the creation of scalable, interpretable instruments for detecting financial anomalies in practical settings, especially inside under-explored SME environments.

The proposed methodology offers an effective screening tool for auditors, financial controllers, and regulatory bodies. Its form is flexible for various institutional contexts and

nations, making it appropriate for cross-national comparisons or standardised regulatory frameworks, such as those in the EU or OECD. The capacity to identify abnormalities by industry, location, or legal structure offers helpful tips for risk-orientated audit planning and focused policy execution. This method can also be applied in internal business environments to identify high-risk entities or transactions. Given that financial transparency is fundamental to economic credibility, including anomaly detection methods in supervisory processes may enhance early warning systems and reduce the incidence of financial malfeasance.

Notwithstanding its merits, the study possesses multiple limitations. The approach has a set threshold (top 10%), which may not accurately represent optimal sensitivity across various sectors. It employs a distance-based anomaly index, which may neglect complex relationships among financial ratios. The analysis is confined to accounting data and excludes macroeconomic variables, behavioural indicators, or textual analyses from narratives or disclosures. While a fixed percentile threshold was employed, alternative thresholding techniques, such as dynamic z-score filters or distance quantiles modified by cluster density, might be investigated to enhance the detection of abnormal observations.

Future studies will explore sophisticated machine learning techniques, such as ensemble models and neural networks, to enhance the accuracy of anomaly categorisation. Furthermore, incorporating elements of profitability, liquidity, activity, and leverage would provide a more holistic perspective on financial conduct. Subsequent enquiries may concentrate on sector-specific models and the influence of macroeconomic and regulatory changes on the creation of anomalies. Furthermore, since the dataset includes eight years of financial reporting, future studies could use longitudinal modelling tools to examine temporal patterns, identify recurring anomalies, and distinguish between temporary financial shocks and structural irregularities. These guidelines would enhance detection techniques and facilitate the continuous development of resilient, flexible frameworks for evaluating financial integrity. A future study may benefit from interdisciplinary collaboration, integrating financial analytics with legal reasoning, behavioural science, or text mining to comprehensively address financial anomalies and reporting habits. We advise that institutions engaged in financial oversight and auditing explore trial deployments of anomaly detection frameworks within their risk assessment methodologies.

The model relies exclusively on historical financial indicators, devoid of contextual or behavioural characteristics, and should be regarded as a preliminary framework for anomaly detection. Future development must emphasise the integration of supplementary qualitative data, including audit findings and disclosure quality, while involving domain experts to authenticate the model's conclusions. This would markedly enhance the model's accuracy and expand its utility in regulatory or internal control contexts.

**Author Contributions:** Conceptualization, R.B. and L.D.; methodology, R.B.; software, R.B.; validation, R.B. and L.D.; formal analysis, R.B.; investigation, L.D.; resources, R.B.; data curation, L.D.; writing—original draft preparation, R.B.; writing—review and editing, L.D.; visualization, R.B.; supervision, R.B.; project administration, L.D.; funding acquisition, R.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available from the corresponding author upon request.

**Acknowledgments:** This research was financially supported by the Slovak Grant Agency for Science (VEGA) under grant No. 1/0509/24.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Beneish, M.D. The detection of earnings manipulation. *Financ. Anal. J.* **1999**, *55*, 24–36. [CrossRef]
2. Sabău, A.I.; Mare, C.; Safta, I.L. A statistical model of fraud risk in financial statements. case for Romania companies. *Risks* **2021**, *9*, 116. [CrossRef]
3. Durana, P.; Blazek, R.; Machova, V.; Krasnan, M. The use of Beneish M-scores to reveal creative accounting: Evidence from Slovakia. *Equilib. Q. J. Econ. Econ. Policy* **2022**, *17*, 481–510. [CrossRef]
4. Urdaneta-Camacho, R.; Guevara-Pérez, J.C.; Martín Vallespín, E.; Le Clech, N. The other side of the "League of Stars": Analysis of the financial situation of Spanish Football. *Int. J. Financ. Stud.* **2022**, *11*, 3. [CrossRef]
5. Abubakr, A.A.M.; Khan, F.; Alomari, K.M.; Sahal, M.S.G.; Yousif, N.A.I.; Yousif, H.K. The role of the internal auditor, stakeholders, and external auditor in discovering and reducing creative accounting practices in UAE companies. *Secur. J.* **2025**, *38*, 31. [CrossRef]
6. Blue, G.; Chahrdahcheriki, M.; Rezaee, Z.; Khotanlou, M. A model for predicting creative accounting in emerging economies. *Int. J. Account. Inf. Manag.* **2025**, *33*, 1–31. [CrossRef]
7. Balaciu, D.; Bogdan, V.; Vladu, A.B. A brief review of creative accounting literature and its consequences in practice. *Ann. Univ. Apul. Ser. Oeconomica* **2009**, *11*, 13.
8. Schiff, M. Accounting Tactics and the Theory of the Firm. *J. Account. Res.* **1966**, *4*, 62–67. Available online: https://www.jstor.org/stable/2490141 (accessed on 9 July 2025). [CrossRef]
9. Copeland, R.M. Income Smoothing. *J. Account. Res.* **1968**, *6*, 101–116. Available online: https://ideas.repec.org/a/bla/joares/v6y1968ip101-116.html (accessed on 9 July 2025).
10. Ondráš, F. Analýza Převodu Českých a Slovenských Účetních Výkazů na Výkazy Podle IFRS. Master's Thesis, Masaryk University, Brno, Czech Republic, 2016.
11. Stolowy, H.; Breton, G. Accounts manipulation: A literature review and proposed conceptual framework. *Rev. Account. Financ.* **2004**, *3*, 5–92. [CrossRef]
12. Bertolus, J.J. L'art de Truquer un Bilan. *Sci. Vie Econ.* **1988**, *40*, 17–23.
13. Lignon, M. L'art de calculer ses bénéfices. *Entreprise* **1989**, *50*, 17–18+20.
14. Gounin, I. L'art de Présenter un Bilan. *Tribune* **1991**, *28*, 11.
15. Ledouble, D. La créativité en comptabilité. *Sem. Jurid. (JCP E)* **1993**, *224*, 1–17.
16. Murineanu, A.M. Creative Accounting: A Literature Review. In Proceedings of the International Conference on Business Excellence, Bucharest, Romania, 21–23 March 2024; Sciendo: De Gruyter, Poland, 2024; Volume 18, pp. 2112–2121. Available online: https://sciendo.com/article/10.2478/picbe-2024-0178 (accessed on 9 July 2025).
17. Honková, I.; Myšková, R. Techniques, objectives and motivations of creative accounting: Evidence from the Czech Republic. *Natl. Account. Rev.* **2024**, *6*, 333–351. [CrossRef]
18. Khaneja, S.; Bhargava, V. A comprehensive review of literature on creative accounting. *Int. J. Bus. Insights Transform.* **2016**, *10*, 46–60.
19. Dixit, A.; Shukla, A. Identifying the Latent Factors Stimulating Creative Accounting Practices. *Pac. Bus. Rev. Int.* **2023**, *16*. Available online: http://www.pbr.co.in/2023/2023_month/October/2.pdf (accessed on 9 July 2025).
20. Jameson, M. *A Practical Guide to Creative Accounting*; Kogan Page: London, UK, 1988; ISBN 978-1850916116.
21. Ado, A.B.; Rashid, N.N.M.; Mustapha, U.A.; Ademola, L.S. Audit Quality and Creative Accounting Strategy: Evidence from Nigerian Public Listed Companies. *Australas. Account. Bus. Financ. J.* **2022**, *16*, 40–54. [CrossRef]
22. Kliestik, T.; Hrosova, L.; Valaskova, K.; Svabova, L. Do firm in the tourism sector manage earnings? The case of the V4 countries. *J. Tour. Serv.* **2022**, *13*, 120–136. [CrossRef]
23. Ruddy, T.; Everingham, G.K. Creative accounting, accounting errors, and the ability of users to detect and adjust for them. *S. Afr. J. Account. Res.* **2008**, *22*, 45–95. [CrossRef]
24. Naser, K.H. *Creative Financial Accounting: Its Nature and Use*; Prentice Hall: Englewood Cliffs, NJ, USA, 1993; p. 250. ISBN 0130617636.
25. Merchant, K.A.; Rockness, J. The ethics of managing earnings: An empirical investigation. *J. Account. Public Policy* **1994**, *13*, 79–94. [CrossRef]
26. Jones, M.J. *Creative Accounting, Fraud and International Accounting Scandals*; John Wiley & Sons: Hoboken, NJ, USA, 2010. ISBN 978-0470057650.
27. Aharony, J.; Wang, J.; Yuan, H. Tunneling as an incentive for earnings management during the IPO process in China. *J. Account. Public Policy* **2010**, *29*, 1–26. [CrossRef]

28. Veiga, J.F.; Golden, T.D.; Dechant, K. Why managers bend company rules. *Acad. Manag. Perspect.* **2004**, *18*, 84–90. [CrossRef]
29. Siekelova, A.; Androniceanu, A.; Durana, P.; Michalikova, K.F. Earnings management (EM), initiatives and company size: An empirical study. *Acta Polytech. Hung.* **2020**, *17*, 41–56. [CrossRef]
30. Setyoputri, L.S.; Mardijuwono, A.W. The impact of firm attributes on earnings management. *Pol. J. Manag. Stud.* **2020**, *22*, 502–512. [CrossRef]
31. Poradova, M.; Siekelova, A. Analysis of Factors with Impact on Earnings and Their Management in Commercial Companies. 2020. Available online: https://digilib.uhk.cz/bitstream/handle/20.500.12603/275/Poradova,%20Siekelova.pdf?sequence=1&isAllowed=y (accessed on 9 July 2025).
32. Milesi-Ferretti, G.M. Good, bad or ugly? On the effects of fiscal rules with creative accounting. *J. Public Econ.* **2004**, *88*, 377–394. [CrossRef]
33. Ozkaya, A. Creative accounting practices and measurement methods: Evidence from Turkey. *Economics* **2014**, *8*, 20140029. [CrossRef]
34. Melo, M.A.; Pereira, C.; Souza, S. Why do some governments resort to 'creative accounting' but not others? Fiscal governance in the Brazilian federation. *Int. Political Sci. Rev.* **2014**, *35*, 595–612. [CrossRef]
35. Maltritz, D.; Wüste, S. Determinants of budget deficits in Europe: The role and relations of fiscal rules, fiscal councils, creative accounting and the Euro. *Econ. Model.* **2015**, *48*, 222–236. [CrossRef]
36. Reischmann, M. Creative accounting and electoral motives: Evidence from OECD countries. *J. Comp. Econ.* **2016**, *44*, 243–257. [CrossRef]
37. Hirota, H.; Yunoue, H. Fiscal Rules and Creative Accounting: Evidence from Japanese Municipalities. *J. Jpn. Int. Econ.* **2022**, *63*, 101172. [CrossRef]
38. Al-Khoury, A.; Haddad, H.; Ali, M.A.; Al-Ramahi, N.M.; Almubaydeen, T.H. A bibliometric analysis using Scopus database of the literature on creative accounting trends. *Montenegrin J. Econ.* **2024**, *20*, 79–98.
39. Marcuta, A.; Radoi, D.; Nuta, A.C.; Nuta, F.M.; Marcuta, L. Bibliometric Study on the Importance of Using Creative Accounting in Financial Reporting. *Sci. Pap. Ser. Manag. Econ. Eng. Agric. Rural Dev.* **2024**, *24*. Available online: https://managementjournal.usamv.ro/index.php/scientific-papers/3600-bibliometric-study-on-the-importance-of-using-creative-bibliometric-study-on-the-importance-of-using-creative (accessed on 9 July 2025).
40. Kliestik, T.; Novak Sedlackova, A.; Bugaj, M.; Novak, A. Stability of profits and earnings management in the transport sector of Visegrad countries. *Oeconomia Copernic.* **2022**, *13*, 475–509. [CrossRef]
41. Blazek, R.; Durana, P.; Valaskova, K. Creative accounting as an apparatus for reporting profits in agribusiness. *J. Risk Financ. Manag.* **2020**, *13*, 261. [CrossRef]
42. Durana, P.; Valaskova, K.; Blazek, R.; Palo, J. Metamorphoses of earnings in the transport sector of the V4 region. *Mathematics* **2022**, *10*, 1204. [CrossRef]
43. Nagy, M.; Valaskova, K.; Durana, P. The effect of CSR policy on earnings management behavior: Evidence from Visegrad publicly listed enterprises. *Risks* **2022**, *10*, 203. [CrossRef]
44. Gajdosikova, D.; Valaskova, K.; Durana, P. Earnings management and corporate performance in the scope of firm-specific features. *J. Risk Financ. Manag.* **2022**, *15*, 426. [CrossRef]
45. Svabova, L.; Kramarova, K.; Chutka, J.; Strakova, L. Detecting earnings manipulation and fraudulent financial reporting in Slovakia. *Oeconomia Copernic.* **2020**, *11*, 485–508. [CrossRef]
46. Michalkova, L.; Cepei, M.; Valaskova, K.; Vincurova, Z. Earnings quality and corporate life cycle before the crisis. A study of transport companies across Europe. *Amfiteatru Econ.* **2022**, *24*, 782–796. [CrossRef]
47. Gajdosikova, D.; Valaskova, K. The impact of firm size on corporate indebtedness: A case study of Slovak enterprises. *Folia Oeconomica Stetin.* **2022**, *22*, 63–84. [CrossRef]
48. Durana, P.; Michalkova, L.; Privara, A.; Marousek, J.; Tumpach, M. Does the Life Cycle Affect Earnings Management and Bankruptcy? *Oeconomia Copernic.* **2021**, *12*, 425–461. [CrossRef]
49. Jones, J.J. Earnings Management During Import Relief Investigations. *J. Account. Res.* **1991**, *29*, 193–228. [CrossRef]
50. Dechow, P.M.; Sloan, R.G.; Sweeney, A.P. Detecting Earnings Management. *Account. Rev.* **1995**, *70*, 193–225.
51. Kliestik, T.; Valaskova, K.; Nica, E.; Kovacova, M.; Lazaroiu, G. Advanced methods of earnings management: Monotonic trends and change-points under spotlight in the Visegrad countries. *Oeconomia Copernic.* **2020**, *11*, 371–400. [CrossRef]
52. Kovacova, M.; Hrosova, L.; Durana, P.; Horak, J. Earnings management model for Visegrad Group as an immanent part of creative accounting. *Oeconomia Copernic.* **2022**, *13*, 1143–1176. [CrossRef]
53. Kliestik, T.; Nica, E.; Suler, P.; Valaskova, K. Innovations in the company's earning management: The case for the Czech Republic and Slovakia. *Mark. Manag. Innov.* **2020**, *3*, 332–345. [CrossRef]
54. Yin, F.; Zheng, Q. Other comprehensive income and earnings management-an empirical analysis based on modified jones model. In Proceedings of the International Conference on Transformations and Innovations in Management (ICTIM 2017), Shanghai, China, 9–10 September 2017; Atlantis Press: Cambridge, MA, USA, 2017; pp. 885–894.

55. Ball, R.; Shivakumar, L. The role of accruals in asymmetrically timely gain and loss recognition. *J. Account. Res.* **2006**, *44*, 207–242. [CrossRef]
56. Kothari, S.P.; Lewellen, J.; Warner, J.B. Stock returns, aggregate earnings surprises, and behavioral finance. *J. Financ. Econ.* **2006**, *79*, 537–568. [CrossRef]
57. Bilan, Y.; Jurickova, V. Detection of earnings management by different models. In *SHS Web of Conferences*; EDP Sciences: Rajecke Teplice, Slovakia, 2021; Volume 92, p. 02005.
58. Kovalová, E.; Michalíková, K.F. The creative accounting in determining the bankruptcy of Business Corporation. In Proceedings of the SHS Web of Conferences, EDP Sciences, Rajecke Teplice, Slovakia; 2020; Volume 74, p. 01017. Available online: https://www.shs-conferences.org/articles/shsconf/abs/2020/02/shsconf_glob2020_01017/shsconf_glob2020_01017.html (accessed on 9 July 2025).
59. Străpuc, C. The Accuracy of the Message-A Quality of Financial Reporting in the Assessment of the Economic-Financial Reality. *Int. J. Commun. Res.* **2023**, *13*, 294–300.
60. Trevisan, P.; Mouritsen, J. Compromises and compromising: Management accounting and decision-making in a creative organisation. *Manag. Account. Res.* **2023**, *60*, 100839. [CrossRef]
61. Papík, M.; Papíková, L. Automated Machine Learning in Bankruptcy Prediction of Manufacturing Companies. *Procedia Comput. Sci.* **2024**, *232*, 1428–1436. [CrossRef]
62. Káčer, M.; Alexy, M. SMEs' failure prediction: Are machine learning models superior? *Stud. Econ. Financ.* 2025, *ahead of print*. [CrossRef]
63. Lee, H.; Kim, N.W.; Lee, J.G.; Lee, B.T. Performance-related internal clustering validation index for clustering-based anomaly detection. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 20–22 October 2021; IEEE: Piscataway NJ, USA, 2021; pp. 1036–1041.
64. Foorthuis, R. On the nature and types of anomalies: A review of deviations in data. *Int. J. Data Sci. Anal.* **2021**, *12*, 297–331. [CrossRef] [PubMed]
65. Samariya, D.; Thakkar, A. A comprehensive survey of anomaly detection algorithms. *Ann. Data Sci.* **2023**, *10*, 829–850. [CrossRef]
66. Hennig, C. Clustering strategy and method selection. *Handb. Clust. Anal.* **2015**, *9*, 703–730.
67. Van Mechelen, I.; Hennig, C.; Kiers, H.A. Onset of a conceptual outline map to get a hold on the jungle of cluster analysis. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2024**, *14*, e1547. [CrossRef]
68. Şchiopu, D. Applying TwoStep cluster analysis for identifying bank customers' profile. *Buletinul* **2010**, *62*, 66–75.
69. Bolton, R.J.; Hand, D.J. Statistical fraud detection: A review. *Stat. Sci.* **2002**, *17*, 235–255. [CrossRef]
70. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
71. West, J.; Bhattacharya, M. Intelligent financial fraud detection: A comprehensive review. *Comput. Secur.* **2016**, *57*, 47–66. [CrossRef]
72. Goldstein, M.; Uchida, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE* **2016**, *11*, e0152173. [CrossRef]
73. Cai, Y.; Ma, Y.; Yang, H.; Hang, H. Bagged Regularized *k*-Distances for Anomaly Detection. *arXiv* **2023**, arXiv:2312.01046.
74. Bakumenko, A.; Elragal, A. Detecting anomalies in financial data using machine learning algorithms. *Systems* **2022**, *10*, 130. [CrossRef]
75. Crépey, S.; Lehdili, N.; Madhar, N.; Thomas, M. Anomaly detection in financial time series by principal component analysis and neural networks. *Algorithms* **2022**, *15*, 385. [CrossRef]
76. Hernandez Aros, L.; Bustamante Molano, L.X.; Gutierrez-Portela, F.; Moreno Hernandez, J.J.; Rodríguez Barrero, M.S. Financial fraud detection through the application of machine learning techniques: A literature review. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1130. [CrossRef]
77. Li, Y. Anomaly detection in financial data. In *Smart Finance and Contemporary Trade (ESFCT 2023), Proceedings of the 2nd International Conference on Economics, Dali, China, 28–30 July 2023*; Springer Nature: Berlin/Heidelberg, Germany; Volume 261, pp. 419–426. Available online: https://www.researchgate.net/publication/374617144_Anomaly_detection_in_Financial_Data (accessed on 9 July 2025).
78. Thudumu, S.; Branch, P.; Jin, J.; Singh, J. A comprehensive survey of anomaly detection techniques for high dimensional big data. *J. Big Data* **2020**, *7*, 42. [CrossRef]
79. Savran, E.; Karpat, E.; Karpat, F. Energy-Efficient Anomaly Detection and Chaoticity in Electric Vehicle Driving Behavior. *Sensors* **2024**, *24*, 5628. [CrossRef]
80. Bakumenko, A.; Hlaváčková-Schindler, K.; Plant, C.; Hubig, N.C. Advancing Anomaly Detection: Non-Semantic Financial Data Encoding with LLMs. *arXiv* **2024**. arXiv.2406.03614. [CrossRef]
81. Lokanan, M.; Tran, V.; Vuong, N.H. Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian J. Account. Res.* **2019**, *4*, 181–201. [CrossRef]

82. Yıldız, K.; Dedebek, S.; Okay, F.Y.; Şimşek, M.U. Anomaly detection in financial data using deep learning: A comparative analysis. In Proceedings of the 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), Antalya, Turkey, 7–9 September 2022; pp. 1–6. [CrossRef]
83. Majumder, R.Q. A Review of Anomaly Identification in Finance Frauds Using Machine Learning Systems. *Int. J. Adv. Res. Sci. Commun. Technol.* **2025**, *5*, 101–110. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.