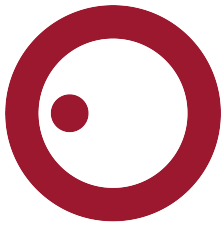

escola superior de tecnologia e gestão
instituto politécnico de leiria

# Anomaly Detection in Numerical Data based on Benford's Law

## Application of Divergence Metrics and Evaluation of Classification Performance

**Patrícia Isabel Santos Martinho**

School of Technology and Management
MSc in Data Science

Leiria, November 2025

*This page intentionally left blank.*

escola superior de tecnologia e gestão
instituto politécnico de leiria

# Anomaly Detection in Numerical Data based on Benford's Law

## Application of Divergence Metrics and Evaluation of Classification Performance

**Patrícia Isabel Santos Martinho**

**Supervisor:** Professor Mário Antunes
*Department of Computer Engineering*
*School of Technology and Management,*
*Polytechnic University of Leiria*

Professor Rui Santos
*Department of Mathematics*
*School of Technology and Management,*
*Polytechnic University of Leiria*

School of Technology and Management
MSc in Data Science

*Project*

Leiria, November 2025

*This page intentionally left blank.*

*This page intentionally left blank.*

# Acknowledgements

*"If you want to go fast, go alone; if you want to go far, go together."* – African Proverb

This proverb perfectly captures the spirit of this journey, which would not have been possible without the invaluable support, encouragement, and guidance of those who walked alongside me.

I would like to start by thanking Professors Rui Santos and Mário Antunes for giving me the opportunity to have them as my advisors and for all the support, guidance, and wisdom shared throughout this journey. I also thank to professor Pedro Fernandes who was always available to share ideas, tips and his knowledge.

A very special thanks to my husband, whose constant support over these two years was crucial. This work would not have been possible without his encouragement, presence, and courage, which allowed me to continue even in the most challenging moments.

I am also deeply grateful to my family, especially my mother, who taught me the importance of persistence and determination. It is thanks to her that I was able to maintain the strength and motivation to keep moving forward, never giving up on my goals.

"For my son, to whom I dedicate this work." *from mom*

*This page intentionally left blank.*

# Resumo

Este projeto centrou-se na deteção de anomalias através da aplicação da lei de Benford, explorando a sua capacidade para identificar desvios estatísticos de forma eficiente e precisa. A abordagem adotada baseou-se nesta lei, amplamente reconhecida pela sua utilidade na deteção de fraudes, especialmente em dados financeiros, ao analisar a distribuição dos primeiros dígitos.

A escassez de dados públicos de qualidade dificultava a avaliação rigorosa de modelos estatísticos. Para superar esta limitação, desenvolveu-se um gerador de dados sintéticos parametrizável, capaz de reproduzir padrões correspondentes tanto a eventos normais como a manipulações realistas. A aplicação desenvolvida permitiu simular condições diversas e aproximar os testes a situações do mundo real, facilitando a análise do desempenho e do comportamento dos métodos estatísticos.

Com os dados simulados obtidos, tornou-se possível avaliar a eficácia de diferentes métodos estatísticos em condições mais próximas da realidade. Neste contexto, a lei de Benford assumiu um papel central, destacando-se pela sua utilidade na deteção de anomalias em múltiplos cenários. Para explorar de forma sistemática esta capacidade, desenvolveu-se um modelo estatístico como alternativa aos modelos tradicionais de *machine learning*, que apresentam elevadas taxas de falsos positivos e grandes exigências computacionais. A proposta assentou na aplicação da lei de Benford combinada com medidas de dissemelhança, permitindo quantificar o desvio entre as distribuições observadas e a distribuição esperada segundo esta lei.

Realizaram-se simulações com o gerador desenvolvido para criar conjuntos de dados conformes e não conformes com a lei de Benford, obtendo-se assim dados classificados. Para medir o desvio, utilizaram-se o qui-quadrado, o desvio médio absoluto, o teste de Kolmogorov–Smirnov, a distância euclidiana, a distância de Hellinger, a divergência de Kullback-Leibler e a combinação dos valores-$p$ dos testes através do método de Fisher. O desempenho das diferentes medidas de dissemelhança foi avaliado com recurso a métricas de classificação como a precisão, *recall* e F1-*score*, os mesmos critérios utlizados em *machine learning*, permitindo comparar o desempenho do modelo em estudo com modelos de *machine learning*. A análise foi complementada pela matriz de confusão e pela curva ROC, ferramentas que permitem uma avaliação mais detalhada do comportamento do modelo, possibilitando a comparação do seu desempenho com o de modelos de *machine learning*.

**Palavras-Chave:** Lei de Benford, detecção de irregularidades, modelos estatísticos, medidas de dissemelhança, avaliação de desempenho, distribuição do 1º dígito, dados sintéticos.

# Abstract

The scarcity of quality public data makes it difficult to rigorously evaluate statistical models. To overcome this limitation, this work develops a parametrizable synthetic data generator capable of reproducing realistic patterns, noises and manipulations. This tool allows you to simulate diverse conditions and approximate the tests to real world situations, facilitating the analysis of performance and behavior of statistical methods.

With the simulated data obtained, it is possible to evaluate the effectiveness of different statistical methods in conditions closer to reality. In this context, Benford's Law assumes a central role, standing out for its usefulness in the detection of anomalies in multiple scenarios. To systematically explore this capacity, a statistical model was developed as an alternative to the traditional models of machine learning which have high false positive rates and large computational requirements. The proposal is based on the application of Benford's Law combined with dissimilarity measures, allowing to quantify the deviation between the observed distributions and the expected distribution according to Benford's law.

Simulations were performed where, using the developed generator, compliant and non-compliant datasets were generated, allowing to obtain classified data. To measure the deviation, we used the chi-square, the mean absolute deviation, the Kolmogorov-Smirnov test, the Euclidean distance, the Hellinger distance, the Kullback-Leibler divergence and the combination of the $p$-values of the tests made through the Fisher method. The performance of the different measures of divergence is evaluated using classification metrics such as precision, *recall* and F1-*score*, the same criteria used in machine learning, which allows to compare the performance of the model under study with machine learning models. The analysis was complemented by the confusion matrix and ROC curve, tools that allow a more detailed evaluation of the behavior of the model, allowing the comparison of its performance with that of machine learning models.

**Keywords:** Benford's law, anomaly detection, statistical models, dissimilarity measures, performance evaluation, first-digit distribution, synthetic data

*This page intentionally left blank.*

# Contents

# List of Figures

*This page intentionally left blank.*

# List of Tables

# Acronyms

| | |
|---|---|
| $\alpha$ | probability of making a Type I error and level of significance |
| $\beta$ | probability of making a Type II error |
| $\chi^2$ | chi-square statistic |
| $q_1$ | first quartile |
| $q_3$ | third quartile |
| **ANN** | artificial neural networks |
| **AUC** | Area Under the Curve |
| **BL** | Benford's Law |
| **CDF** | cumulative distributions function |
| **CNN** | convolutional neural networks |
| **DCT** | discrete cosine transform |
| **DT** | decision trees |
| **FN** | False Negative |
| **FP** | False Positive |
| **FPR** | False Positive Rate |
| $\mathbf{H}_0$ | null hypothesis |
| $\mathbf{H}_1$ | alternative hypothesis |
| **IQR** | interquartile range |
| **J** | Youden's statistic |
| **JPEG** | Joint Photographic Experts Group |
| **KL** | Kullback-Leibler divergence |
| **KS** | Kolmogorov-Smirnov |
| **LIME** | Local Interpretable Model-agnostic Explanations |

**MAD** mean absolute deviation
**ML** machine learning

**ROC** Receiver Operating Characteristic

**SARIMA** Seasonal Autoregressive Integrated Moving Average
**SHAP** SHapley Additive exPlanations
**SMOTE** Synthetic Minority Oversampling TEchnique
**SPC** Statistical Process Control
**SVM** support vector machines

**TN** True Negative
**TP** True Positive
**TPR** True Positive Rate

# 1 Introduction

*"All models are wrong, but some are useful."* - George Edward Pelham Box

This famous quote, often attributed to the British statistician George E. P. Box, encapsulates a fundamental truth about the nature of statistical modelling. Many researchers in the field of statistics seek to develop theoretical models that aim to predict the behavior of certain processes - from the trend of sales of a product to the number of tourists in a city. The essence of this quote lies in the recognition that each model will inevitably be an imperfect approximation of reality, never able to represent exactly the real behavior observed. However, even if you cannot describe reality exactly, a model can be extremely useful if it is close enough to the truth [5].

This imperfection inherent in the models, far from constituting a limitation, can become a powerful tool of discovery. When data deviate systematically from a model's predictions, this divergence can reveal much more than simple inaccuracies and may indicate the presence of hidden anomalies. Statistics is, par excellence, the science that seeks to understand the patterns and variability observed in real-world data, distinguishing between what results from natural variation and what may be a sign of deeper phenomena worthy of investigation.

For this distinction to be effective in practice, it is essential to understand how statistical methods behave in the face of the diversity of scenarios that may arise in real contexts. The need to simulate data with different handling characteristics and degrees of complexity is justified by the natural variability of real data, whose understanding is decisive for the effectiveness of any statistical method.

Indeed, behind the apparent randomness of numbers often hides a surprising mathematical order (or its absence) when unexpected phenomena occur in the data. This variability can have multiple origins: natural processes, human behavior, random errors or intentional manipulations. For statistical methods to be effective in identifying legitimate patterns and, above all, in detecting anomalies, it is essential that their evaluation be carried out in contexts that reflect the complexity and unpredictability of reality.

In this context, Benford's Law, which describes the expected distribution of the first significant digit in datasets, has established itself as a valuable tool in anomaly detection. Its ability to report statistical deviations in areas as diverse as accounting, finance, electoral data or scientific records is due to its robustness in contexts where the data were not intentionally manipulated.

In recent years, there has been growing interest in using natural laws, in particular Benford's Law (BL), to detect anomalies in various datasets. The discovery of this law has an interesting history, which deviates a little from traditional scientific discoveries.

The initial observation that gave rise to Benford's Law dates back to the 19th century. In 1881, the mathematician Simon Newcomb, when studying logarithmic tables, noticed that the numbers on the first pages of these tables (where the first digits were smaller) were easier to read and apparently more used than those on later pages [40]. This led him to formulate the idea that numbers with smaller digits (like 1, 2 and 3) appear more often as first digits than larger numbers (like 8 and 9). This behavior contradicts the common expectation that all digits would occur with the same frequency, making the phenomenon surprising and intellectually stimulating.

Newcomb's initial discovery raised intriguing questions about the behavior of numbers in practical contexts. However, it was only several decades later that this idea would gain greater visibility and scientific foundation, thanks to the work of Frank Benford in 1938, who deepened and formalized the phenomenon based on comprehensive empirical evidence. Benford realized that the phenomenon applied not only to logarithmic numbers, but to a wide range of empirical data. Published an article ([6]), in the journal *Proceedings of the American Philosophical Society*, where he demonstrated that the distribution of the first digits in various datasets (such as geographic areas, populations, financial figures, etc.) follows a specific statistical rule. Benford has proven, through evidence and practical examples, that the first digits of many observed datasets (regardless of the unit of measurement) follow a specific logarithmic distribution.

Since then, the Benford's Law (BL) has aroused growing interest, not only as a mathematical curiosity, but as a statistical tool with practical applications. Indeed, over the following decades, several studies have demonstrated its potential in the analysis of real data, which contributed to BL becoming one of the most promising approaches for detecting anomalies in datasets, for several reasons:

- Successful application in financial audits and accounting for identifying potential fraud or data manipulation.
- Increasing use in diverse areas, such as election fraud detection, scientific data analysis, and big data verification.
- The development of increasingly sophisticated statistical methodologies for applying BL, improving accuracy and reliability.
- The theoretical foundation provided by Hill's Theorem and other mathematical works explaining when and why conformity to BL is expected.

Benford's law, also known as "first digit law", establishes an expected statistical distribution for the initial digits of numbers in many natural datasets. According to this law, smaller digits, such as 1, 2 or 3, tend to occur more frequently than larger digits, such as 8 or 9, following a well-defined mathematical pattern. By comparing the actual data with this theoretical distribution, it is possible to measure the distance between them. Significant deviations from this expected distribution may signal the presence of anomalies, such as manual input errors, fraudulent manipulation or artificial data generation.

Benford's law has been applied to detect anomalies in numerical data, particularly in financial and accounting contexts, with significant methodological evolution over time. This includes the first statistical tests, such as the chi-square, followed by the introduction of statistical distance measures, such as the Kolmogorov-Smirnov distance, by Nigrini [41]. George Judge and Laura Schechter [32] applied statistical dissimilarity techniques to assess the quality of economic data, while Elena Badal-Valero, José A. Alvarez-Jareno and Jose M. Pavía [4] innovated by combining machine learning with dissimilarity measures.

Despite these advances, significant challenges remain, such as the natural non-conformity of many datasets with the Benford distribution due to structural and contextual factors, data quality issues and the need for integrated approaches in complex scenarios. Opportunities for research include the development of optimized dissimilarity measures, multi-digit analysis, domain-specific adaptations, creation of adaptive thresholds, integration with other analytical techniques, evaluation of robustness against purposeful manipulation, improvement of interpretability, implementation in real-time detection, exploration of the theoretical limitations of the method and validation with confirmed cases of fraud, thus representing a favourable area for theoretical and practical advances.

In this context, the main objectives of this research are defined in order to respond to some of these challenges and contribute to the development of more effective and robust approaches in anomaly detection based on Benford's Law.

The main objectives of this research are:

- To create a controlled synthetic data generator to simulate different standards of compliance and non-compliance with Benford's law.
- To implement a line-by-line approach to verify compliance with the Benford Law, allowing direct identification of instances that present anomalies. Unlike traditional approaches where compliance verification is performed at the level of the dataset as a whole, this study proposes a line-by-line approach, allowing the identification of specific instances with irregular behavior. This granularity offers significant practical advantages, namely in the precise location of potential frauds or errors.
- To evaluate the performance of different measures of divergence, such as chi-square test, mean absolute deviation, Kolmogorov-Smirnov test, Euclidean dis-

tance, Hellinger distance, Kullback-Leibler divergence and Fisher's method to combine $p$-values.
- To measure the effectiveness of the detection model using metrics such as precision, *recall* and F1-*score*, complemented by the analysis of the confusion matrix;
- To investigate more efficient alternatives in terms of computing to traditional models of machine learning, through the use of statistical methods based on Benford's Law, with a view to developing lightweight, fast and interpretable solutions.
- To evaluate a statistical approach that can be combined with machine learning techniques, creating hybrid models that increase the flexibility and accuracy of anomaly detection.
- To propose one more accurate and reliable tool for auditors, accountants, and investigators, enhancing fraud detection across multiple areas.

The effectiveness of Benford's Law depends on its ability to be applied to situations that reflect the challenges of real data, not just to idealized or artificial cases. This requires a deep understanding of the variability inherent in empirical data and the importance of statistical methods that can handle the diversity and uncertainty present in real-world applications.

Thus, it became essential to develop a data generator capable of simulating varied and controlled scenarios, allowing a rigorous evaluation of the robustness and sensitivity of the tests applied to Benford's Law.

To replicate data following this law, we used the expression $x = 10^{uniform}$, an already established technique to generate first digit distributions according to Benford's Law. This is based on the principle that if the logarithm of a variable (in base 10) is distributed evenly, then its first digits will follow the Benford distribution [30].

On the other hand, to simulate data that does not follow Benford's Law, parameterizable expressions capable of generating alternative distributions were developed, including uniform distributions, gaussian noise, uniform noise and outliers.

There are some scientific works that describe synthetic data generators to test compliance with Benford's Law, including the generation of compliant and non-compliant data. Although the authors do not always explicitly call "generator", several studies develop experimental methods or simulators. For example, Nigrini in [42] describes how to generate synthetic datasets that follow or violate the expected distribution, and uses this to test fraud detection techniques. Another example, Durtschi, C., Hillison, W. and Pacini, C. in [19] describe examples of Benford-compliant data and cases with simple manipulations (digit changes) to simulate frauds. Although there are already several studies where custom data generation simulation routines are created to test compliance with the Benford's Law, there are not many articles dedicated to the development of automated and parameterizable generators for data according to Benford's law and non-conforming. This project fills this gap. The developed data generator works as a benchmarking tool, enabling the creation of synthetic datasets with well-

defined characteristics (data following the Benford distribution, data following a uniform distribution, Gaussian noise, uniform noise and outliers), which allows to test, compare and validate the performance of different models with different objectives.

This type of controlled data generation is essential not only to test compliance with Benford's Law, but also to simulate more realistic scenarios where anomalies occur. In practice, such anomalies in a set of numerical data can have several sources, such as human errors (for example, mistyping or misformatting), intentional manipulations with fraudulent purposes, technical problems (such as sensor failures, data corruption, synchronization errors or poorly made integrations between different sources), or even the natural variability of the data, as occurs in physical measurements where the presence of noise is common.

Anomalies in a numeric dataset can have various sources, such as human errors (for example, incorrect typing or misformatting), intentional manipulations with fraudulent purposes, technical problems (such as sensor failures, data corruption, synchronization errors or poorly made integrations between different sources) or even the natural variability of the data, as occurs in physical measurements, where the presence of noise is common.

A concrete example can be observed in digital images, which can also be represented as sets of numerical features, essentially matrices (or tensors) of intensity values per channel (R, G, B, etc.). When transforming a dataset of images into numerical values, specific potential anomalies arise. For example, manipulated images (such as deepfakes) may show changes in pixel values unnaturally due to cloning, excessive smoothing, or other artificial edits. Another relevant example is that of JPEG (Joint Photographic Experts Group) compression with severe losses. Because this format uses Discrete Cosine Transform (DCT) to transform pixel blocks into frequency values, the resulting coefficients can be quantified in a way that eliminates visual details. When compression is excessive or applied multiple times, the DCT coefficients fail to follow the expected statistical distribution, which may indicate recompression, tampering, or image manipulation.

Therefore, anomalies detection is an important research topic, with applications in various fields including information security, financial auditing and industrial quality control. The identification of anomalous patterns in data is essential to prevent fraud, cyber attacks and other anomalies that can compromise the integrity of systems.

Traditional models of anomaly detection are often based on machine learning techniques such as artificial neural networks (ANNs), convolutional neural networks (CNNs) and support vector machines (SVMs), among others. However, these approaches present significant challenges, including high false positive rates and computationally intensive requirements, which can hinder the efficiency of real-time analysis. On the other hand, statistical models emerged as a viable alternative due to their speed, ease of implementation and accessibility.

Given these limitations of traditional models, it becomes relevant to explore lighter and interpretable statistical approaches, especially in contexts where simplicity and

efficiency are fundamental. In this sense, this project proposes a methodology based on a complete pipeline, from data generation to unsupervised classification and model evaluation, with the objective of comparing and validating different scenarios.

The controlled synthetic data generator allowed to simulate different patterns of compliance and non-compliance with Benford's law, in order to create a classified dataset, essential for the validation of the proposed system. It allows control over data quality and the injection of noise to simulate real-world conditions. Initial analyses revealed that the publicly available datasets did not fully conform to the expected Benford distribution, which led to the development of this solution. The creation of the synthetic data generator makes a significant contribution by providing rigorous control over the data used to evaluate the model. The generation was designed in a parametrizable way, allowing the manipulation of several variables such as number of instances, number of attributes and type of anomalies introduced.

An unsupervised classification model based on statistical tests of conformity with the theoretical distribution of Benford's Law was also developed.

To support this approach, divergence metrics play a central role in quantifying the difference between the observed distribution of digits and the expected distribution according to Benford's Law. Measures such as the chi-square test, mean absolute deviation, Kolmogorov-Smirnov statistic, Euclidean distance, Hellinger distance and Kullback-Leibler divergence were used. The consolidation of the results obtained by these metrics, through the Fisher methodology based on individual $p$-values, allows to gather robust statistical evidence about the compliance of the analyzed data.

Based on these metrics, the model was designed to evaluate each instance of the dataset individually. Instead of resorting to aggregate analysis, as the traditional methods of applying Benford's Law do, this methodology adopts a more granular approach. The proposed statistical model applies the principles of the law at the level of individual record, allowing the immediate detection of anomalies in specific transactions or entries. This record level analysis allows to accurately locate the exact points where statistically significant deviations occur, offering a clear advantage over conventional methodologies, which only indicate potential anomalies globally. Thus, this model provides a more effective and practical detection capacity, with potential for real-time applications and with specific location of anomalies.

The pipeline also includes evaluating model performance through classic metrics. To rigorously validate the performance of this model, it was applied the same evaluation metrics used in machine learning models. The results were systematized in confusion matrices, allowing a direct and objective comparison between this statistical approach based on Benford's Law and conventional ML algorithms. This standardized assessment framework not only quantifies the accuracy, sensitivity and specificity of this model, but also facilitates comparative benchmarks with other fraud detection techniques, establishing a complete picture of its relative effectiveness. However, it is important to discuss the particularities involved in applying these metrics to an unsupervised nature model.

Although the classification model developed is of unsupervised nature, the use of synthetic data with prior labeling allowed the evaluation of its performance through typical metrics of supervised classification, such as the confusion matrix. However, this approach involves an important conceptual misalignment: while the model identifies anomalies without previously knowing the classes, the confusion matrix assumes explicit knowledge of the true classes to compare the decisions of the model. This discrepancy is overcome in this experimental context, since the generated data provide labels that serve as a reference to validate the model's ability to detect deviations from Benford's Law. The definition of thresholds to convert test statistics into binary decisions, although necessary for the construction of the matrix, can introduce arbitrarities that influence performance metrics. In addition, the statistical nature of the method implies that subtle manipulations can go unnoticed, generating false negatives that reflect the limits of the sensitivity of the model and not the failures themselves. The unbalanced distribution between handled and unhandled cases also impacts the interpretation of metrics, making it essential to analyze complementary measures such as precision, *recall*, and F1-*score*. Despite these limitations, a robust experimental validation provides the basis for practical application of the model in real situations where labels are not known, allowing efficient detection of anomalies in new data in an unsupervised way. To ensure this robustness in the evaluation, this work resorted to controlled simulations that generate datasets both compatible and not compatible with Benford's Law, enabling a detailed examination of the effectiveness of divergence metrics in identifying anomalies.

In this work, simulations were performed in which datasets compatible and not compatible with Benford's Law are generated, allowing a controlled evaluation of the effectiveness of divergence metrics in detecting anomalies. To measure the performance of detection models, common evaluation metrics are used for unbalanced datasets applied in data science, such as precision, *recall*, F1-*score* and confusion matrix.

The results obtained showed that the performance of statistical methods depends essentially on the number of cases treated, the proportion of changes per row and the dimensionality of the data. The results indicate that performance improves with increasing sample size, stabilizing from about 600 rows and 1100 columns. Methods such as the chi-square test and some divergence metrics showed good precision and balance between accuracy and *recall*, although some tests, such as the Fisher combination, have shown greater sensitivity and tendency to false positives. In scenarios with low proportion of manipulations or subtle changes, the overall performance decreases, highlighting the need for complementary approaches to ensure reliable detections.

The results obtained during this work reinforce the relevance and usefulness of the proposed contributions. The finding that the performance of statistical methods strongly depends on the number of cases handled, the proportion of changes per row and the dimensionality of the sample showed the need for a synthetic data generator robust and flexible. This tool allowed the simulation of controlled and varied scenarios, enabling comparative evaluation between different approaches for detecting anoma-

lies, which responds directly to the limitation of previous studies that were based on real data without control over the parameters. In addition, the tests revealed limitations in cases with subtle manipulations or reduced samples, underlining the importance of exploring the theoretical limits of detection based on Benford's Law as the minimum sample size. The development of a deviation-sensitive classification algorithm, capable of acting at the level of each row, allowed not only to identify the presence of anomalies, but also to locate them with greater precision, filling a gap in previous approaches. The diversity of statistical metrics analyzed also demonstrated that different measures capture distinct aspects of deviations from the Benford distribution, which validates the option to integrate multiple tests and combinations, such as the Fisher method. Finally, the results obtained show that the proposed solution is computationally efficient, scalable and adaptable to different fields of application, which paves the way for its integration into existing audit systems, as well as the creation of hybrid models that combine statistical and machine learning techniques, promoting more flexible and accurate approaches to anomaly detection.

Based on these results, it is possible to identify several relevant contributions of this work for research and practical application in anomaly detection:

- a benchmarking tool, allowing the creation of synthetic datasets with well-defined characteristics (data following Benford distribution, following uniform distribution, gaussian noise, uniform noise and outliers), which allows testing, comparing and validating the performance of different models with different objectives;
- the synthetic data generator developed in this project is a valuable contribution to the research community, providing a benchmarking tool for evaluating and comparing different fraud detection methods, thereby encouraging the development of new approaches and solutions (https://github.com/PatriciaMartinho/Anomaly-Detection-in-Numerical-Data/);
- creation of an open source solution to promote access within the scientific community and support replicability of results;
- creation of a deviation-sensitive classification algorithm in the first digit distributions;
- identification of the rows where the anomaly is, as previous studies focus on finding fraud in the dataset without specifying the location;
- exploration of the theoretical limits of BL-based detection, as minimum sample size;
- investigation of how different dissimilarity measures capture different aspects of the Benford distribution deviations;
- proposed a more accurate and reliable tool for auditors, accountants and investigators, improving fraud detection in various areas;
- creation of an adaptable framework for application in various fields, including finance, economics and social sciences;
- implementation of a more efficient computational model that allows the analysis of large datasets in real time;

- development of a scalable algorithm suitable for integration into existing audit systems;
- a statistical approach model that can be combined with machine learning techniques, creating hybrid models that increase the flexibility and accuracy of anomaly detection.

In short, the results obtained throughout this work validate the proposed statistical approach, evidencing its potential for anomaly detection in different contexts. Through the combination of controlled simulations, sensitivity analysis and development of dedicated tools, it was possible to explore the limits and capabilities of the application of Benford's law. The contributions presented here not only reinforce the relevance of this approach but also pave the way for new investigations and practical applications in real scenarios, namely in areas where the detection of anomalies is critical.

This study belongs to an academic research context and aims to contribute to the advancement of knowledge in the field and explore new approaches for real-time anomaly detection using Benford's Law as a central analytical tool.

# 2
# Background and Related Work

This chapter begins with a demystification of Benford's Law, addressing the mathematical bases that explain its occurrence in naturally generated data, contrary to the perception that it is an enigmatic or exceptional phenomenon. The main practical applications of the law are then explored, as well as its limitations, namely the requirements and conditions under which the Benford distribution is reliably applicable.

It presents the divergence measures used as tools to quantify the deviation between the observed distribution of digits and Benford's law. The role of hypothesis tests in the statistical evaluation of data conformity is discussed, with emphasis on the combination of $p$-values through the Fisher methodology.

The chapter continues with a reflection on the concept of detection of anomalies, contextualizing it within the scope of data science and discussing the main existing approaches, from classical statistical techniques to models based on machine learning. Finally, the state of the art regarding the application of Benford's Law and measures of divergence in the detection of anomalies is presented, identifying gaps and opportunities still little explored in the literature, as the line-by-line analysis proposed in this work, which aims to improve the accuracy and practical usefulness of existing methodologies.

## 2.1 The Benford's Law

Benford's Law can be used as a statistical approach for the identification of expected patterns or detection of anomalies in numerical datasets. It is an unsupervised method that can be used to detect anomalies by analyzing the distribution of the digits in numerical values. This property is particularly valuable in contexts where there is no clear label on what constitutes anomaly and where supervised machine learning techniques become impractical. Since naturally generated datasets generally exhibit the digit distribution predicted by Benford's Law, significant departures from this pattern can signal potential anomalies. In addition, Benford's Law offers a light approach from the computational point of view, facilitating its application in resource-limited environ-

ments or in analyzing large volumes of data. In this work, its innovative application at the record level further reinforces its usefulness, allowing a detailed and localized analysis of anomalies, which is not common in traditional approaches.

Benford's Law has aroused the interest of researchers from various areas not only for its practical usefulness, but also for the apparently counterintuitive and even mysterious nature of its manifestation. The fact that many real datasets, coming from such different areas as finance, geography, physics or demography, spontaneously follow a logarithmic distribution in the digits challenges the common statistical intuition, which would expect a uniform distribution. This emerging regularity raises profound questions about the structure of natural numerical data and about the mechanisms underlying its generation.

### 2.1.1 Demystifying Benford's Law from a mathematical perspective

Throughout the history of mathematics, it is not uncommon to come across discoveries that surprise, whether by challenging intuition, the curious way they were found or the unexpected applications they make possible. Among these discoveries, the so-called Benford's Law stands out. Emerging from a seemingly casual observation, this law continued, over more than a century, to reveal remarkable properties and eventually consolidated itself as a powerful tool in anomaly detection.

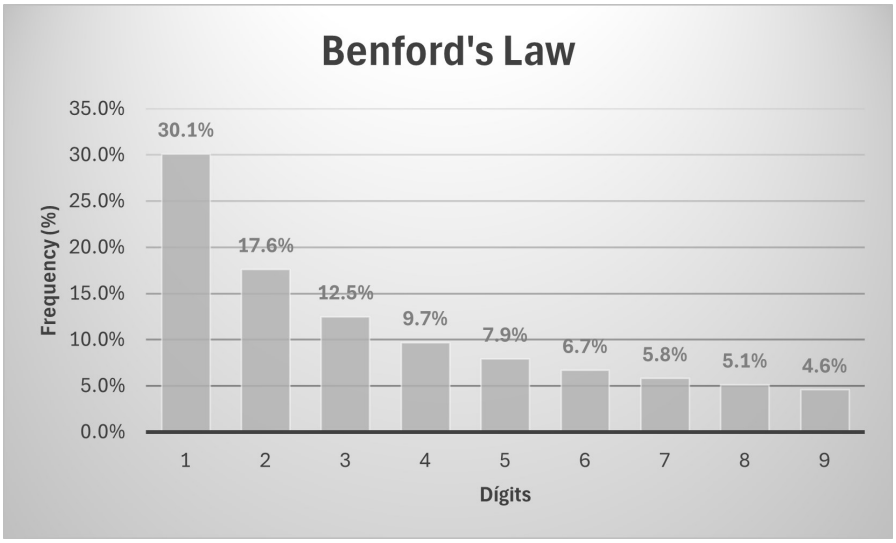


**Figure 2.1:** *Frequency of digits 1 through 9 that is expected to occur in first position*

Benford's Law (BL) is the observation that, in many collections of numbers, whether they originate from mathematical tables, real-world datasets, or a combination of both, the leading significant digits are not uniformly distributed, as might be intuitively expected. Instead, there is a marked tendency for lower digits to appear more frequently as the leading digits [7]. Briefly explained, Benford's Law maintains that the numeral 1 will be the leading digit in a genuine dataset of numbers 30.1% of the time; the numeral 2 will be the leading digit 17.6% of the time; and each subsequent numeral, 3 through

9, will be the leading digit with decreasing frequency. This expected occurrence of leading digits can be depicted in the Figure 2.1. [13]

First observed by Newcomb [40], Benford's Law only gained prominence 57 years later, following a new publication by Benford [6]. In its simplest form, this law states that the first significant digit of a number drawn from a nearly arbitrary statistical dataset does not follow a uniform distribution over the set $\{1, \cdots, 9\}$. On the contrary, according to this law, the digit 1 appears far more frequently, while the digit 9 is the least common [28], see Figure 2.2.

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

| Group | Title | First Digit | | | | | | | | | Count |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| A | Rivers, Area | 31.0 | 16.4 | 10.7 | 11.3 | 7.2 | 8.6 | 5.5 | 4.2 | 5.1 | 335 |
| B | Population | 33.9 | 20.4 | 14.2 | 8.1 | 7.2 | 6.2 | 4.1 | 3.7 | 2.2 | 3259 |
| C | Constants | 41.3 | 14.4 | 4.8 | 8.6 | 10.6 | 5.8 | 1.0 | 2.9 | 10.6 | 104 |
| D | Newspapers | 30.0 | 18.0 | 12.0 | 10.0 | 8.0 | 6.0 | 6.0 | 5.0 | 5.0 | 100 |
| E | Spec. Heat | 24.0 | 18.4 | 16.2 | 14.6 | 10.6 | 4.1 | 3.2 | 4.8 | 4.1 | 1389 |
| F | Pressure | 29.6 | 18.3 | 12.8 | 9.8 | 8.3 | 6.4 | 5.7 | 4.4 | 4.7 | 703 |
| G | H.P. Lost | 30.0 | 18.4 | 11.9 | 10.8 | 8.1 | 7.0 | 5.1 | 5.1 | 3.6 | 690 |
| H | Mol. Wgt. | 26.7 | 25.2 | 15.4 | 10.8 | 6.7 | 5.1 | 4.1 | 2.8 | 3.2 | 1800 |
| I | Drainage | 27.1 | 23.9 | 13.8 | 12.6 | 8.2 | 5.0 | 5.0 | 2.5 | 1.9 | 159 |
| J | Atomic Wgt. | 47.2 | 18.7 | 5.5 | 4.4 | 6.6 | 4.4 | 3.3 | 4.4 | 5.5 | 91 |
| K | $n^{-1}, \sqrt{n}, \cdots$ | 25.7 | 20.3 | 9.7 | 6.8 | 6.6 | 6.8 | 7.2 | 8.0 | 8.9 | 5000 |
| L | Design | 26.8 | 14.8 | 14.3 | 7.5 | 8.3 | 8.4 | 7.0 | 7.3 | 5.6 | 560 |
| M | *Digest* | 33.4 | 18.5 | 12.4 | 7.5 | 7.1 | 6.5 | 5.5 | 4.9 | 4.2 | 308 |
| N | Cost Data | 32.4 | 18.8 | 10.1 | 10.1 | 9.8 | 5.5 | 4.7 | 5.5 | 3.1 | 741 |
| O | X-Ray Volts | 27.9 | 17.5 | 14.4 | 9.0 | 8.1 | 7.4 | 5.1 | 5.8 | 4.8 | 707 |
| P | Am. League | 32.7 | 17.6 | 12.6 | 9.8 | 7.4 | 6.4 | 4.9 | 5.6 | 3.0 | 1458 |
| Q | Black Body | 31.0 | 17.3 | 14.1 | 8.7 | 6.6 | 7.0 | 5.2 | 4.7 | 5.4 | 1165 |
| R | Addresses | 28.9 | 19.2 | 12.6 | 8.8 | 8.5 | 6.4 | 5.6 | 5.0 | 5.0 | 342 |
| S | $n^1, n^2 \cdots n!$ | 25.3 | 16.0 | 12.0 | 10.0 | 8.5 | 8.8 | 6.8 | 7.1 | 5.5 | 900 |
| T | Death Rate | 27.0 | 18.6 | 15.7 | 9.4 | 6.7 | 6.5 | 7.2 | 4.8 | 4.1 | 418 |
| Average..... | | 30.6 | 18.5 | 12.4 | 9.4 | 8.0 | 6.4 | 5.1 | 4.9 | 4.7 | 1011 |
| Probable Error | | ±0.8 | ±0.4 | ±0.4 | ±0.3 | ±0.2 | ±0.2 | ±0.2 | ±0.2 | ±0.3 | — |

**Figure 2.2:** *Benford's original data* [6]

More specifically, Benford's Law states that the significant digits in many datasets follow a specific logarithmic distribution. In its most common form, referring to the first significant digits in base 10, wich will be denoted by $D_1$, this law is also known as the First-Digit Law and is expressed by Equation (2.1) [7]

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right), \tag{2.1}$$

where $d_1 \in \{1, 2, 3, \ldots, 9\}$.

The Benford's Law is considered mysterious for several reasons, mainly due to its unexpected applicability to various datasets and its counter-intuitive nature. Here are some of the main factors that contribute to this mystery [28]:

- *Counter-intuitive distribution of leading digits:* According to Benford's Law, num-

bers that start with the digit "1" are much more frequent than numbers starting with "9". For example, about 30% of numbers in real-world datasets begin with the digit "1", while only around 5% start with the digit "9". This contradicts the expectation that all digits should occur with the same probability.

- *Universal applicability:* Benford's Law is not limited to a specific type of data. It applies to a wide variety of real-world datasets, such as city populations, stock prices, scientific measurements, and even bank account numbers. The idea that a simple mathematical rule can apply so broadly feels mysterious and counter-intuitive.
- *Lack of a simple explanation:* Although Benford's Law was discovered by Frank Benford in 1938, it doesn't have a simple or intuitive explanation. Some factors, like logarithmic distribution and the scale transformation of numbers, can explain the law in mathematical terms, but this still feels mysterious to many, especially when applied to real-world contexts.
- *Scale-invariance property:* Benford's Law is invariant under scale transformation, meaning it applies to numbers in any unit of measurement, as long as the data is well-distributed and not bound by a fixed lower or upper limit. This strange behaviour, where the law holds regardless of the scale of the data, is part of what makes Benford's Law seem "mysterious".

Despite all the mystery associated with Benford's Law, it can be easily explained by analyzing the logarithms of numbers and observing the behaviour of their fractional (or decimal) part (mantissas).

Therefore, the mantissa is the fractional part of the logarithm of a number in a given base. In other words, if we take the logarithm of a real positive number $x$, it can be separated into the integer part plus the fractional part.

Benford's law is closely linked to the distribution of the mantissas of the logarithms of numbers. One of the fundamental properties of Benford's Law is that these mantissas are uniformly distributed over the interval $]0,1[$.

This means that, when analyzing a dataset that conforms to Benford's Law, the logarithms of these numbers will have mantissas uniformly distributed between 0 and 1. This behavior provides the mathematical basis for explaining why the leading digits of these numbers follow Benford's distribution.

If a set of positive numbers, $x$, is distributed such that their logarithms follow an approximately uniform distribution across several orders of magnitude, then the logarithm of a number can be expressed as

$$\log_{10}(x) = k + M, \tag{2.2}$$

where $k = \text{Int}(\log_{10}(x))$ and $M = \text{Frac}(\log_{10}(x))$ which is the mantissa in the interval $]0, 1[$ ($\text{Int}(x)$ denotes the integer part of $x$, i.e., the value obtained by truncating the decimal part of $x$, and $\text{Frac}(\text{x})$ represents the decimal part of $x$). When converting back from the logarithm to the original numbers, is obtained

$$x = 10^{k+M} = 10^k \times 10^M. \tag{2.3}$$

The term $10^M$ controls the first $M$ digits. Mantissas $M$ are uniformly distributed over the interval $]0,1[$, the factor $10^M$ assumes values logarithmically distributed between 1 and 10. This means that the probability of $M$ having a first digit $d_1$ is equal to the probability that $10^M$ falls within the interval

$$d_1 \leq 10^M < d_1 + 1.$$

This way, it is concluded that

$$\log_{10}(d_1) \leq M < \log_{10}(d_1 + 1).$$

As $M$ is uniformly distributed between 0 and 1, the fraction of values that fall between $\log_{10}(d_1)$ and $\log_{10}(d_1 + 1)$ is exactly

$$P(D_1 = d_1) = \log_{10}(d_1 + 1) - \log_{10}(d_1) = \log_{10}\left(\frac{d_1 + 1}{d_1}\right) = \log_{10}\left(1 + \frac{1}{d_1}\right). \tag{2.4}$$

When the analyzed values assume negative values, to test whether they are in accordance with Benford's Law, the symmetrical value is used (using the module to eliminate negative values).

However, Benford's Law is not limited to the first digit and it can also be extended to the second digit or beyond. For the second digit or beyond, the distribution is also non-uniform, although it is less skewed than for the first digit. Hence, using analogous reasoning, the probability of the second digit $D_2$ being equal to $d_2 \in \{0, 1, 2, \cdots, 9\}$ is given by

$$P(D_2 = d_2) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{10 \times d_1 + d_2}\right), \tag{2.5}$$

where the probabilities of the second digit being equal to $d_2$ and the first assuming all possibilities from 1 to 9 are added together, since the probability of simultaneously $D_1 = d_1$ and $D_2 = d_2$ is given by

$$P(D_1 = d_1, D_2 = d_2) = \log_{10}\left(1 + \frac{1}{10 \times d_1 + d_2}\right). \tag{2.6}$$

For example, the probability of the second digit being zero is given by

$$\begin{aligned} P(D_2 = 0) &= \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{10 \times d_1 + 0}\right) \\ &= \log_{10}\left(1 + \frac{1}{10}\right) + \log_{10}\left(1 + \frac{1}{20}\right) + \cdots + \log_{10}\left(1 + \frac{1}{90}\right) \\ &\approx 0.1196793. \end{aligned}$$

In general, the probability of the $k$-th digit $D_k$, with $k \geq 2$, being being equal to $d_k \in \{0, 1, \cdots, 9\}$ is given by

$$P(D_k = d_k) = \sum_{n=10^{k-2}}^{10^{k-1}-1} \log_{10}\left(1 + \frac{1}{10n + d_k}\right). \tag{2.7}$$

Here, $D_2, D_3, D_4, \cdots$ represent the second, third, fourth, and so on, significant digits.

From the fourth or fifth significant digit onwards, the frequencies become very close to 10%, tending toward a uniform distribution. In other words:

- *First digit:* highly skewed (ex: $P(D_1 = 1) \approx 30.103\%$).
- *Second digit:* skewed, but less so (digits 0 to 9 with frequencies between 8% and 12%).
- *Third digit and beyond:* gradual convergence toward 10%.

Table 2.1 presents the approximate probabilities (with 5 decimal places) for each possible result for the first $n$ significant digits according to Benford's Law, illustrating how the distribution progressively approaches uniformity as $n$ increases

| Digit | 1st place | 2nd place | 3th place | 4th place | 5th place | 6th place |
|---|---|---|---|---|---|---|
| 0 | — | 0.11968 | 0.10178 | 0.10018 | 0.10002 | 0.10000 |
| 1 | 0.30103 | 0.11389 | 0.10138 | 0.10014 | 0.10001 | 0.10000 |
| 2 | 0.17609 | 0.10882 | 0.10097 | 0.10010 | 0.10001 | 0.10000 |
| 3 | 0.12494 | 0.10433 | 0.10057 | 0.10006 | 0.10001 | 0.10000 |
| 4 | 0.09691 | 0.10031 | 0.10018 | 0.10002 | 0.10000 | 0.10000 |
| 5 | 0.07918 | 0.09668 | 0.09979 | 0.09998 | 0.10000 | 0.10000 |
| 6 | 0.06695 | 0.09337 | 0.09940 | 0.09994 | 0.09999 | 0.10000 |
| 7 | 0.05799 | 0.09035 | 0.09902 | 0.09990 | 0.09999 | 0.10000 |
| 8 | 0.05115 | 0.08757 | 0.09864 | 0.09986 | 0.09999 | 0.10000 |
| 9 | 0.04576 | 0.08500 | 0.09827 | 0.09982 | 0.09998 | 0.10000 |

**Table 2.1:** *Expected Frequencies Based on Benford's Law*

Benford's Law can be generalized not only to each digit, but to the combination of the first $n$ significant digits. In this case, the probability of a number starting with a specific sequence of digits $d_1, d_2, \cdots, d_n$ (where the first digit $d_1 \in \{1, 2, \cdots, 9\}$ and the following $d_j$, with $j = 2, \ldots, n$, can be from $d_j \in \{0, 1, 2, \cdots, 9\}$) is given by

$$P(D_1 = d_1, D_2 = d_2, \cdots, D_n = d_n) = \log_{10}\left(1 + \frac{1}{N}\right), \tag{2.8}$$

where

$$N = d_1 \times 10^{n-1} + d_2 \times 10^{n-2} + \cdots + d_n = \sum_{i=1}^{n} d_i \times 10^{n-i}.$$

In general terms, it represents a statement about the joint distribution of all decimal digits. Hence, the following equation holds for any positive integer $n$

$$P(D_1 = d_1, D_2 = d_2, \cdots, D_n = d_n) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{n} d_i \times 10^{n-i}}\right). \tag{2.9}$$

For example, if we want to know the probability of a number starting with 23, just substitute $d_1 = 2$ and $d_2 = 3$ into the formula

$$P(D_1 = 2, D_2 = 3) = \log_{10}\left(1 + \frac{1}{23}\right) \tag{2.10}$$

and we will get

$$P(23) \approx \log_{10}(1.043478) \approx 0.01848.$$

In other words, about 1.85% of the numbers in a set that follows Benford's Law will have the first two digits 23.

The three fundamental properties of Benford's Law, which help explain its occurrence across many real-world datasets, are as follows [8]:

- *Scale Invariance:* If a dataset follows Benford's Law, it will continue to follow the law even if all values are multiplied by a positive constant. This means that the distribution of leading digits is independent of the unit of measurement used.
- *Base Invariance:* Benford's Law can be formulated for any numerical base $b$, while preserving the probabilistic structure of the leading digits.
- *Uniform Distribution of Logarithmic Mantissas in the Interval (0,1):* This property arises from the relationship between Benford's Law and the distribution of the fractional parts (mantissas) of the logarithms of numbers, as previously demonstrated.

These fundamental properties of Benford's Law (scale invariance, base invariance and invariance under certain transformations) provide a theoretical framework for understanding the mathematical behaviour of this distribution. However, for many years, a rigorous statistical explanation was lacking to support the widespread occurrence of Benford's Law in real-world data. It was in this context that the work of Theodore P. Hill emerged as a decisive contribution, offering a statistical derivation that explains why natural datasets follow this apparently counter-intuitive law.

In [30], Theodore P. Hill presents a mathematical derivation of Benford's Law. This important theoretical contribution offers a new perspective for understanding why many natural datasets follow this peculiar statistical law. Hill develops a different approach from previous ones, which were mainly based on scale invariance or considerations of logarithm mantissas. Instead, he uses a new concept, distribution of distributions, to explain why Benford's Law is present in so much real-world data. The article establishes what has become known as the Hill's Theorem, demonstrating that when numbers are randomly selected from multiple statistical distributions (as often occurs in nature and human processes), the resulting distribution of the first significant digits tends to follow Benford's Law.

Theodore P. Hill [30] develops a statistical model of what he calls the "natural probability space" for the significance of digits, proving that this space leads naturally to the Benford distribution. Using measure and probability theories, the article formalizes the precise conditions under which Benford's Law emerges as an inevitable mathematical consequence. This work explains why Benford's Law appears in such diverse datasets, from physical constants to financial records, and provides a solid mathematical basis for using this law to detect anomalies in data. Theodore P. Hill [30] demonstrates that the Benford distribution is the only digit distribution that remains invariant under certain transformations and changes of scale, explaining its remarkable stability in real data. The article also clarifies the conditions under which we can expect data to follow Benford's Law, helping to avoid misapplications. The author employed an approach based on measure theory, using the concept of "random mixing" applied to significant digits, and established a countably additive probabilistic framework for the analysis of significant digits, using Borel sets and invariant measures to prove his results. This work represented a fundamental evolution in the theoretical understanding of Benford's Law, transforming it from a mere mathematical curiosity into a well-founded statistical phenomenon with important practical applications in data analysis and anomaly detection.

### 2.1.2 Applications and Limitations of Benford's Law in Practice

The Benford's Law, while initially perceived as a mathematical curiosity, has found numerous practical applications in diverse fields. One of its most prominent usage is in forensic accounting and financial auditing [13, 18, 19, 41, 42, 43], where it serves as a tool for detecting anomalies and potential fraud. Financial records that deviate significantly from the expected distribution of leading digits may indicate data manipulation or fabrication.

In addition to financial applications, Benford's Law has been applied to anomaly detection and data analysis in other domains such as electoral fraud detection [16], demographic and census data [32] verification and validation of datasets [17, 49]. Deviations from the expected distribution can reveal inconsistencies, errors, or deliberate data manipulation.

Another important application is in anomaly detection within information systems [14, 22, 25, 33]. By monitoring the distribution of values in operational logs or database records, anomalous patterns that do not conform to Benford's distribution may signal security breaches or abnormal system behaviours.

The law is also employed in validating the outputs of predictive models and simulations. In situations where simulated data is expected to reflect natural phenomena, conformity to Benford's Law can be used as a diagnostic measure for assessing the plausibility of results.

Moreover, public sector auditing and control bodies increasingly apply Benford's Law to evaluate tax declarations, budgetary reports, and public spending records [42,

43]. The methodology has also been extended to emerging domains such as cryptocurrency transaction analysis, where the irregular distribution of significant digits can indicate potential manipulative activities [54].

These applications highlight the versatility of Benford's Law as a practical statistical tool for identifying anomalies, validating data integrity, and supporting decision-making processes in both public and private sectors.

While the practical applications of Benford's Law are numerous and diverse, its widespread occurrence across seemingly unrelated datasets naturally raises a fundamental theoretical question: *why does this law appear so frequently in real-world data?*. Addressing this question requires a deeper understanding of the mathematical mechanisms underlying the law. A significant breakthrough in this regard was achieved by Theodore P. Hill, whose seminal theorem provides a rigorous probabilistic explanation for the emergence of Benford's Law in many empirical contexts. Hill's work formalizes the conditions under which the distribution of significant digits converges to Benford's Law, particularly through the concepts of scale invariance, base invariance, and the mixing of distributions.

The Hill's Theorem [30] establishes a precise mathematical mechanism that explains why Benford's Law naturally emerges in many datasets, and its main mechanisms are:

- *Base Invariance:* The theorem demonstrates that the only digit distribution that is invariant under changes in base (such as converting numbers from base 10 to base 8) is Benford's distribution. This property is fundamental to understanding why Benford's Law is so universal.
- *Scale Invariance:* Hill demonstrated that Benford's distribution is the only digit distribution that remains invariant under multiplication by positive constants. This means that if a dataset follows Benford's Law and all values are multiplied by a constant, the distribution of the leading digits will remain the same.
- *Mixture of Distributions:* The central mechanism of Hill's theorem is that when numbers are randomly selected from a mixture of distributions (i.e., when numbers from multiple different distributions are combined), the resulting distribution of the leading digits converges to the Benford's Law. Mathematically, this means that if we have several different probability distributions, and we randomly select one of these distributions and then randomly select a number from the chosen distribution, repeating this process many times, the distribution of the leading digits of the resulting numbers will follow Benford's Law.

Hill's theorem [30] also helps us to understand the specific conditions under which we can expect deviations from Benford's Law:

- *Absence of Distribution Mixture:* When data comes from a single well-defined statistical distribution (rather than a mixture of several), they may not follow Benford's Law. For example, numbers generated from a uniform distribution between 1 and 100 will not follow Benford's Law.

- *Purely Additive Processes:* Data generated by strictly additive processes (as opposed to multiplicative ones) may deviate significantly from Benford's Law.
- *Sample Insufficiency:* Even when theoretical conditions are met, small samples may exhibit significant deviations due to natural sample variability.
- *Artificial Restrictions on Data:* When data is artificially constrained, such as in cases of imposed limits or sequentially assigned numbers, the distribution may deviate from Benford's Law.
- *Intentional Manipulation:* One of the most important applications of Hill's Theorem is in fraud detection. Deliberate data manipulation (such as falsifying financial records) often results in detectable deviations from Benford's Law because there is a tendency to distribute digits more uniformly when numbers are fabricated.
- *Noise and Measurement Errors:* Empirical data often contain measurement errors, censoring, or missing data, which can alter the distribution of the leading digits.

The Benford's Law is a statistical tendency, but not a strict rule that applies to all datasets. While many financial, scientific, and demographic data tend to follow this law approximately, factors such as external constraints, sample size, and underlying processes can lead to some variations.

## 2.2 Tools for Measuring Distribution Divergence in Benford's Law Applications

The evaluation of the degree of conformity between an observed distribution and the theoretical distribution foreseen by Benford's Law requires the use of robust quantitative tools that allow to objectively measure the deviation between both.

This section presents a set of divergence metrics widely used in studies that apply Benford's Law. Each of these tools is analyzed in terms of its theoretical foundation, its practical characteristics and its applicability in the context of non-conformity detection.

In addition, the role of statistical hypothesis tests in formal evaluation of conformity with Benford's Law is discussed, including the combination of $p$-values through the Fisher technique, allowing the integration of multiple metrics into a single statistical inference.

A thorough understanding of these tools is essential not only to correctly interpret the results of compliance tests, but also to compare the performance of the various numerical approaches in detecting anomalies in information systems. This section thus serves as a theoretical basis for the analysis and comparison of the metrics used in the model proposed in this work.

### 2.2.1 Evaluating Conformity to Benford's Law through Dissimilarity Measures

Dissimilarity measures are fundamental tools in data analysis, used to quantify the difference between observed objects and theoretical distributions. In the context of Benford's Law, these measures are essential for assessing the degree of conformity between an observed dataset and the theoretical distribution predicted by the law. Comparing distributions in this way is a widely used approach for identifying patterns, anomalies, or significant deviations.

Specifically, when applied to Benford's Law, dissimilarity measures make it possible to quantify how much a dataset diverges from the expected distribution, supporting the detection of potential data manipulations or anomalies.

In this project we used chi-square test, mean absolute deviation, Kolmogorov-Smirnov test, Euclidean distance, Hellinger distance and Kullback-Leibler divergence [44].

Pearson's chi-square statistic is one of the most widely used statistics in statistical tests to assess the independence or the goodness of fit of observed frequency distributions in relation to expected ones. It was introduced by Karl Pearson in 1900 and is widely applied in inferential statistics.

This measure is calculated as

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i\text{-}E_i)^2}{E_i}, \tag{2.11}$$

where, $O_i$ denotes the observed frequency and $E_i$ the expected frequency.

The higher the value of $\chi^2$, according to the chi-square Test of Goodness of Fit, the greater the discrepancy between observed values and expected values, indicating incompatibility between what was observed and the theoretical model analyzed. In the chi-square Independence Test, one of the best-known hypothesis tests in Statistics, it would mean that the variables are related, as the observed frequencies are far from what would be expected considering independence between the variables.

The mean absolute deviation, MAD, is a simple extension of the absolute variation. It adds up the absolute variations and divides the result by the number of records. The mean absolute deviation is a statistical error that calculates the average distance between observed proportions (relative frequencies) and proportion (probabilities) defined by Benford's Law. It is calculated as follows

$$\text{MAD} = \frac{1}{9} \sum_{i=1}^{9} \left| f_i - f_i^{(\text{BL})} \right|, \tag{2.12}$$

where $f_i$ is the proportion of observations where $d_1 = i$ and $f_i^{(\text{BL})}$ is the expected proportion of observation with $d_1 = i$ under Benford's Law, i.e., $f_i^{BL} = \log_{10}\left(1 + \frac{1}{i}\right)$.

The MAD measures the average of the distances between the proportion of observations and the expected proportion (under Benford's Law) for each possible digit. Higher values indicate greater distance between the observed observations and Ben-

ford's Law.

The Kolmogorov-Smirnov (KS) distance is a statistical measure used to compare two probability distributions. It is widely used to test the hypothesis that two samples come from the same distribution or to compare an empirical sample with a theoretical distribution. The KS distance can be defined by

$$D = \sup_x |F_1(x) - F_2(x)|, \tag{2.13}$$

where $\sup_x$ represents the supremum (i.e. the largest value) of the difference between the two cumulative distributions (CDFs) at all points $x$; $F_1(x)$ and $F_2(x)$ are the CDFs of the distributions being compared. In our application, one will correspond to the CDF of Benford's Law and the other to the empirical CDF of the data under analysis.

The Euclidean distance is one of the most common ways of measuring the proximity between two points in a multidimensional space. It is based on the Pythagorean theorem, which defines the distance between two points as the length of the straight line that connects them. Hence, it is calculated using the formula

$$d(P, Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{2.14}$$

for two points $P = (x_1, y_1)$ e $Q = (x_2, y_2)$ on the Cartesian plane (2D). In our investigation, since $d_1 \in \{1, 2, \cdots, 9\}$, the applied formula to compute the Euclidean distance between observed data (OD) and Benford's Law (BL) is

$$d(\text{OD}, \text{BL}) = \sqrt{\sum_{i=1}^{9} \left(f_i - f_i^{(\text{BL})}\right)^2}. \tag{2.15}$$

Again, the larger the value, the further away the distribution of the observed data is from the Benford's Law.

The Hellinger distance is a metric used to measure the dissimilarity between two probability distributions. It is often used to compare histograms, statistical models and discrete or continuous probability distributions.

For two discrete probability distributions $P = (p_1, p_2, \cdots, p_n)$ and $Q = (q_1, q_2, \cdots, q_n)$, the Hellinger distance is defined as

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}, \tag{2.16}$$

which is related to the Euclidean distance between the square root vectors $\sqrt{P}$ and $\sqrt{Q}$, where $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \cdots, \sqrt{p_n})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \cdots, \sqrt{q_n})$. Hence, for the $D_1$

assessment, the applied formula was

$$H\left(\mathrm{OD}, \mathrm{BL}\right) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{9} \left(\sqrt{f_i} - \sqrt{f_i^{(\mathrm{BL})}}\right)^2}. \tag{2.17}$$

The Kullback-Leibler divergence (KL) is another statistical measure that quantifies the difference between two probability distributions. It is asymmetric and indicates how much a distribution $Q$ deviates from a reference distribution $P$. It measures the loss of information when using $Q$ to approximate $P$.

For the two discrete probability distributions $P$ and $Q$, where $P = \left(p_1, p_2, \cdots, p_n\right)$ and $Q = \left(q_1, q_2, \cdots, q_n\right)$, where $p_i$ and $q_i$, for $i = 1, \cdots, n$, represent the probability of assuming the value $i$ respectively in the distribution $P$ and $Q$, the KL divergence is given by

$$D_{KL}\left(P \| Q\right) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right). \tag{2.18}$$

Thus, the formula applied in the investigation of $D_1$ is given by

$$D_{KL}\left(\mathrm{BL} \| \mathrm{OD}\right) = \sum_{i=1}^{9} f_i^{(\mathrm{BL})} \log\left(\frac{f_i^{(\mathrm{BL})}}{f_i}\right). \tag{2.19}$$

### 2.2.2 The Role of Hypothesis Testing in Detecting Deviations from Benford's Distribution

After quantifying how far a given set of data deviates from the expected distribution, we need to define a criterion (or cut-off point) to define the limit between what will be considered close to or far from the expected behavior.

To establish this criterion, it is essential to use statistical methods that allow an objective decision. In this context, hypothesis tests appear as an essential tool to evaluate whether the observed results are statistically significant or can be attributed to chance.

Hypothesis tests is a statistical methodology to make decisions about one or more characteristics of the population, based on the information obtained from the sample. In general, hypothesis tests evidence for or against the questions raised by the researcher, that is, it is a statistical procedure that allows to make a decision, reject or not the null hypothesis ($\mathrm{H}_0$), between two hypotheses (null hypothesis $\mathrm{H}_0$ and alternative hypothesis $\mathrm{H}_1$), using the observed data of a given experiment. They are used to determine which results of a scientific study may lead to the rejection of the null hypothesis, $\mathrm{H}_0$, at a pre-established level of significance $\alpha$. The study of probability theory and the determination of the correct statistical test are fundamental for the consistency of a hypothesis test. If the test hypotheses are not assumed correctly, the result will be incorrect and the information will be inconsistent with the scientific study question.

Hypotheses are statements or assumptions always formulated about the population

and that can be tested based on observed data (from samples). $H_0$ is usually a simple hypothesis, where only one value is specified for the parameter/distribution under study, whereas $H_1$ is usually a compound hypothesis, i.e., it is specified more than one value for the parameter/distribution.

As we are trying to validate a statement made for the population from sample data, the decision is subject to errors, see Table 2.2. The type I error is about rejecting $H_0$ when $H_0$ is true; while in type II error the hypothesis $H_0$ is not rejected, but $H_0$ is false [38, 39, 45].

| Decision | Situation | |
|---|---|---|
| | $H_0$ is true | $H_0$ is false |
| Reject $H_0$ | Error of type I | Right decision |
| Don't reject $H_0$ | Right decision | Error of type II |

**Table 2.2:** *Decisions and types of error in hypothesis tests*

Therefore, being the error of type I, the error made when rejecting the null hypothesis ($H_0$) in the cases where that hypothesis is true, its probability is called the level of significance $\alpha$ and is defined by

$$\alpha = \mathrm{P}\left(\text{reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ true}\right). \tag{2.20}$$

Similarly, the probability of making a Type II error is given by

$$\beta = \mathrm{P}\left(\text{Don't reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ false}\right). \tag{2.21}$$

The decision maker, by setting a significance level $\alpha$, is establishing the maximum acceptable probability of rejecting the null hypothesis when it is true, that is, of committing a type I error.

In the decision-making process, it would be ideal to minimize the probability of both types of error. However, for a fixed sample size, decreasing the probability of an type I error in a hypothesis test results in a higher probability of a type II error, and the opposite also occurs, a well known trade-off. In addition, to calculate the probability $\alpha$, it is assumed that $H_0$ is true, which is specific information (the law is known as the null hypothesis has only one possibility), while to calculate the probability $\beta$ of error type II, it is assumed that the alternative hypothesis is true, which usually includes several laws of probability or values for the parameters. Therefore, for each distribution or parameter value (different from the one proposed in the null hypothesis) the probability of a type II error is different. Thus, it is impossible, or at least much more complex, to control the type II error probability.

Under the assumption that the null hypothesis $H_0$ is true, the $p$-value represents the probability of obtaining a result equal to or more extreme than that observed by the statistical test. It is important to note that the $p$-value does not indicate the probability that the null hypothesis is true or false, but rather if there is evidence to question $H_0$.

Thus, low $p$-values indicate incompatibility between the observed data and the null hypothesis in $H_0$ and, as such, the decision to test the hypotheses is made through:

- If $p$-value $< \alpha$: there is statistical evidence to reject $H_0$ at a level of significance $\alpha$;
- If $p$-value $\geq \alpha$: there is no statistical evidence to reject $H_0$ at a level of significance $\alpha$.

To calculate the $p$-value, it is essential to know three elements: the null hypothesis in analysis, the probabilistic model associated with the statistical test and a way of ordering data according to any measure that assesses the dissimilarity between the observed data and the distribution associated with the null hypothesis, allowing to identify those that exceed a certain threshold under $H_0$. This ordering is usually established on the basis of a statistic test.

In this project we also used the Fisher method [27] to combine the $p$-value obtained with the calculated distances. It is a statistical technique used to aggregate evidence from multiple independent statistical tests (see [9, 10] for more information about combining $p$-values). It relies on chi-square statistics to assess whether there is a significant effect considering multiple sources of evidence.

If we have $k$ independent statistical tests, each with a $p$-value, $p_i$, the Fisher method calculates a combined statistic as follows

$$\chi^2 = -2 \sum_{i=1}^{k} \ln \left( p_i \right) . \tag{2.22}$$

If the observed value $\chi^2$ is sufficiently large, we reject the null hypothesis, indicating that at least one of the tests showed a significant effect, otherwise there is insufficient evidence to reject the null hypothesis.

To apply this method it is necessary that the $p$-value are derived from independent tests and the validity of the final conclusion depends on the quality of the individual tests.

### 2.2.3 Confusion Matrix-based Metrics and Epidemiology Metrics

A confusion matrix is a table that summarizes the performance of a classification model by comparing the predictions made by the model with actual results. In the case of a binary classification, such as distinguishing between normal (negative) and anomalous (positive) data, the matrix is organized into four categories:

- TP means true positive cases which represent correctly identified positive cases (anomalous data),
- TN means true negative cases which are the correctly classified negative cases (normal data),
- FP means false positive cases which correspond to the incorrectly classified negative cases as positive (false alarms – normal data classified as anomalous data),

- FN means false negative cases which are the positive cases that the model failed to detect (anomalous data classified as normal).

| | | **Predicted Label** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Actual Label** | **Positive** | TP | FN |
| | **Negative** | FP | TN |

**Table 2.3:** *Confusion Matrix for binary classification*

The confusion matrix (Table 2.3) is an example of a binary classification but it can also be applied to multiclass classifications.

The simplest and most intuitive metric that can be taken from a confusion matrix is accuracy. This metric gives us the proportion of correctly ranked cases among the total number of cases and, therefore, is given by

$$\text{Accuracy} = \frac{\textbf{cases correctly classified}}{\text{total of cases}}. \tag{2.23}$$

However, this metric should not be used in unbalanced datasets because it can give a false sense of high performance even if the model is completely ignoring the minority class. Taking as an example a dataset with 95% negatives and 5% positives, if the model classifies everything as negative, it hits 95% of the time, that is, has 95% accuracy, but in fact does not detect any positive. We have high accuracy in a useless model.

Hence, important metrics such as preditive value (precision), sensitivity (or *recall* of the positive class), F1-*score* and specificity (or *recall* of the negative class) are fundamental in problems with unbalanced classes where accuracy can be misleading. These metrics are used to evaluate the performance of classification models, especially in binary classification contexts (ex: normal vs. anomalous), but also apply to multiclass classifications.

Precision is the proportion of correctly classified cases among cases classified in the class, in multiclass classifications is given by

$$\text{Precision} = \frac{\textbf{cases correctly classified in the intended class}}{\text{total of cases classified in the intended class}}. \tag{2.24}$$

In a binary classification, precision (of the positive class) is the ratio of predicted positive cases that are actually positive and is given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2.25}$$

In a multiclass classification, *recall* represents the ratio of correctly classified cases in the actual class, given by

$$\text{Recall} = \frac{\textbf{cases correctly classified in the actual class}}{\text{total of cases in the actual class}}. \tag{2.26}$$

In a binary classification we can look at the *recall* (of the positive class) as the ratio of positive cases correctly classified as positive, wich is given by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.27}$$

The precision corresponds to the positive predicted value (PPV) in epidemiology. Moreover, the negative predicted value (NPV) would corresponds to the precision of the negative class and is defined by

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}, \tag{2.28}$$

which gives the proportion of correct classification between the negative classifications.

F1-*score* (of the positive class) is the harmonic mean between precision and *recall*, given by

$$\text{F1-}score = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}. \tag{2.29}$$

Two important epidemiology metrics are the sensitivity and specificity. The sensitivity is the same as *recall* (of the positive class), and the specificity measures the ability of the model to correctly identify negatives (true negatives) and corresponds to the *recall* of the negative class. Hence, specificity is given by

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}. \tag{2.30}$$

Another fundamental measure for analyzing the performance of a classification model is the ROC (Receiver Operating Characteristic) curve which is a visual tool to asses the binary classification models. The ROC curve graphically represents the relationship between sensitivity, or *recall* or True Positive Rate (TPR), and False Positive Rate (FPR), where FPR = 1 − Specificity. It shows how the model behaves as we vary the classification threshold. Thus, the ROC curve allows you to visualize the evolution of sensitivity and specificity when the cut-off point passes through all possible values, from the point at which all individuals are classified as positive to the other extreme at which all individuals are classified as negative [47]. It is especially useful when we want to compare models or analyze the compromise between detecting positive cases (high sensitivity) and avoiding false alarms (high specificity).

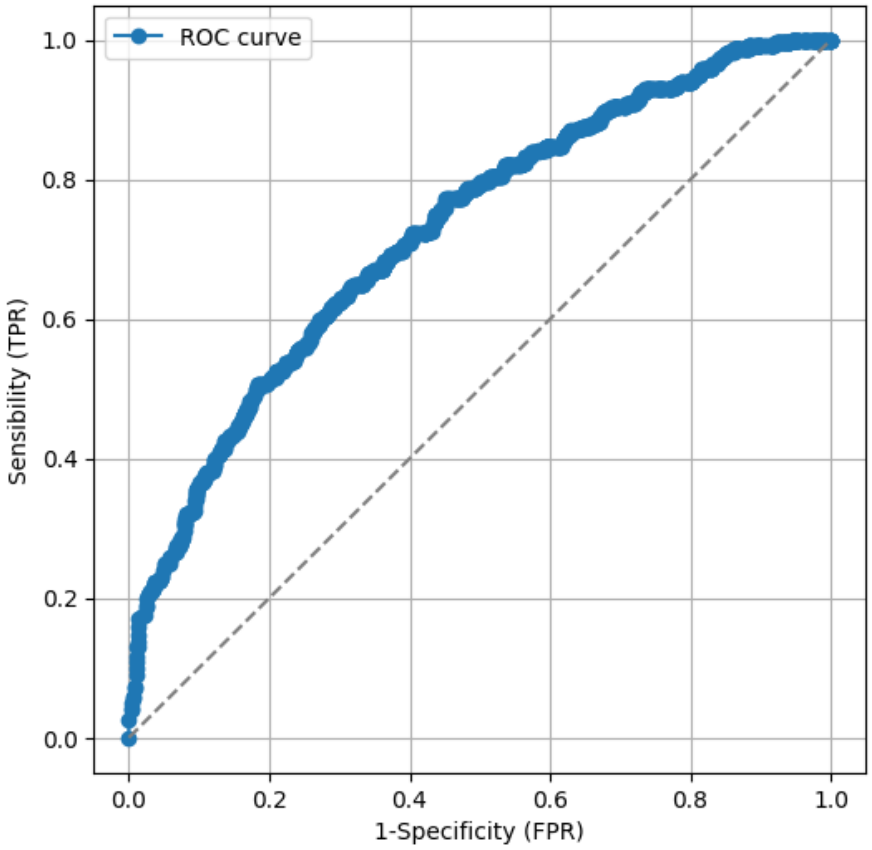


**Figure 2.3:** *Example of a ROC curve graph*

To find the cut-off point that best balances sensitivity and specificity, we can consider the point with the smallest distance to the point (0,1) of of the ROC curve represented in Figure 2.3. The point $(0, 1)$ represents the perfect classification point, with accuracy, sensitivity and specificity equal to 1 (absence of classification errors), so we are choosing the point closest to the (usually unattainable) optimal point.

Another method to find the cut-off point that best balances sensitivity and FPR is the Youden's statistic, J, which is defined as

$$\mathrm{J} = \text{Sensibility} + \text{Specificity} - 1 = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} - 1.$$

The best cut-off point is the highest value of J [47]. Note that the index J corresponds to adding the two recalls (the one associated with the positive class and the one associated with the negative class) minus 1.

Moreover, since the ROC curve represents sensitivity (recall) across all possible cut-off values, the area under the ROC curve (AUC) can be interpreted as the average sensitivity over the full range of specificities. An AUC of 0.5 indicates no discriminative ability, equivalent to random classification, while an AUC of 1.0 corresponds to perfect classification without error. The AUC is one of the most widely used metrics for evaluating the accuracy of binary classification methods.

In short, metrics based on the confusion matrix provide a comprehensive view of

the performance of a classification model, allowing to evaluate both the ability to correctly identify positive cases and avoid false alarms. They facilitate the comparison of the performance of statistical models, such as this project, with the performance of machine learning models. The balance between sensitivity and specificity, metrics widely used in epidemiology, as in the ROC curve, is particularly important because it represents the model's ability to correctly identify positive cases without compromising the correct identification of negative ones. In critical contexts, such as medical diagnosis, this balance is essential to minimize both false negatives (undetected dangerous cases) and false positives (unjustified alarms), ensuring more reliable and responsible decisions.

## 2.3 Anomaly Detection

Anomaly detection, often used interchangeably with outlier detection, is the process of identifying data points, observations, or patterns that significantly deviate from the expected behavior within a dataset. These anomalies may indicate critical incidents such as fraud, security breaches, or operational faults, but can also result from benign causes such as data entry errors or rare but legitimate events.

Anomaly detection can be broadly classified into:

- Supervised methods, where labelled examples of normal and anomalous instances are available, and are used to train the model to distinguish examples between the two classes.
- Unsupervised methods, which detect deviations based only on the structure of the data, without prior labels.
- Semi-supervised methods leverage both labeled and unlabeled data, using the labeled data to guide learning while exploiting patterns in the unlabeled data to improve generalization.

Due to the rarity and diversity of anomalous instances, often referred to as the class imbalance problem, anomaly detection presents significant challenges in accuracy.

Traditionally, anomaly detection have relied on:

- Machine learning models such as artificial neural networks (ANNs), support vector machines (SVMs), and decision trees (DT).
- Statistical methods including distributional analysis, outlier detection tests, and distance-based techniques.
- Rule-based systems, where domain-specific rules flag suspicious activity.

In this project, a purely statistical approach is proposed, leveraging Benford's Law and dissimilarity measures to quantify deviations from expected distributions, providing a computationally efficient and interpretable alternative to complex machine learning models.

### 2.3.1 Traditional Statistical Approaches

Traditional statistical approaches to anomaly detection include a set of mathematical methods based on statistical theories established decades ago, using rigorous principles of probability, inference and analysis of distributions. Unlike machine learning-based approaches, which often operate as black boxes by automatically learning complex patterns from large volumes of data, traditional statistical methods offer greater transparency and interpretability, based on explicit assumptions and formal tests with clearly defined statistical significance.

A significant advantage of these traditional techniques is their reduced consumption of computational resources, making them particularly suitable for real-time implementations and in systems with hardware limitations, enabling fast decision-making with low latency in critical applications. While machine learning techniques tend to be excellent in capturing complex non-linear relationships, traditional statistical approaches provide a solid theoretical foundation, allow intuitive understanding of results, facilitate the explanation of decisions, and often do not require computational power for effective implementation, being valuable in contexts where operational efficiency, interpretability and rigorous statistical justification are as important as predictive accuracy.

It is in this type of approach that the Benford's Law, the measures of dissimilarity and hypothesis tests play a fundamental role in anomaly detection, establishing the groundwork for more sophisticated statistical analysis [37].

Among the traditional statistical approaches, correlation analysis with the Pearson and Spearman coefficients identifies abnormally strong or weak relationships between variables and may indicate data manipulation [24, 23]. There is also the linear regression that models expected relationships between variables and allows to identify very large residues (differences between observed and predicted values), which can indicate anomalies [51].

Another relevant approach is Statistical Process Control (SPC), which employs control charts to monitor variables over time by establishing upper and lower control limits. Data points that fall outside these limits or exhibit non-random patterns may indicate process anomalies. [38].

Traditional outliers analysis is also an established statistical approach, and one of the most used methods is the standard score or $z$-score which calculates how many standard deviations a value is far from the average, as denoted in Equation 2.31

$$z = \frac{x - \mu}{\sigma}, \tag{2.31}$$

where $\mu$ denotes the mean and $\sigma$ the standard deviation of $x$. Values with $|z| > 3$ (or other defined limit) are considered potential outliers because they are significantly far from the average of the data in terms of standard deviations. This criterion is based on the assumption that the data follows a normal distribution, in which the values are distributed symmetrically around the average. In a normal distribution, about 99.73%

of the values are in the range between $-3$ and $+3$ standard deviations, so values with |z| greater than 3 are statistically very rare. This rarity suggests that such observations may not belong to the same underlying pattern as the remaining dataset. In contexts such as the detection of anomalies, these unusual deviations may reflect atypical or suspicious behaviors, justifying their identification as outliers.

The quartile-based method, based on the InterQuartile Range (IQR), which is defined by IQR = $q_3 - q_1$ where $q_i$ represents the $i$-th quartile results in outliers below $q_1 - 1.5 \times$ IQR or above $q_3 + 1.5 \times$ IQR. This method is a more robust approach because it does not depend on the normality of the data. Therefore, the quartile method, based on the interquartile interval, is often considered more robust than the $z$-score method, especially when the data does not follow a normal distribution. While the $z$-score assumes an approximately normal distribution to identify outliers based on distance in standard deviations from the average, the IQR method uses the quartiles of the data distribution itself, that is, the 1st quartile ($q_1$) and the 3rd quartile ($q_3$). Thus, the main advantage of this method is that it does not depend on assumptions about the form of the distribution of data, so it is suitable even when the distribution is asymmetric or contains extreme values. In addition, since it is based on position measures (quartiles) and not on scattering measures sensitive to outliers (such as mean and standard deviation), it is less influenced by extreme values, which makes it particularly useful in contexts with heterogeneous or biased data.

Another traditional statistical approach is the time series analysis, where we fit, for example, the Seasonal Autoregressive Integrated Moving Average models (SARIMA) that model time dependence structures to identify observations that do not follow established historical patterns. Another method is the decomposition of time series that consists in separating a series into trend, seasonality and residual components, allowing to identify anomalies in the residual component [56].

These traditional approaches have the advantage of being mathematically well established, relatively simple to implement and interpret, offering transparent and auditable results. They are particularly valuable in regulatory contexts where ease of interpretation is critical and continue to be the basis for many modern anomaly detection systems, often used in conjunction with more advanced machine learning techniques.

### 2.3.2 Machine Learning Approaches

The detection of anomalies using classification-based machine learning approaches relies fundamentally on distinguishing between normal and anomalous behavior. This distinction is made using labelled historical data as training examples (supervised learning methods). These approaches are especially useful in contexts such as financial systems, insurance, telecommunications or accounting audit, where data is available and past anomaly labelling allows predictive models to be trained [1, 2, 15, 46].

The most common supervised classification algorithms include logistic regression and decision trees, which are simple and interpretable, and random forests, which

capture non-linear relationships between variables and are quite robust to noisy data.

One of the major challenges in this area is class imbalance, since anomalies are much rarer than legitimate behaviours. To mitigate this problem, techniques such as the oversampling of the minority class, for example with the Synthetic Minority Oversampling TEchnique algorithm, SMOTE, or the undersampling of the majority class are used. Alternatively, many algorithms allow to incorporate different penalties for errors in each class, making them more sensitive to anomaly detection, i.e. with high penalties for false negative results. In this type of problems, the overall accuracy is often misleading, so we use metrics such as precision, *recall*, F1-*score*, ROC curve and particularly AUC, the area under the ROC curve, which is more informative in strongly unbalanced scenarios.

For the models to perform well, a careful pre-processing process and feature engineering is essential. This includes the normalization of data, proper coding of categorical variables and, above all, the construction of derived variables with predictive value, for example, time since the last transaction, changes in the amount, frequency of operations in a given period, among others. These variables often carry more signal than the raw data.

When there are not many examples of labelled data, semi-supervised or even unsupervised approaches can be applied. Here, the focus is on anomaly detection, patterns that deviate significantly from the norm. Models such as SVM, isolation forest or auto-encoders trained to reconstruct normal patterns are used to identify instances that deviate from the expected data.

Another critical aspect is the adaptability of models over time. Anomaly is by nature dynamic, the context and pattern of the data changes over time. Therefore, it is essential that the models are updated regularly and incorporate concept drift detection mechanisms, i.e., changes in behavior patterns over time that may affect the validity of the model.

Finally, in such a sensitive field as anomaly detection, the explainability of models is essential. Techniques such as SHapley Additive exPlanations, SHAP, and Local Interpretable Model-agnostic Explanations, LIME, help to interpret the decisions of the algorithms, explaining in a local or global way which variables are most relevant for a given forecast. This is especially important when the results are used to justify legal actions, account freezes or internal audits.

## 2.4 State of the Art in Anomaly Detection Using Benford's Law

The application of Benford's Law to anomalies detection has been an active field of research for several decades. One of the pioneers in this area was Mark Nigrini, who, in the 1990s, demonstrated how the distribution of digits in accounting data could reveal signs of manipulation and anomalies [41]. His work released the foundation for using digit analysis as a forensic tool in auditing and financial investigations.

Following Nigrini's initial contributions, several researchers developed systematic

methodologies to apply Benford's Law more rigorously. For example, C. Durtschi, W. Hillison and C. Pacini [19] provided important guidelines for auditors on how to properly interpret deviations from Benford's expected distributions, cautioning against blind reliance without considering contextual factors.

A key theoretical advancement was made by Theodore P. Hill [30], who mathematically proved that the distribution of first digits tends toward Benford's Law when data is drawn from a random mixture of distributions. Hill's work provided a solid theoretical justification for the presence of Benford behaviour in real-world datasets.

Comprehensive summaries like those provided by Andreas Diekmann [17] have outlined not only classical applications of the Benford's Law in detecting anomalies in numerical data produced by scientific research, such as statistical coefficients, experimental results, and the reported measurements, but also highlighted the importance of extending Benford-based techniques beyond financial datasets, emphasizing their relevance in broader scientific and academic contexts.

In more recent years, research has evolved towards more sophisticated statistical and computational approaches. For instance, George Judge and Laura Schechter [32] applied Benford's Law to detect potential fraudulent behavior in survey data, showing the method's versatility beyond financial applications. Furthermore, R. M. Fewster [26] offered a thorough statistical explanation for Benford's Law, strengthening understanding of its underlying probabilistic principles and reinforcing its applicability in anomaly detection.

An innovative extension of Benford's Law applications has been the work of Pedro Fernandes and Mário Antunes [21], who proposed a method for detecting manipulated digital images using Benford's Law. Their study, demonstrated that image manipulations can disrupt the natural distribution of discrete cosine transform (DCT) coefficients, which normally follow Benford's Law. This research opened new avenues for applying Benford's Law beyond financial datasets, particularly into cybersecurity and digital forensics.

Another important development has been the integration of Benford's Law with machine learning techniques. Jose A. Alvarez-Jareño and Jose M. Pavia [1] presented a pioneering approach by combining Benford-based features with supervised learning models to detect money laundering activities. Their study, "Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish court case", provided a real-world application of this methodology, demonstrating its effectiveness in judicial contexts and reinforcing the practical importance of hybrid models that leverage both statistical laws and modern computational techniques [4].

Extending the application scope even further, Laleh Arshadi and Amir Hossein Jahangir [3] explored the behaviour of Internet traffic data in relation to Benford's Law. In their work, "Benford's Law Behaviour of Internet Traffic", they showed that many metrics from network traffic naturally conform to Benford's distribution, suggesting potential applications for anomaly and intrusion detection in network security environments. This highlights that Benford's Law is a powerful tool for monitoring and

securing information systems.

In conclusion, the applicability of Benford's Law is becoming increasingly widespread and remains a highly active area of research remains a highly active area of research [35], as evidenced by recent studies in diverse fields such as tax irregularities and financial fraud detection [12, 20, 48], data manipulation [50, 57, 36], AI-generated text identification [55], detection of non-authentic digital images [34], anomaly recognition in electricity consumption patterns [31], digital pathology [11], electroencephalogram analysis [53], as well as astrophysical [29], and seismic (accelerogram) analyses [52], among others.

## 2.5 Research Gaps and Opportunities

Despite substantial progress in understanding and applying Benford's Law for anomaly detection and analysis, several important research gaps remain, offering valuable opportunities for further investigation:

1. Validation with confirmed cases of anomalies: while many studies demonstrate that certain datasets deviate from Benford's Law, few validate these findings against datasets with confirmed anomalous cases. Validation using confirmed anomalous cases is essential to move Benford's Law from a theoretical anomaly-detection tool to a practically trusted method. This requires building or gaining access to appropriate datasets and systematically evaluating performance.
2. Practical benchmarking and standardization: there is a lack of universally accepted benchmarks or standardized procedures for applying Benford's Law in practice. More comparative studies, public datasets, and clear guidelines could help practitioners implement these methods more reliably and consistently.
3. Exploration of the theoretical limitations of the method: while Benford's Law has been widely applied in detecting anomalies, especially in accounting and financial fraud, its theoretical foundations and limitations are often not addressed deeply in applied studies. This presents a significant research gap.
4. The development of optimized dissimilarity measures: current approaches often rely on conventional statistical tests (e.g., chi-square, mean absolute deviation) that may not fully capture the nuanced patterns of anomalous activities.
5. Creation of adaptive thresholds: the creation of sophisticated adaptive threshold systems has the potential to transform Benford's Law from a relatively blunt instrument into a precision tool capable of identifying subtle anomalies while respecting the natural statistical variations present in legitimate financial data.
6. Integration with other analytical techniques: the strategic integration of Benford's Law with complementary analytical approaches has the potential to address the limitations inherent in any single method while leveraging their combined strengths, creating more robust and comprehensive anomaly detection systems.

7. Evaluation of robustness against purposeful manipulation: it shifts the focus from passive detection to active anticipation of evolving fraud strategies, potentially initiating an "arms race" between increasingly sophisticated detection and evasion techniques that ultimately strengthens fraud prevention capabilities.
8. Improvement of interpretability: enhancing interpretability is not merely about technical communication but fundamentally about transforming statistical indicators into actionable insights that support decision-making in audit, investigation, and compliance contexts. This research direction has significant potential to bridge the gap between advanced statistical detection capabilities and practical anomaly prevention applications.
9. Implementation in real-time detection: the transition from retrospective to real-time Benford analysis represents a paradigm shift that could transform anomaly detection from a post-facto investigation tool to a preventive control mechanism capable of identifying and interrupting anomalous activities before they fully materialize, significantly reducing organizational exposure to operational disruptions, financial inefficiencies, compliance risks, and other adverse impacts caused by diverse kind of anomalies.
10. Adaptation to complex data structures: most studies applying Benford's Law focus on relatively simple datasets, such as financial transactions or accounting figures. However, modern datasets often involve complex structures, for example, hierarchical, networked, or time-series data. There is a need for new methods that can adapt Benford analysis to these complex environments, preserving its diagnostic power without oversimplifying the data.
11. Robustness to noise and data imperfections: real-world data is rarely clean, it often contains missing entries, measurement errors, or mixed sources. Although some robustness improvements have been proposed, developing more resilient statistical tests that maintain good detection capabilities under imperfect conditions remains an open challenge.
12. Integration with advanced machine learning techniques: while there are preliminary efforts to combine Benford's Law with machine learning, systematic integration remains limited. There are opportunities to develop hybrid models where Benford-based features enhance the explainability and performance of deep learning or ensemble methods in anomaly detection.
13. Domain-specific applications: the use of Benford's Law has been heavily concentrated in financial auditing. However, domains such as healthcare, cybersecurity, energy consumption monitoring, and environmental data also present promising grounds for application. Tailoring Benford-based methods to the specific statistical characteristics of different fields could yield powerful new tools. This is especially relevant because naïvely applying Benford's Law across domains without adaptation often leads to misleading results.
14. Theoretical extensions: most theoretical work focuses on the first digit or on simple extensions to the second and third digits. Deeper exploration of the joint

distribution of sequences of digits, and more general settings (e.g., non-decimal bases, generalized digit laws), is still relatively rare and could significantly expand the applicability of Benford-type analyses.

In short, despite the wide application of Benford's Law, significant gaps persist that limit its effectiveness and credibility. Among these, the first three issues that this project intends to address directly stand out: the need for practical benchmarking and standardization, the formulation of clear guidelines for consistent application and in the in-depth exploration of the theoretical limitations of the method.

In this project, the critical gap related to the validation of Benford's Law through confirmed cases of anomalies was addressed in an innovative and methodologically rigorous way, representing a significant advance in the transition of this tool from a predominantly theoretical status to a practically reliable application.

The strategy of controlled generation of synthetic data that has developed constitutes an ingenious solution to the fundamental problem of scarcity of datasets with confirmed anomalies. Unlike most studies that limit themselves to identify statistical deviations without confirmation of the anomalous nature of the data, the data generator developed in this project allows to create scenarios where the presence, location and intensity of anomalies are known a priori. This approach eliminates the uncertainty inherent in real data, where deviations from Benford's Law may have legitimate origins unrelated to irregularities.

The parameterization of the ratio of manipulated lines and the ratio of manipulations per line allows a systematic and granular validation of the performance of methods based on Benford's Law. This ability to precisely control where and with what intensity the anomalies are introduced makes it possible to assess not only whether the method detects anomalies, but also how effectively it does so in different controlled scenarios. This is a fundamental contribution, because it allows to quantify performance metrics such as accuracy, precision, *recall*, F1-*score* as well as sensitivity and specificity in an objective and replicable way.

The integrated classification model represents an important evolution in validation, because it goes beyond the mere identification of statistical deviations to implement a classification system that can be evaluated through standard performance metrics. This approach allows a systematic evaluation of performance that responds directly to the need identified in the research gap. The comparison of model classifications with known labels of synthetic data provides concrete empirical evidence on the practical reliability of Benford's Law.

The diversity of experimental settings created through systematic manipulation of the four main parameters (number of columns $m$, number of rows $n$, ratio of anomalous rows $t_{\mathrm{B}}$ and ratio of anomalies per anomalous row $t_m$) generates a comprehensive spectrum of validation scenarios. This multidimensional approach allows to evaluate the robustness of methods in various conditions, identifying not only when they work, but also under which specific conditions they may fail. This information is crucial for

establishing practical guidelines on the reliable applicability of Benford's Law.

Although it has not been yet explored, the developed generator has the ability to simulate different types of anomalies, which is particularly relevant, allowing to evaluate whether the methods based on Benford's Law are effective in detecting specific anomalies that are intended to be identified in real applications. This targeted validation is essential to establish the practical reliability of the method in specific contexts.

This project also establishes an important methodological precedent by demonstrating how controlled data generation can be used for systematic validation. This approach can be replicated and adapted by other researchers, creating a standardized framework for the validation of anomaly detection methods based on Benford's Law.

The developed data pipeline that combines controlled generation of synthetic data with a classification model for anomaly detection directly addresses one of the main limitations of this research area, the absence of standardized methodologies and reference datasets, offering practical benchmarking and standardization.

The synthetic data generation component developed is a fundamental tool for the scientific and professional community. The ability to simulate in a controlled manner different standards of compliance and non-compliance with Benford's Law allows creating a standardized test environment, something that has been lacking in the literature. This parameterizable approach, which allows the manipulation of variables such as number of instances, attributes and specific types of irregularities, offers the necessary flexibility to test diverse and realistic scenarios.

The synthetic data generation component developed is a fundamental tool for the scientific and professional community. The ability to simulate in a controlled manner different standards of compliance and non-compliance with Benford's Law allows creating a standardized test environment, something that has been lacking in the literature. This parameterizable approach, which allows the manipulation of variables such as number of instances, attributes and specific types of irregularities, offers the necessary flexibility to test diverse and realistic scenarios.

The integrated classification model not only demonstrates the practical applicability of these developments, but also sets a methodological precedent that other researchers can follow, adapt and expand. This holistic approach, which ranges from controlled data generation to the implementation of detection systems and their performance evaluation, provides a framework that can serve as a reference for future work in the area.

The experimental process carried out in this project establishes a direct and fundamental link with the research gap on the theoretical limitations of Benford's Law, contributing substantially to its completion through a rigorous and systematic methodological approach.

The controlled manipulation of the parameters columns and rows allows to explore how the dimensionality of the data affects the theoretical applicability of Benford's Law. This analysis is crucial to understand the limitations related to the size and structure of datasets, issues that are often not addressed in applied studies. By systemati-

cally varying these parameters, it is possible to investigate the theoretical limits of law at different scales of data, contributing to a deeper understanding of when and how the law remains valid.

The parameter used for the proportion of manipulated lines is particularly relevant to explore the theoretical limitations related to the prevalence of anomalies. This variable allows investigating fundamental questions such as "what is the theoretical threshold from which the presence of irregularities begins to significantly affect the expected distribution?". How do different levels of data contamination impact the theoretical robustness of tests based on Benford's Law? These are deep theoretical questions that are rarely addressed systematically in the applied literature.

The parameterization of the ratio of manipulations per line offers a unique perspective on the theoretical limitations at the granular level. This parameter allows to explore how the localized intensity of irregularities affects the theoretical foundations of law. By controlling this variable, it is investigating whether there is a theoretical threshold beyond which the localized irregularities compromise the global applicability of Benford's Law, a theoretical issue of great relevance that remains largely unexplored.

The systematic combination of these four parameters creates a multidimensional experimental space that allows to map the theoretical boundaries of Benford's Law in an empirical way. This approach contributes significantly to filling the identified gap in the following ways:

- Empirically quantifying theoretical limitations that until now were only speculative or based on case observations. This project transforms abstract theoretical questions into concrete and replicable measurements.
- Establishing causal relationships between data characteristics and the theoretical robustness of law, providing fundamental insights into when and why Benford's Law may theoretically fail.
- Creating an experimental framework that allows other researchers to systematically explore other theoretical dimensions of the law, establishing a methodological precedent for the investigation of theoretical limitations.

Thus, this project deeply investigated the theoretical foundations of Benford's Law through controlled experimentation, filling a critical gap in the literature by providing empirical evidence on the limitations and conditions of applicability of the law.

Therefore, this project represents a multifaceted and transformative contribution to the field of application of Benford's Law, simultaneously filling three critical gaps in current research.

# 3
# Framework

The proposed structure implements a modular architecture designed to systematically evaluate anomaly detection methods based on Benford's Law. The system addresses a key challenge in anomaly detection research: the difficulty of objectively comparing different methods due to the lack of terrain veracity in real-world datasets. The structure consists of a synthetic data generator that produces datasets with known anomaly patterns, allowing controlled experimentation and objective evaluation of performance. This approach allows researchers to manipulate specific variables (types of anomalies, proportions, data dimensions) while keeping all other factors constant, facilitating rigorous comparative analysis.

The system implements six statistical methods for anomaly detection: chi-square test, mean absolute deviation (MAD), Kolmogorov-Smirnov test, Euclidean distance, Hellinger distance and Kullback-Leibler divergence. Each method captures different aspects of distribution divergence - from global goodness-of-fit to specific distance measures and theoretical information divergences. In addition, the Fisher combination method aggregates individual test results to potentially improve overall detection performance. The framework generates comprehensive performance evaluations, including precision, *recall*, F1-*scores*, ROC curves and sensitivity analyses in various experimental conditions. This systematic approach allows the identification of optimal operating conditions for each method and provides insights into its relative strengths and limitations. To enable such assessments under controlled conditions, it was necessary to design a data pipeline capable of generating synthetic datasets and supporting the classification of anomalies, ensuring a robust validation environment for the proposed methods.

This chapter describes the development process of the data pipeline, which integrates two central components of this work: the synthetic data generator and the classification model for anomaly detection. The construction of controlled synthetic data allowed to simulate different standards of compliance and non-compliance with Benford's law, allowing the creation of a classified dataset, essential for the validation of the BL based anomaly detection system. The generator is customizable, allowing the

manipulation of several variables such as number of instances, number of attributes and type of anomalies introduced.

In addition, an unsupervised classification model was developed based on statistical tests of conformity with the theoretical distribution of Benford's Law. The model evaluates each instance of the dataset individually, applying a diverse set of statistical metrics in order to identify significant deviations that may indicate anomalies. The pipeline also includes evaluating the performance of the model through classic metrics.

The overall pipeline, from data generation to results aggregation, constitutes the methodological basis of the present work, allowing testing, comparing and validating different scenarios.

## 3.1 Development of the Data Generation module

This project includes a data generator that generates a dataset with a quantity of features defined by parameter $m$ and a quantity of instances controlled by parameter $n$. The generated dataset contains a $t_B$ ratio of anomalous rows with a $t_m$ ratio of anomalous elements.

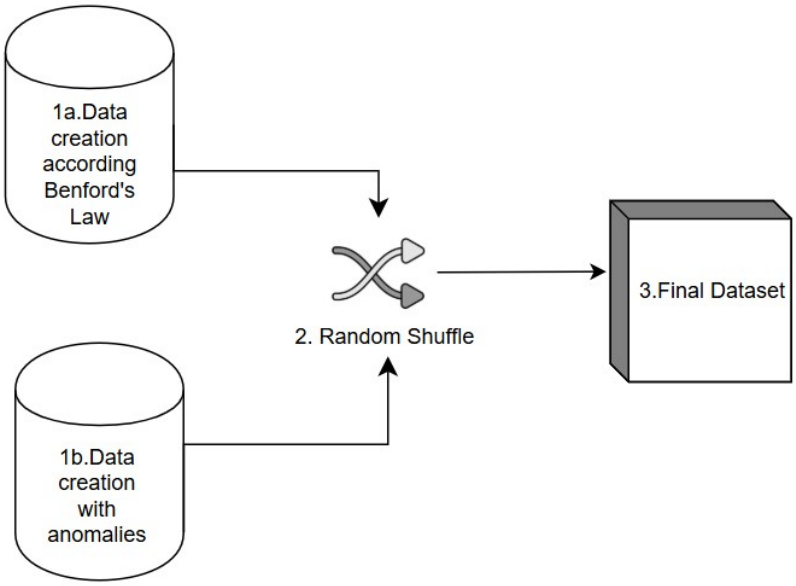


**Figure 3.1:** *Data generator Pipeline*

As we can see in Figure 3.1, the data generator begins by producing $(1 - t_B)\, n$ instances that follow Benford's Law (step 1a.Data creation according Benford's Law), as well $t_B n$ that not (step 1b.Data creation with anomalies). Each generated instance of numbers corresponds to a row in the dataset, with each number within the row occupying one column. The number of rows is controlled by the parameter $n$ and the column by the parameter $m$. The extra column shows the classification attribute. After that, the data is mixed to intersperse the numbers following Benford's Law with those

not following (step 2.Shuffle), thus creating the final dataset (step 3.Final Dataset).

*Step 1a.Data creation according to Benford's Law*

A data row is created with numbers generated from a uniform distribution between 0 and 1, see Figure 3.2, which are then transformed by raising 10 to these values, so that they follow the Benford's Law. That is, if $x \in \mathcal{U}(0,1)$ then $\log_{10}(x)$ follows a logarithmic distribution.

```
def gerar_num_benford (n):
    uniforme=np.random.uniform(low=0, high=1, size=n)
    bfl_numb=10**uniforme
    return bfl_numb
```

**Figure 3.2:** *Function in python to generate data according to Benford's Law*

Using $10^{\mathcal{U}(0,1)}$ allows the resulting numbers to be distributed logarithmically, which is a characteristic of BL. This method of generating random data according BL was studied by Arno Berger and Theodore P. Hill [8].

The method used in this project is detailed below in example 10 [8]:

*Example 10: Let $U$ be uniformly distributed in $[0, 1]$.*
*(1)(...)*
*(2) $X = 10^U$ is Benford, since $S(X) = X$, and $P(S(X) \leq t) = P(X \leq t) = P(10^U \leq t) = P(U \leq \log_{10} t) = \log_{10} t$ for all $t \in [1, 10)$. In fact, this construction provides an excellent way of generating random data that follows Benford's law on a digital computer: Use any standard program to generate $U$, and then raise 10 to that power.* [8]

After the generation of the row, or instance, according to the BL, it receives a label with the value 0, indicating that it does not contain anomalies.

Then, several other rows without anomalies are generated, always using the same method, by adding the 0 label to these rows and ensuring that the values are positive, until the ratio given by 1-$t_B$ is achieved (i.e., $(1 - t_B)\, n$ rows without anomalies).

*Step 1b. Data creation with anomalies*

After that, the generator creates rows that simulate instances with anomalies whose amount depends on the parameter $t_B$(ratio of anomalous rows). For these, it generates a combination of numbers based on Benford's Law, in proportion of 1-$t_m$ ($(1 - t_m)\, m$ numbers), and anomalous numbers according to the parameter $t_m$ (ratio of anomalies in each anomalous row, which corresponds to $t_m m$ numbers).

The anomalous row can originate in a uniform distribution, Gaussian noise, uniform noise, mixed noise (uniform + Gaussian) or they can be simple random outliers. Uniform distribution anomalies are obtained by a uniform continuous distribution between 0 and 100000. To create the anomalies based on Gaussian noise, the generator first generates data according to the Benfords Law, then generates the Gaussian noise

based on a standard distribution between 0 and the $intensity$ parameter and at the end adds up the conforming data, with the noise. Generator creates the anomalies based on Uniform noise in the same way but instead of a normal distribution it uses a uniform distribution between $-intensity$ and $+intensity$. To generate mixed noise, the generator sums Benfords Law data with Gaussian noise data and Uniform noise data. All of anamalous rows are labeled 1, indicating the presence of anomalies.

*Step 2.Shuffle*

All generated data is organized into a DataFrame, which is then shuffled randomly to ensure that the examples with and without anomalies are mixed together.

*Step 3.Final dataset*

Finally, the DataFrame obtained is converted into a NumPy array, making the data ready for use. Figure 3.3 shows an example of a generated dataset.

```
array([[6.16727549e+01, 5.19075421e+03, 2.36592793e+01, ...,
        4.44293928e+05, 6.67026738e+05, 0.00000000e+00],
       [1.91004563e+02, 5.41196970e+03, 1.64800044e+05, ...,
        4.17054451e+00, 1.38235303e+00, 1.00000000e+00],
       [2.31138226e+01, 1.32537244e+09, 1.12888602e+07, ...,
        9.76302447e+00, 4.33837899e+00, 1.00000000e+00],
       ...,
       [1.94467113e+05, 8.79702267e+08, 1.20516994e+00, ...,
        4.14937877e+00, 8.62349032e-01, 1.00000000e+00],
       [5.90301545e+01, 9.86029318e+08, 2.15014750e+03, ...,
        2.42349759e+05, 7.00042264e+04, 0.00000000e+00],
       [1.40464578e+03, 7.03929612e+01, 1.46285093e+05, ...,
        1.24339280e+05, 2.42749694e+08, 0.00000000e+00]])
```

**Figure 3.3:** *Example of a generated dataset*

The Algorithm 1 summarizes the data generation process previously described. The developed data generator is a robust and versatile benchmarking tool, capable of simulating near-reality scenarios through the controlled generation of synthetic datasets. This ability to reproduce features and patterns found in real data allows researchers to test and validate their algorithms under controlled conditions, ensuring more consistent and comparable results.

One of the main advantages of this tool is that it allows to explore different scenarios and experimental conditions in a systematic way, allowing more comprehensive comparative studies and the identification of limitations and strengths of the proposed methods. The flexibility in the parameterization of the generated data offers researchers the opportunity to test their algorithms under specific conditions, contributing to a more rigorous and informed evaluation.

In short, this tool represents a significant contribution to the scientific community, providing an efficient and reliable means for benchmark data generation, facilitating reproducible research and promoting the advancement of knowledge in the area.

**Algorithm 1:** Data Generator

**Input:** Number of columns $m$, number of rows $n$, proportion of anomalous rows $t_B$, proportion of anomalous elements per anomalous row $t_m$
**Output:** Final shuffled dataset `final_data`

**Initialize:**;
$qt\_BL \leftarrow n \times (1 - t_B)$;
$A\_count \leftarrow m \times t_m$;
$BL\_count \leftarrow m \times (1 - t_m)$;
`result` $\leftarrow$ one Benford row from $10^{U(0,1)}$ with label $0$ (no fraud);

**for** $i \leftarrow 1$ **to** $qt\_BL$ **do**
  Compute Benford numbers as $10^{(\text{uniform samples})}$;
  Take absolute values;
  Append label $0$ (no fraud);
  Append row to `result`;

**switch** *anomaly* **do**
  **case** *uniform* **do**
    **for** $i \leftarrow qt\_BL + 1$ **to** $n$ **do**
      Generate $t_m$ uniform random values in $[0, 10^5]$ (fraud);
      Concatenate Benford numbers with anomalies;
      Append label $1$ (anomaly) and add to `result`;
  **case** *gaussian_n* **do**
    **for** $i \leftarrow qt\_BL + 1$ **to** $n$ **do**
      $ben\_aux \leftarrow t_m$ Benford numbers from $10^{(\text{uniform samples})}$;
      Take absolute values;
      Generate $t_m$ normal random values;
      Sum $ben\_aux$ with normal values;
      Concatenate with Benford numbers;
      Append label $1$ and add to `result`;
  **case** *uniform_n* **do**
    **for** $i \leftarrow qt\_BL + 1$ **to** $n$ **do**
      $ben\_aux \leftarrow t_m$ Benford numbers from $10^{(\text{uniform samples})}$;
      Take absolute values;
      Generate $t_m$ uniform random values;
      Sum $ben\_aux$ with uniform values;
      Concatenate with Benford numbers;
      Append label $1$ and add to `result`;
  **case** *outliers* **do**
    **for** $i \leftarrow qt\_BL + 1$ **to** $n$ **do**
      Generate $t_m$ values in defined outlier band;
      Append label $1$ and add to `result`;

Shuffle `result` randomly (fixed seed);
Reset indices;
Convert to numpy array `final_data`;
**return** `final_data`.

## 3.2 Classification Model Design and Implementation

The developed classification model implements an unsupervised approach, focused on the detection of anomalies based on statistical discrepancy with respect to the Benford distribution, without the use of supervised training or conventional machine learning.

The model was implemented in Python, using the following libraries:

- `NumPy` for numerical operations and array manipulation,
- `SciPy` for statistical tests (chi-square),
- `Pandas` for data handling and reporting performance metrics,
- Custom functions for metrics such as MAD, Euclidean, Hellinger, and KL divergence.

The system is organized into modular components following a sequential pipeline. It involves the sequential application of different statistical tests and distance measurements, whose results are subsequently combined to generate predictions and performance metrics. Figure 3.4 shows the flow chart that synthesizes the pipeline.

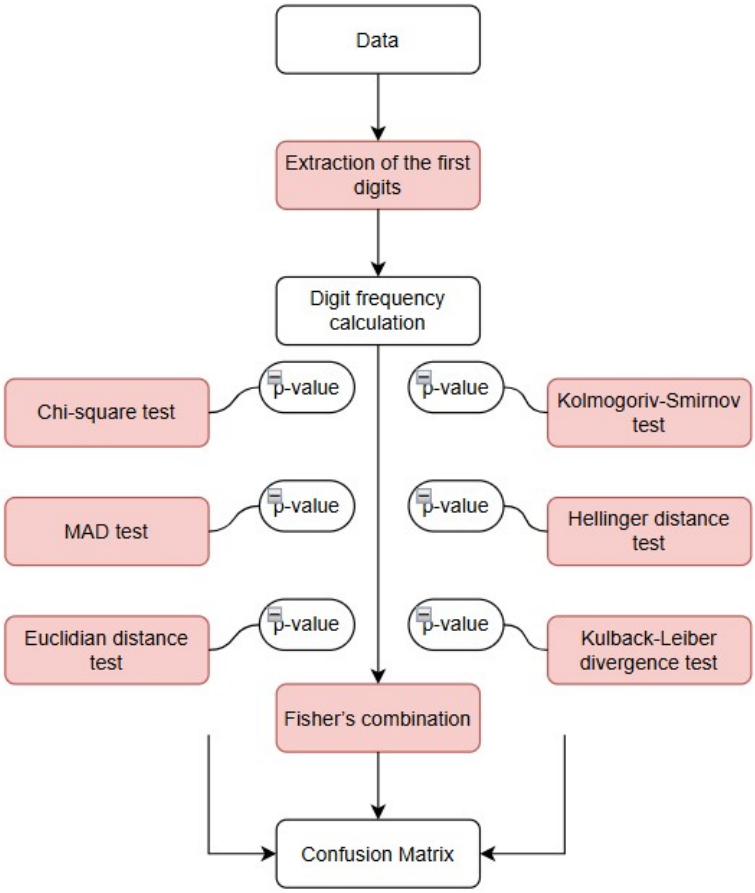


**Figure 3.4:** *Schematic description of the model architecture*

The system was designed to process each observation (row) independently, testing whether the empirical distribution of first digits conforms to the theoretical distribu-

tion defined by Benford's Law. The procedure begins with the extraction of the first digits from each numerical value of the analyzed row (Extraction of the first digits), followed by the calculation of the frequency of occurrence of each digit (Digit frequency calculation). This observed distribution is compared with the theoretical distribution of Benford's Law through six statistical and distance methods:

- Chi-square test (Pearson's chi-square),
- Mean absolute deviation (MAD),
- Kolmogorov-Smirnov (KS),
- Euclidian distance,
- Hellinger distance,
- Divergence of Kullback-Leibler.

Each metric captures different aspects of deviation, from cumulative distribution differences to point-wise distances and divergence measures. The diversity of metrics increases the detection capability under different manipulation patterns. After computing these metrics, their statistical significance is assessed through the corresponding $p$-values. Values lower than the significance level $\alpha$ indicate the rejection of the null hypothesis and, therefore, the classification of the case as manipulated (class 1).

So, for each test, the model:

- Generates the statistic and $p$-value,
- Applies a decision rule ($p$-value $< \alpha$ classified as anomalous),
- Records both the $p$-value and binary classification.

To avoid numerical problems, very low values are truncated to $10^{-15}$.

The $p$-values obtained in the six tests are combined by the Fisher's combination method, resulting in a single aggregated measure. For these tests a significance level, $\alpha$, previously configured is used to produce a final classification. Both the individual results of each test and the combined result are used to generate predictions that, together with real classes, feed the construction of the confusion matrix.

From this matrix, performance metrics such as precision, *recall* and F1-*score* are calculated, allowing to evaluate the effectiveness of the model. In addition, a variation of $\alpha$ (0 to 1, with step of 0.001) is generated to enable the construction of the ROC curves and analysis of the discriminatory capacity of each approach.

This architecture allows independent evaluation of each metric's performance, as well as the combined effect via Fisher's method.

The code was designed to ensure modularity to allow future extension with additional metrics or alternative combination methods.

The system processes each row independently in an iterative loop, storing predictions, $p$-values, and performance metrics in structured arrays and DataFrames for later analysis.

The model's predictions were compared to the known class labels using a confusion matrix for each metric and for Fisher's method. From the confusion matrix, the following performance indicators were computed:

- True Positives (TP): correctly identified anomalous rows,
- True Negatives (TN): correctly identified non-anomalous rows,
- False Positives (FP): normal rows incorrectly flagged as anomalous,
- False Negatives (FN): anomalous rows missed by the model.

From these, precision, *recall*, and F1-*score* were calculated to provide a balanced view of classification performance.

To complement the narrative description and improve the formal presentation of the proposed method, the Algorithm 2 outlines the process step-by-step. This structured representation facilitates replication and provides a clear overview of the workflow, from data input to the computation of performance metrics.

**Algorithm 2:** Classification Model Algorithm

**Input:** DataFrame `result`, significance level $\alpha$, expected distribution `expected_distribution`
**Output:** Per-test forecasts, combined forecast, performance metrics

**foreach** *row n in* `result` **do**
  Extract feature values (exclude actual label);
  Calculate first digits array `array_primeiros_digitos`;
  Calculate observed first-digit frequencies `frequencia_primeiros_digitos`;
  Apply statistical tests and compute $p$-values:;
    $\chi^2$ statistic;
    Mean absolute deviation (MAD);
    Kolmogorov–Smirnov statistic (KS);
    Euclidean distance;
    Hellinger distance;
    Kullback–Leibler divergence (KL);
  **foreach** *test* **do**
    Truncate $p$-values $< 10^{-15}$ (avoid numerical issues);
    Class $\leftarrow 1$ if $p$-value $< \alpha$; else $0$;
  Combine $p$-values via Fisher method;
  Assign combined class $1$ if $p$_fisher $< \alpha$, else $0$;

Build confusion matrix for each test and combined classification;
Compute drecision, *recall*, and F1-*score* for each;
Save metrics to Pandas DataFrame;
Export results to Excel file.

Importantly, the evaluation metrics (precision, *recall*, and F1-*score*) are standard in machine learning classification tasks. Their inclusion ensures that the model's performance is directly comparable to traditional machine learning classifiers, enabling benchmarking and positioning the statistical approach within the broader ML context.

All performance results were organized into a Pandas DataFrame which can be used for reporting and visualization.

## 3.3 Hypothesis Testing Framework

The present work implements a non-parametric hypothesis testing framework, which evaluates the compliance of the observed distributions of the first digits in relation to the theoretical distribution provided by Benford's Law. This approach aims to quantify deviations between the observed data and the expected pattern, using different divergence metrics.

The process is based on Monte Carlo simulation, in which a large number of samples are generated under the null distribution of each test statistic. For each simulation, a sequence of first digits is generated that follows the Benford distribution, with the same size of the observed sample. The corresponding divergence statistic between the simulated values under $H_0$ and the expected frequencies is then calculated. This procedure allows us to observe distances compatible with $H_0$, where the observed distances come from randomness, since we observe the distance between the observed values and the theoretical values in a large number of samples that are in accordance with Benford's law (a known piece of information since we simulated these samples). Then one function generalizes this process, allowing to apply the six different metrics already identified previously. Each metric is implemented in a separate function, ensuring modularity and flexibility. The hypothesis test follows the following steps:

- Define the null hypothesis ($H_0$): data follows the Benford distribution.
- Generate the null distribution: through repeated simulation of $m$ samples of digits following the distribution of Benford.
- Calculate the observed statistic: the divergence metric is calculated between the actual data and the theoretical distribution.
- Calculate the $p$-value or each value of the test statistic: the $p$-value corresponds to the ratio of simulated statistics equal or greater than the observed statistic. That is, it represents the probability of getting such an extreme deviation (or more) under $H_0$.

This process avoids parametric assumptions about the distribution of test statistics, ensuring greater robustness in real data contexts with possible deviations or heterocedasticity.

The choice of multiple metrics allows to capture different aspects of the divergence between distributions, from global differences (e.g., KS) to probabilistic distances (e.g., Hellinger, Kullback-Leibler), increasing the sensitivity of anomaly detection.

This framework thus provides a rigorous statistical basis to identify significant deviations from Benford's law, allowing the evaluation of different hypotheses with complementary metrics.

## 3.4 Experimental Procedures

A sensitivity analysis was conducted in order to evaluate the performance of different methods based on Benford's Law in detecting numerical anomalies. The experimental process consisted of controlled generation of synthetic data, allowing to analyze the sensitivity and robustness of statistical tests in different experimental settings, following the steps that have already been previously explained. The kind of anomaly used in this experiments is the uniform data. In each simulation was played with the parameters explained in Table 3.1.

| **Parameter** | **Symbol** | **Description** |
|---|---|---|
| Number of columns | $m$ | Dataset dimensionality |
| Number of rows | $n$ | Sample size |
| Anomalous row ratio | $t_B$ | Proportion of anomalous instances |
| Anomaly intensity | $t_m$ | Proportion of anomalous elements per row |
| Significance level | $\alpha$ | Statistical threshold |

**Table 3.1:** *System Parameters Configuration and Testing Ranges*

Each row of the dataset was then subjected to a statistical analysis, where the frequencies of the first digits were calculated and seven methods of anomaly detection were applied: the chi-square test, the mean absolute deviation, The Euclidean distance, the Kolmogorov-Smirnov test, the Hellinger distance, the Kullback-Leibler divergence and, to have a combined effect, the $p$-values combination by the Fisher method. For each method, the value of the test statistic and its $p$-value were recorded. The classification of each row as anomalous or normal was made by comparing the $p$-value with a level of significance $\alpha = 0.05$. A value lower than $\alpha$ led to the classification as anomalous. We varied only one parameter under study, keeping all the others fixed. For each value assumed by this parameter, the simulation is repeated, allowing to observe how this change influences the results.

The results of the classification were compared with the real labels of the data (indicating whether or not the row had manipulation), allowing to calculate the performance metrics of each method, namely the values of true positives, true negatives, false positives and false negatives. From these, additional metrics such as precision, *recall* and F1-*score* were derived. For each method, the ROC curve was also plotted through the variation of $\alpha$ in the interval $[0, 1]$, allowing to evaluate the compromise between true positive and false positive rate. From the ROC curve, two optimal cut-off points were determined, by the Youden criterion, corresponding to the maximum difference between TPR and FPR, and by the criterion of the smallest distance to the ideal point $(0, 1)$, without any misclassification problem.

The experimental process was repeated by systematically varying one parameter at a time, keeping the remaining constant, in order to study the impact of different factors on the performance of the methods. The variables tested included the ratio of manipu-

lations per row ($t_m$), the ratio of manipulated rows ($t_B$), the total number of rows ($n$) and the number of columns per row ($m$). Each simulation produced results recorded in Excel files, containing the performance metrics obtained for each experimental configuration.

This experimental design ensured comparability between the methods, since all were evaluated under the same conditions and using the same simulated data in each scenario. The use of synthetic data also allowed to ensure the previous knowledge of the real class of each set, ensuring a rigorous evaluation of the anomaly detection capacity of each tested method.

## 3.5 Computational Implementation

All computational procedures described in this work were implemented using the Python programming language, leveraging its flexibility and the availability of robust scientific libraries. Hence, the data generation, statistical testing, and performance evaluation were fully coded and executed in Python, ensuring reproducibility and allowing for flexible experimentation.

To generate synthetic datasets conforming to Benford's Law, a custom function was developed based on inverse transform sampling using the logarithmic distribution. This allowed the creation of numerical data spanning several orders of magnitude while preserving the expected digit distribution. Functions were implemented to extract the first digit of each number, compute their empirical frequencies, and compare these with the theoretical Benford distribution.

Multiple statistical distances and divergence measures were coded from scratch, including the mean absolute deviation, Kolmogorov-Smirnov, Euclidean distance, Hellinger distance, and Kullback-Leibler divergence. Each test statistic was calculated both for the observed data and for simulated datasets generated under the null hypothesis, enabling Monte Carlo estimation of $p$-values. A dedicated function was implemented to carry out the hypothesis testing process iteratively for each metric, returning the observed statistic and its corresponding $p$-value based on the empirical distribution from simulations.

The combination of $p$-values from different tests was performed using Fisher's method, implemented directly via the log-sum formula and chi-squared distribution. Additionally, custom handling of confusion matrices was developed to accommodate cases where certain classes might be missing from predictions, ensuring consistent extraction of true/false positives and negatives.

For the evaluation of detection performance, functions were written to compute confusion matrices at different decision thresholds, calculate false positive and true positive rates, and generate ROC curves. Visualizations of the ROC curve and confusion matrices were produced using Matplotlib and Seaborn, providing graphical insights into the model's classification performance.

Throughout the implementation, widely-used Python libraries such as `NumPy`, `SciPy`,

`Pandas`, `Scikit-Learn` and `Matplotlib` were employed to facilitate numerical operations, statistical functions, and data visualization. The code was structured into modular functions, enabling reuse and simplifying the integration of additional statistical measures or evaluation metrics in future work.

# 4
# Results and Analysis

The sensitivity analysis is a fundamental step in evaluating the robustness of statistical methods, allowing to understand how far the results obtained remain consistent against changes in structural parameters of the simulation.

In order to address this aspect, this chapter presents and discusses the results obtained with the system described in Chapter 3. The results were obtained with simulations and tests performed throughout the study. The way in which the simulations were carried out is described in the Section 3.4.

The main objective is to evaluate the performance of the developed model and the effectiveness of different divergence metrics in detecting irregularities in numerical data, based on Benford's Law, and understand their performance in different scenarios, varying the number of rows (cases), the number of columns (features), the ratio of cases with anomalies and the ratio of irregularities per anomalous cases.

In order to compare the classification performance of the different approaches, precision, *recall* and F1-*score* were calculated and the confusion matrix for each measurement combination and decision threshold was analyzed. This analysis allows not only to verify which methods are the most effective in identifying irregularities, but also to evaluate their behavior in terms of false positives and negatives.

Finally, the practical implications of the results are discussed, highlighting the advantages and limitations of each approach and proposing recommendations for their application in real contexts.

## 4.1 Sensitivity analysis in relation to the number of cases

This chapter aims to investigate the influence of sample size on the results obtained. To do this, we compare different scenarios, varying the number of rows in the datset, while the other parameters (number of columns, proportion of anomalous rows and anomalies per anomalous row) remain constant. This approach makes it possible to evaluate whether the growth in the number of cases leads to a systematic improvement in test performance, or if there are limits from which additional gains in robustness

become marginal.

### 4.1.1 Ratio of anomalous rows of 30% with significance level of 0.05

As the number of cases (rows) increases, different trends are observed among the different statistical methods used to detect deviations from Benford's Law.

To perform this simulation, the following parameters were established:

- $m$=1000 (quantity of columns),
- $n$, quantity of rows, varies between 100 and 9600 by steps of 500, i.e., $n = 100 + 500 \times k$ with $k = 0, 1, \cdots, 19$.
- $t_B$=0.3 (ratio of anomalous rows),
- $t_m$=0.25 (ratio of anomalies per anomalous row, that is, proportion of columns with anomalies in each row with anomalies),
- $\alpha = 0.05$ (significance level).

The $\chi^2$ test, widely used in studies on Benford's Law, showed a relatively good performance in this study. The Figure 4.1 shows a lower precision with the use of 100 rows. From 600 rows this metric remained between 0.885 and 0.908, indicating that the amount of false positives is very low. The *recall* was almost constant between 0.997 and 1. The F1-*score* showed the influence of precision for less than 600 rows and then remained between 0.939 and 0.952. This measure is stable from 600 rows, having an high precision and *recall*, so, it has a good balance. It suggest that it is good for datsets with more than 600 rows.

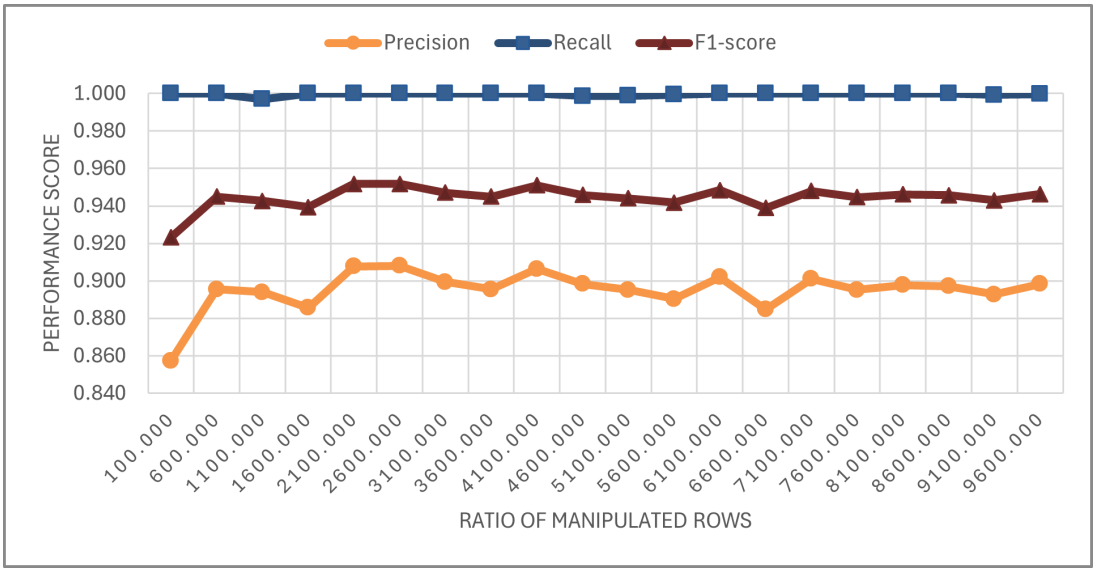


**Figure 4.1:** *Chi-square: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

The analysis to MAD in Figure 4.2 doesn't show us an evident drop in precision above 600 rows. Between 600 and 9600 rows it remains in the range $[0.888; 0.906]$. With this test the *recall* is not so stable but still remains between 0.988 and 1. Among all

the tests performed in this simulation, MAD is the one that presents a stable F1-*score* between 0.939 and 0.950.

Hence, it has F1-*score* consistent and stable. A slightly better precision than $\chi^2$ for 100 rows but with slightly lower *recall* in general. As is achieves great overall performance, the results suggests that it is a good candidate for robust use.

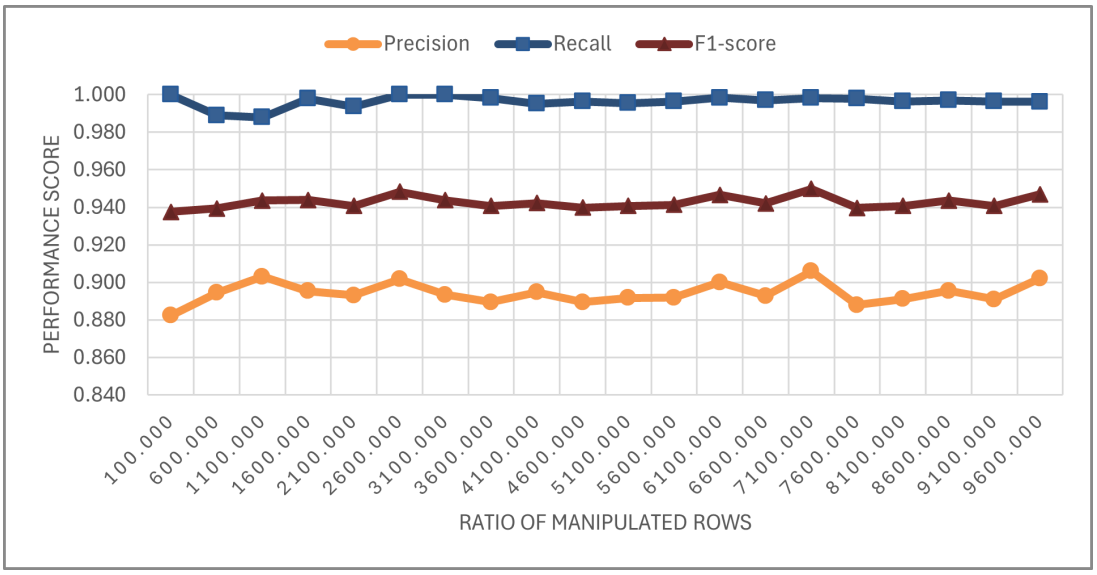


**Figure 4.2:** *MAD: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

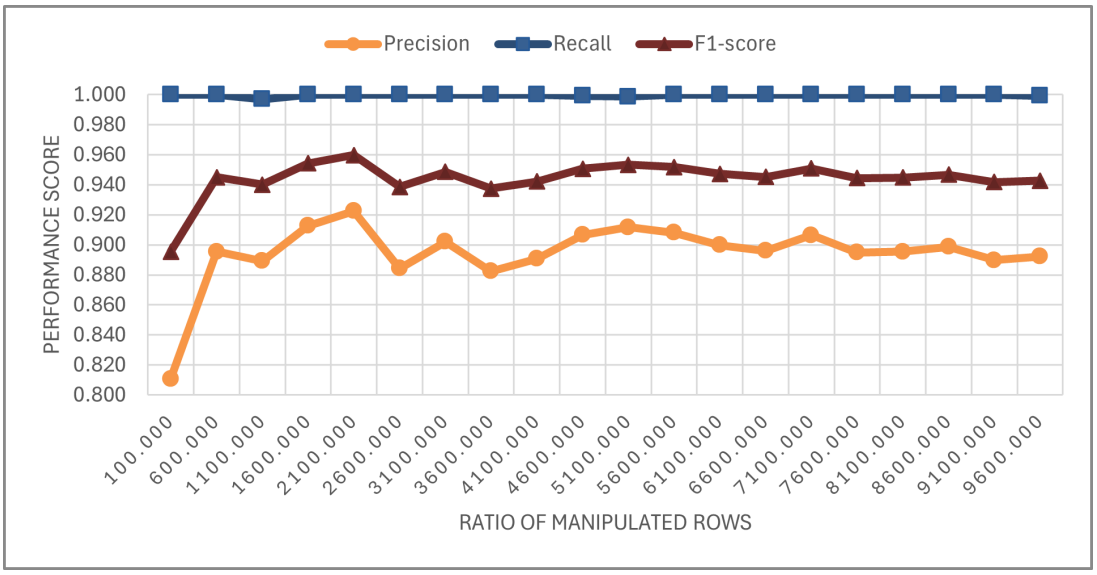


**Figure 4.3:** *KS: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

The Kolmogorov-Smirnov test, KS, is the method that reveals the greatest variability in the precision values with a large break below 600 rows, cf. Figure 4.3. For 100 rows the precision is 0.811 and for 600 rows it is already 0.895. The F1-*score* also shows a lot of instability due to the influence of precision. *Recall* is stable with the value of 1.

This test has high variability and is sensible to sample size. It can generate many false positives with less than 600 rows. It could be useful only with large volumes of data and when used carefully.

In the Euclidean Distance test the precision also has a large variation up to 3600 rows. From there it stabilizes the values between 0.888 and 0.901. The *recall* behavior is identical to previous tests (see Figure 4.4).

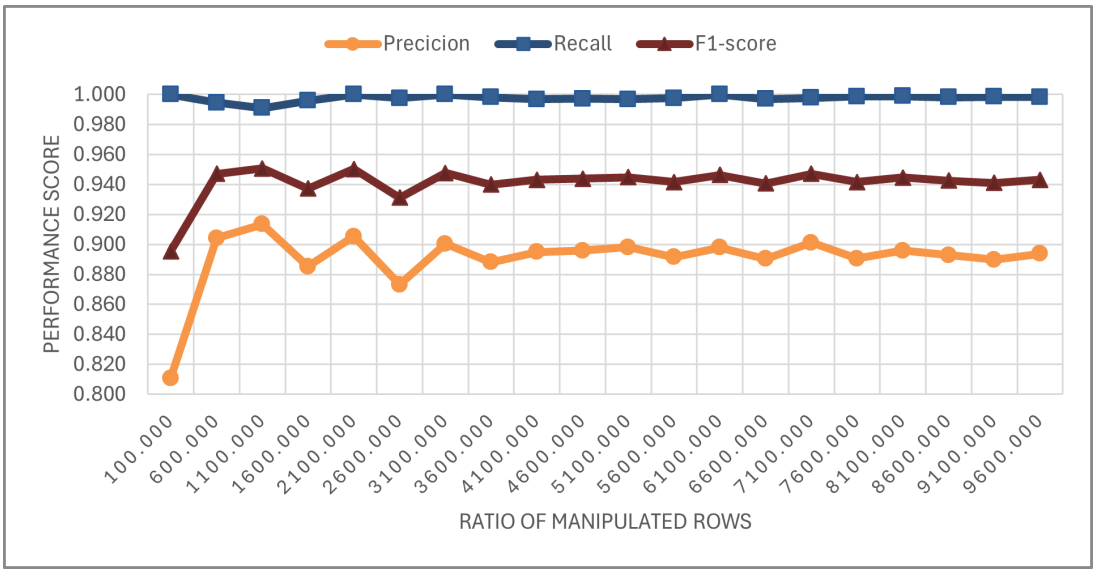


**Figure 4.4:** *Euclidean distance: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

As can be seen in the Figure 4.5, contrary to previous tests, in the Hellinger test the precision has its maximum value on the 100 rows. It has better precision with small samples, which is rare. This metric has a range between 0.878 and 0.909 along the analyzed sample dimensions. The *recall* is stable at 1 as in previous metrics.

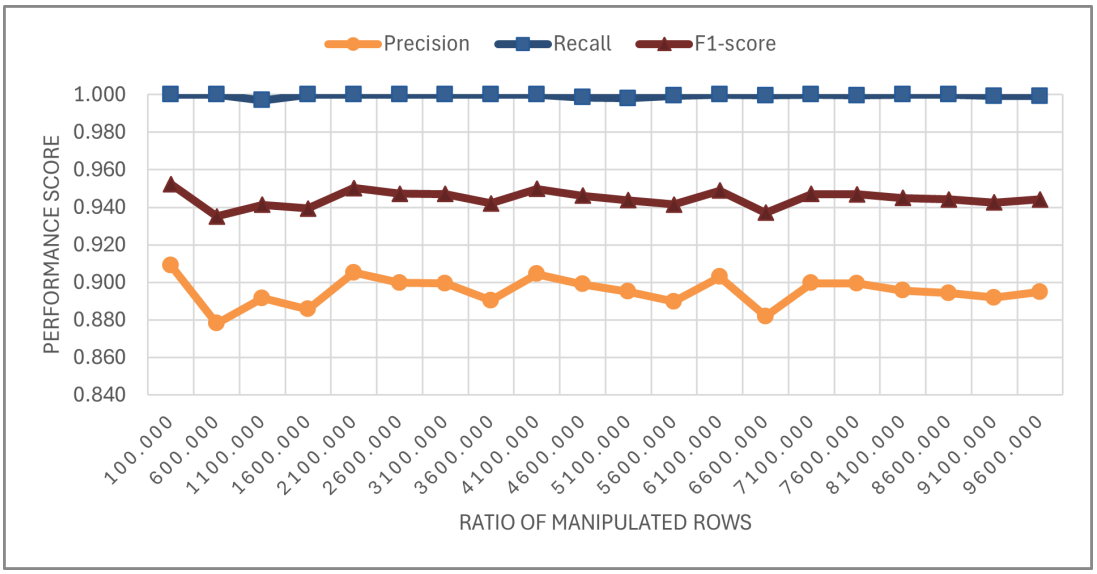


**Figure 4.5:** *Hellinger distance: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

As in the MAD test, in the Kullback-Leibler divergence, there is also no large variation of precision between 100 and 600 rows, cf. Figure 4.6. It values vary between 0.882 and 0.910. *recall* has a value of 1, or very close, as in previous tests (see Figure 4.6). It's a very consistent test, similar to MAD. It's ideal for detecting small deviations with high reliability. It could be a good alternative to the chi-square method.

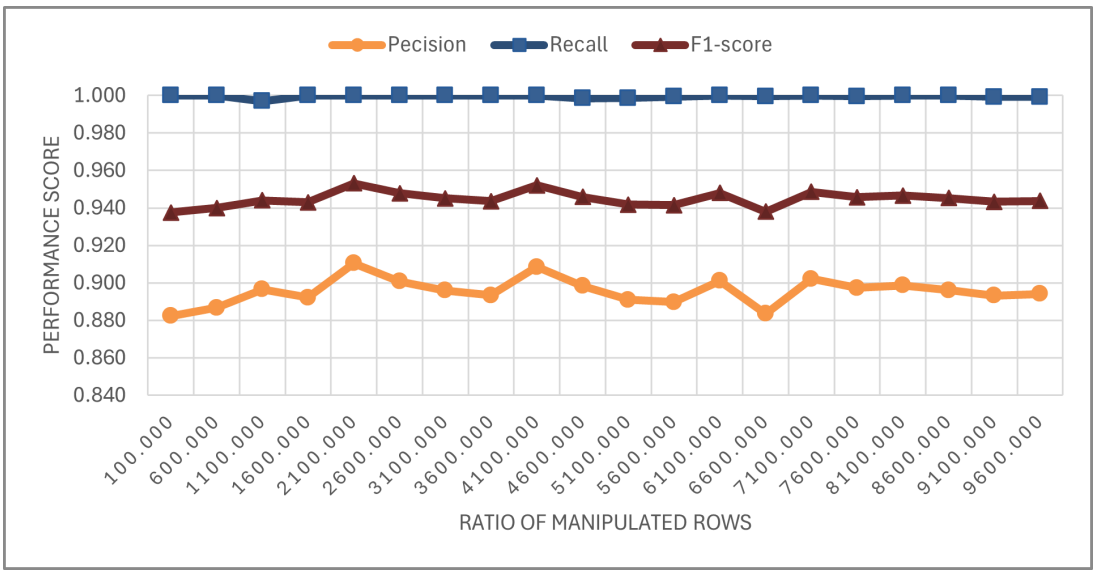


**Figure 4.6:** *KL: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

In the Figure 4.7 we see the combined effect of all previous tests.

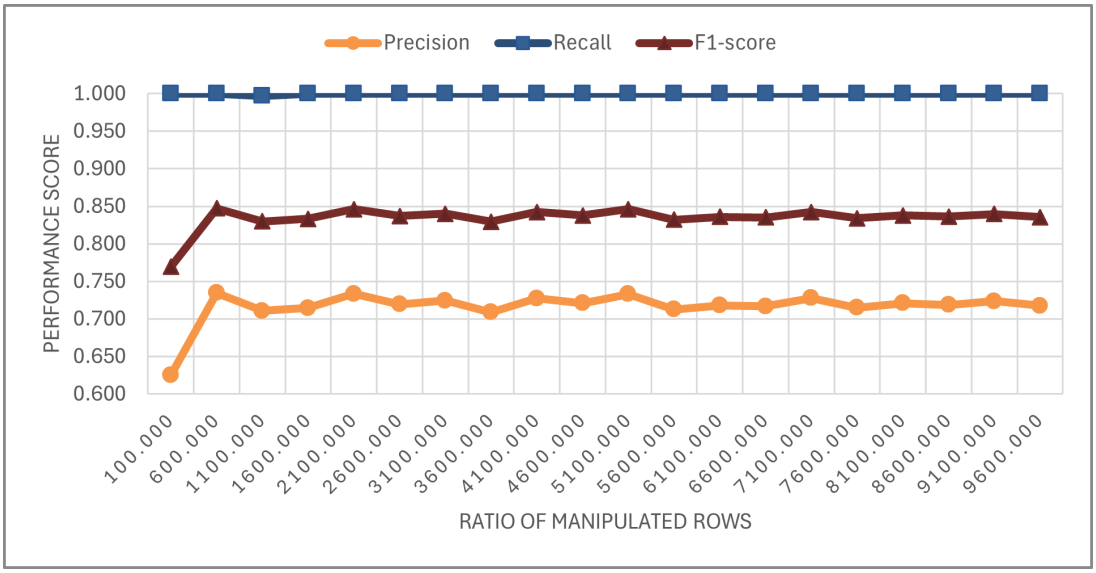


**Figure 4.7:** *Fisher: Sensitivity analysis in relation to the number of cases, 30% of anomalous rows, significance level of 0.05*

The difference in precision is evident for 100 and 600 rows, a difference of 0.110. Between 600 and 9600 rows the value of precision remains between 0.709 and 0.735. The *recall* has a value of 1.

Fisher's combination brings together the $p$-values of different statistical tests, which can make it more sensitive to the presence of any small deviation detected by at least one test. That is, whenever any individual test detects a possible deviation (even by statistical fluctuation), the Fisher combination can consider the case as anomalous, even if most tests do not indicate it. It is concluded that the Fisher combination may be useful as an exploratory or support mechanism, not as a main criterion.

Based on the summary of results for the different statistical tests applied (Table 4.1), it is possible to elaborate a comparative and interpretative analysis of the performance of each method.

| Measure | Chi-square | MAD | Kolmogorov–Smirnov | Euclidean | Hellinger | Kullback–Leibler | Fisher |
|---|---|---|---|---|---|---|---|
| Precision (average) | 0.897 | 0.895 | 0.899 | 0.895 | 0.895 | 0.896 | 0.721 |
| Recall (average) | 1.000 | 0.996 | 1.000 | 0.997 | 1.000 | 1.000 | 1.000 |
| F1-score (average) | 0.946 | 0.943 | 0.947 | 0.943 | 0.944 | 0.945 | 0.838 |
| Accuracy (average) | 0.965 | 0.964 | 0.966 | 0.964 | 0.965 | 0.965 | 0.884 |

**Table 4.1:** *Comparative table of the different methods as the number of cases increasess(30% of anomalous rows, significance level of 0.05)*

These simulations were performed with a significance level of $0.05$. For all rows, the best cut-off point was calculated by the Youden criterion and by the criterion of the smallest distance. In both cases, a cut-off point of $0.249$ was obtained for this analysis.

It was achieved a high and stable accuracy, around 0.965 for almost all measurements except Fisher, showing robustness and equivalent performance.

For all methods (with the exception of the chi-square test), the *recall* is consistently high (very close or equal to 1.00), regardless of the number of rows. This indicates, as expected, that with more than 600 rows, the models maintain or even strengthen their ability to correctly identify positive cases (frauds or deviations).

Precision tends to stabilize or improve slightly as the number of rows increases, this behavior indicates a relative reduction in false positives, suggesting once again that larger samples allow the model to better distinguish between compliance and deviation.

With the *recall* stable at 1 or very close to 1, the F1-*score* follows the variations of the precision but with higher values.

### 4.1.2 Ratio of anomalous rows of 15% with significance level of 0.05

To perform the second simulation of sensitivity analysis in relation to the number of cases, the following parameters were established:

- $m$=1000 (number of columns),
- $n$, quantity of rows, varies between 100 and 9600 by steps of 500, $n = 100 + 500 \times k$ with $k = 0, 1, \cdots, 19$.
- $t_B$=0.15 (ratio of anomalous rows),

- $t_m$=0.25 (ratio of anomalies per anomalous row),
- $\alpha = 0.05$.

The ratio of anomalous rows, $t_B$, was the only parameter that changed (from 0.3 to 0.15).

First of all, it is intended to point out that, in general, the performance of the model has decreased considerably, as can be seen in the Table 4.2.

| Measure | Chi-square | MAD | Kolmogorov-Smirnov | Euclidean | Hellinger | Kullback-Leibler | Fisher |
|---|---|---|---|---|---|---|---|
| Precision (average) | 0.770 | 0.771 | 0.785 | 0.773 | 0.767 | 0.768 | 0.513 |
| Recall (average) | 0.973 | 0.969 | 0.997 | 0.975 | 0.963 | 0.966 | 0.997 |
| F1-score (average) | 0.860 | 0.859 | 0.878 | 0.862 | 0.853 | 0.856 | 0.677 |
| Accuracy (average) | 0.952 | 0.952 | 0.959 | 0.953 | 0.950 | 0.951 | 0.857 |

**Table 4.2:** *Comparative table of the different methods as the number of cases increases (15% of anomalous rows, significance level of 0.05)*

Regarding the accuracy, it is high and consistent ($\geq 0.950$) for all methods, with slight advantage for Kolmogorov-Smirnov. The Fisher method shows a clearly inferior performance.

The *recall* decreased slightly, but nothing significant. While its average, for the different measures, previously was between 0.996 and 1, now it is between 0.966 and 0.997. Regarding the precision, the average for the different measurements is now around 0.77, whereas with 30% of rows anomalous it was around 0.89, the fall has already been a little more pronounced. This resulted from the fact that the number of positive cases decreased, and since the number of cases analyzed remained the same, the number of negative cases increased. Thus, if we assume that the test's ability to identify each category (recall for both categories) remains the same, the number of true positives decreases proportionally to the number of positive cases, and the number of false positives increases proportionally to the number of negative cases. Consequently, precision (of the positive class) decreases, as expected.

## 4.2 Sensitivity analysis in relation to the number of features

Like the number of cases, also the number of features (columns) plays a relevant role in the performance of detection methods based on Benford's Law. A larger number of features may translate into additional information and potentially greater statistical power, but it can also introduce redundancy or noise that compromises the effectiveness of the tests.

In this section, a sensitivity analysis is performed where the number of columns is variable, while the remaining parameters of the simulation process are fixed. For each configuration, the simulations are repeated, allowing to evaluate how the size of the dataset in terms of features affects the consistency of the results obtained. The objective

is to determine if the presence of more features contributes to improve the detection of manipulations or if, from a certain point, the additional gains become marginal.

### 4.2.1 Ratio of anomalous rows of 30% with significance level of 0.05

Analysis of the model's performance based on the number of columns (features) was made with the following parameters:

- $m$, quantity of columns, varies between 100 and 9600 by steps of 500, i.e., $m = 100 + 500 \times k$ with $k = 0, 1, \cdots, 19$,
- $n = 2000$ (quantity of rows),
- $t_B = 0.3$ (ratio of anomalous rows),
- $t_m = 0.3$ (ratio of anomalies by anomalous row),
- $\alpha = 0.05$ (significance level).

This analysis reveals a clear pattern: as the number of columns (features) increases, the model's ability to distinguish between compliant and non-compliant data with the Benford distribution improves substantially as Figure 4.8 clearly shows.

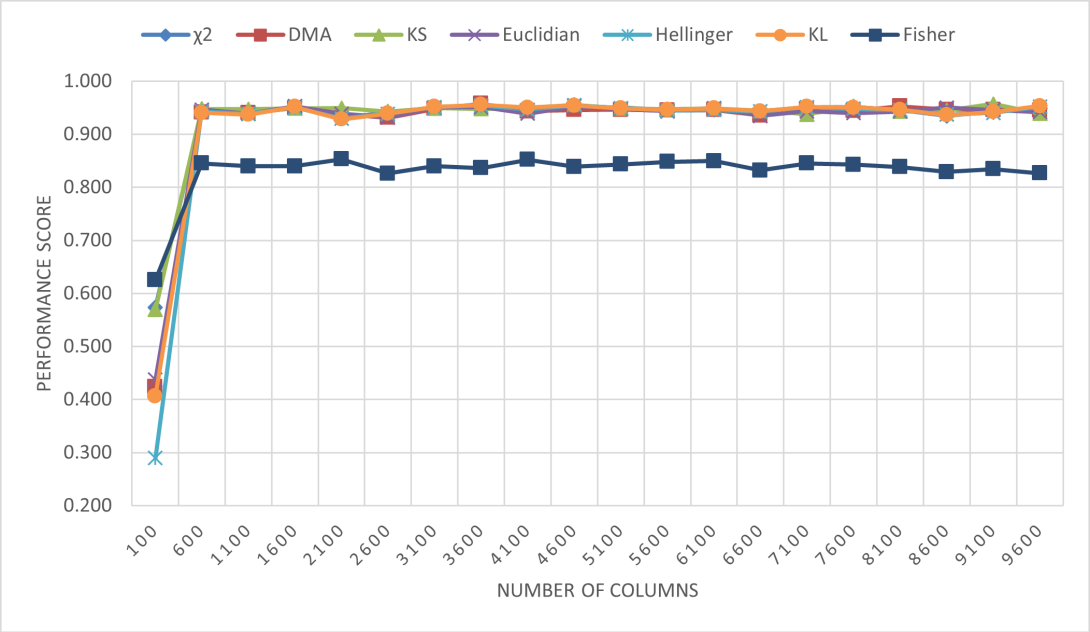


**Figure 4.8:** *F1-score: Sensitivity analysis in relation to the number of features, 30% of anomalous rows, significance level of 0.05*

As the number of columns (features) increases, already from 600 columns, there is a sharp increase in the F1-*score* for all methods, reaching values above 0.90. From about 1100 columns, performance stabilizes at high levels, with F1-scores often above 0.94, and in some cases exceeding 0.95.

In scenarios with a reduced number of columns (e.g., 100), the performance of all methods is significantly weaker, with lower F1-*score* values. This is expected because with fewer data per observation, statistical variability increases and it becomes more

difficult to capture consistent deviation patterns. In this range, the combination of $p$-values by means of the Fisher method was particularly effective, obtaining the best relative performance (F1-*score* $\approx 0.625$), by aggregating weak signals from different tests and making them more robust together. Figure 4.9 details the model performance in this range.

Figure 4.9 shows us that the F1-*score* increases with the number of columns for all methods, which was already noticeable in the previous graph. The Kolmogorov-Smirnov test presents the best overall performance, reaching about 0.93 for $m \geq 500$. The $\chi^2$ test stands out for low values of $m$, but stabilizes before the rest. Distance methods (Euclidean, KL, MAD, Hellinger) converge to results close to 0.90, while the Fisher combination maintains consistent performance but lower than KS for high values of $m$, stabilizing at 0.82 from $m$ equal to 400.

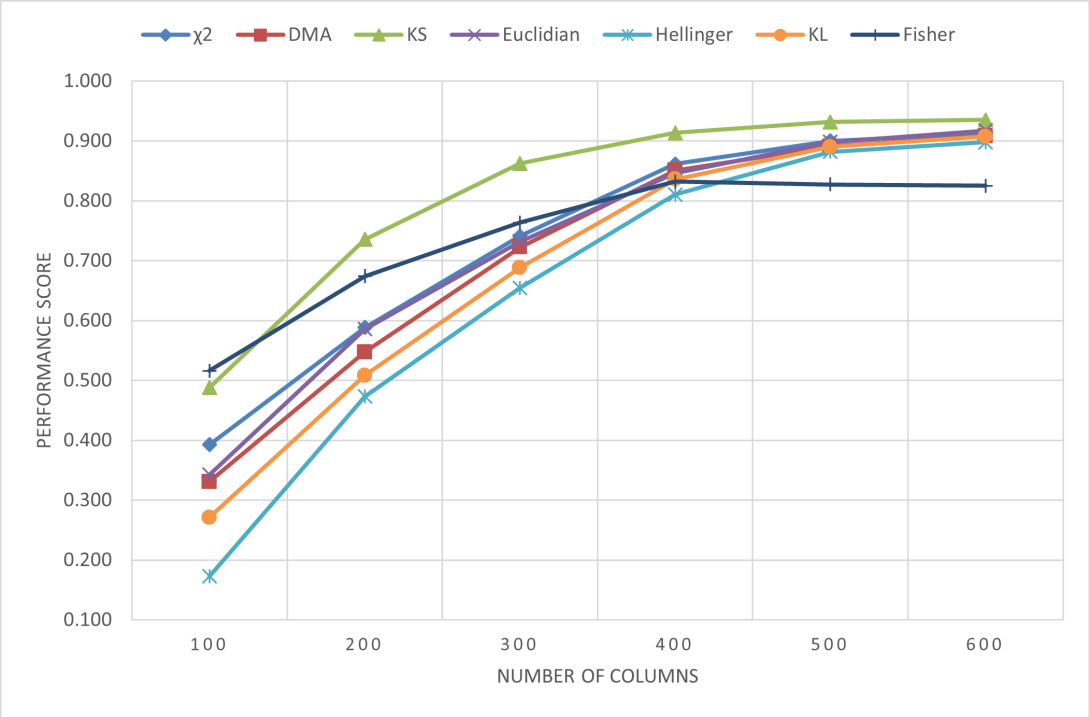


**Figure 4.9:** *F1-score, 100 to 600 columns in detail: Sensitivity analysis in relation to the number of features, 30% of anomalous rows, significance level of 0.05*

Among the methods used, the Kolmogorov-Smirnov (KS) showed the most consistent results, even in scenarios with few features, reduced $m$. The same happens in precision (Figure 4.10).

The precision grows with the number of columns up to about $m = 400$, then stabilising for most methods. The Kolmogorov-Smirnov test consistently outperforms other methods. The distance methods (Euclidean, KL, Hellinger, MAD) have similar values, reaching 0.89. Hellinger's distance starts very low, at 0.48 but approaches quickly and with $m = 300$ reaches the same value as KL. The $\chi^2$ maintains high values but slightly lower than KS. The Fisher combination presents inferior and more unstable performance, not exceeding 0.75.

The behavior of the model in a scenario with more features is shown in Figure 4.11

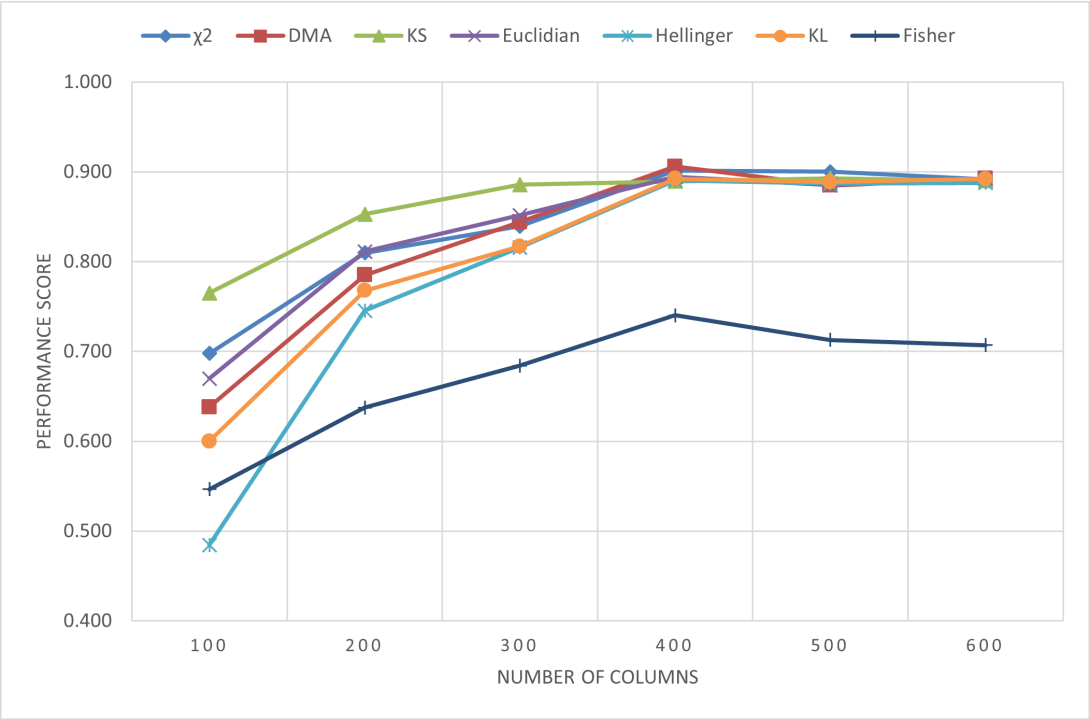


**Figure 4.10:** *Precision, 100 to 600 columns in detail: Sensitivity analysis in relation to the number of features, 30% of anomalous rows, significance level of 0.05*

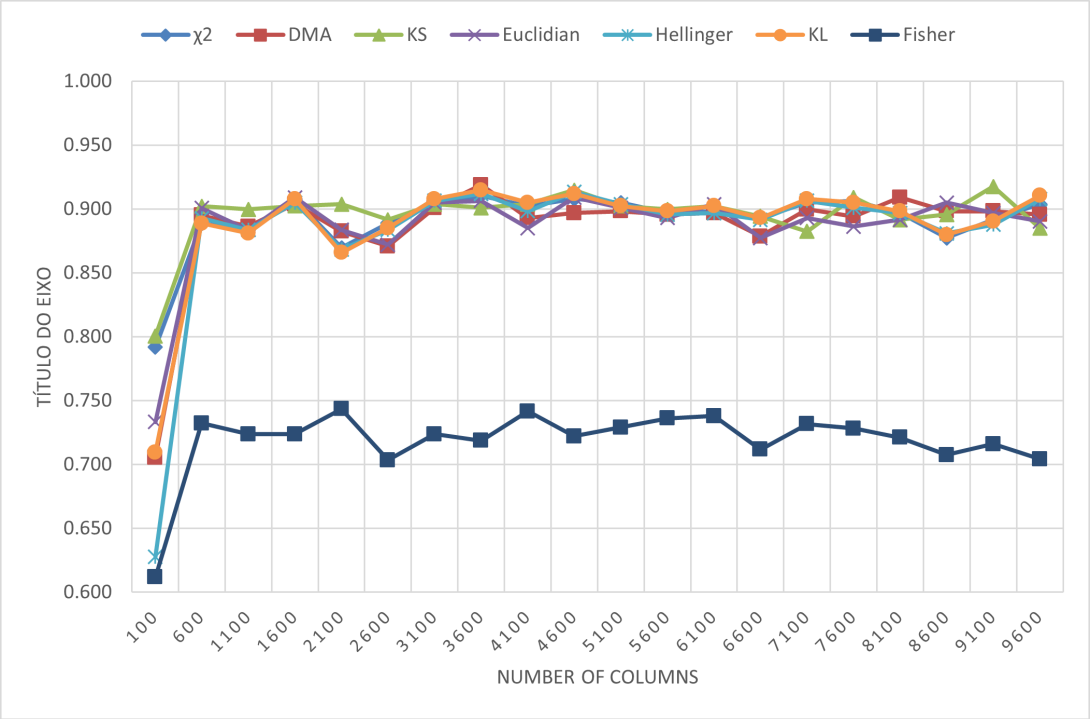


**Figure 4.11:** *Precision: Sensitivity analysis in relation to the number of features, 30% of anomalous rows, significance level of 0.05*

The MAD method, although simpler, had one of the best absolute precision value

(0.92) in an isolated instance, in equality with KS, showing that less complex methods can also perform well in specific scenarios.

However, as noted in Figures 4.8 and 4.11, the combination of $p$-values by Fisher's method, despite having lost relative relevance with the increase of the number of columns it continues to offer competitive performance and can be advantageous in situations with less informative data or in contexts where multiple sources of statistical evidence are desired. In Table 4.3, we have the comparison of the different methods.

| Method | Medium Precision | Medium Recall | Medium F1-score | Medium Accuracy |
|---|---|---|---|---|
| **Chi-square** | 0.897 | 1.000 | 0.945 | 0.965 |
| **MAD (medium absolute dev.)** | 0.895 | 1.000 | 0.945 | 0.965 |
| **Kolmogorov-Smirnov** | 0.900 | 1.000 | 0.947 | 0.967 |
| **Euclidean distance** | 0.894 | 1.000 | 0.944 | 0.964 |
| **Hellinger** | 0.896 | 1.000 | 0.945 | 0.965 |
| **Kullback-Leibler** | 0.898 | 1.000 | 0.946 | 0.966 |
| **Fisher combination** | 0.724 | 1.000 | 0.840 | 0.886 |

**Table 4.3:** *Comparative table of the different methods as the number of columns increases (30% of anomalous rows, significance level of 0.05)*

Therefore, the model presents optimal and stable performance from 1100 columns, with *recall* = 1 and high F1-*score*. The calculation of the optimal cut-off, both by the method of Youden and by the method of the smallest distance is consistent from this point, always giving the value 0.001 (see Figure 4.12). The results suggest using at least 1100 features to ensure robust performance; ideal between 1600 and 3600 features for maximum precision.

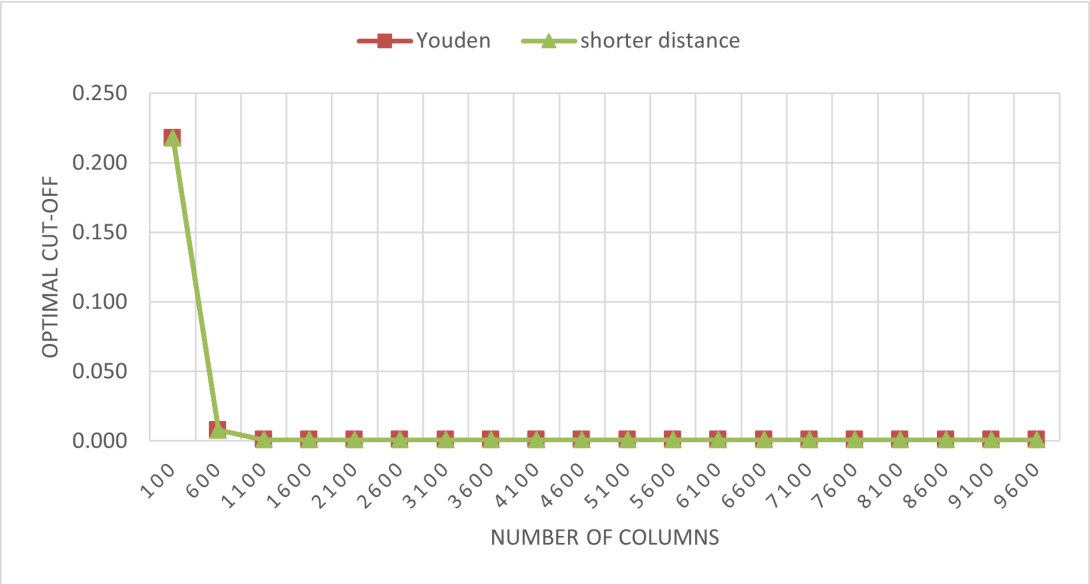


**Figure 4.12:** *Optimal cut-off point as the number of columns increases, 30% of anomalous rows, significance level of 0.05*

From 1100 features, the calculation for the optimal cut-off point (both for the Youden criterion and for the smallest distance) gives us the value of 0.001 and remains constant. This suggests that with more than 1100 columns, which suggests that the model

reaches a stability in cut-off point choice at 0.001.

For a reduced value of features, as shown in Figure 4.13, we can see a fairly consistent reduction of the optimal cut-off point by the method of the smallest distance. For the Youden method, the behavior of the ideal cut-off point is already more unstable.

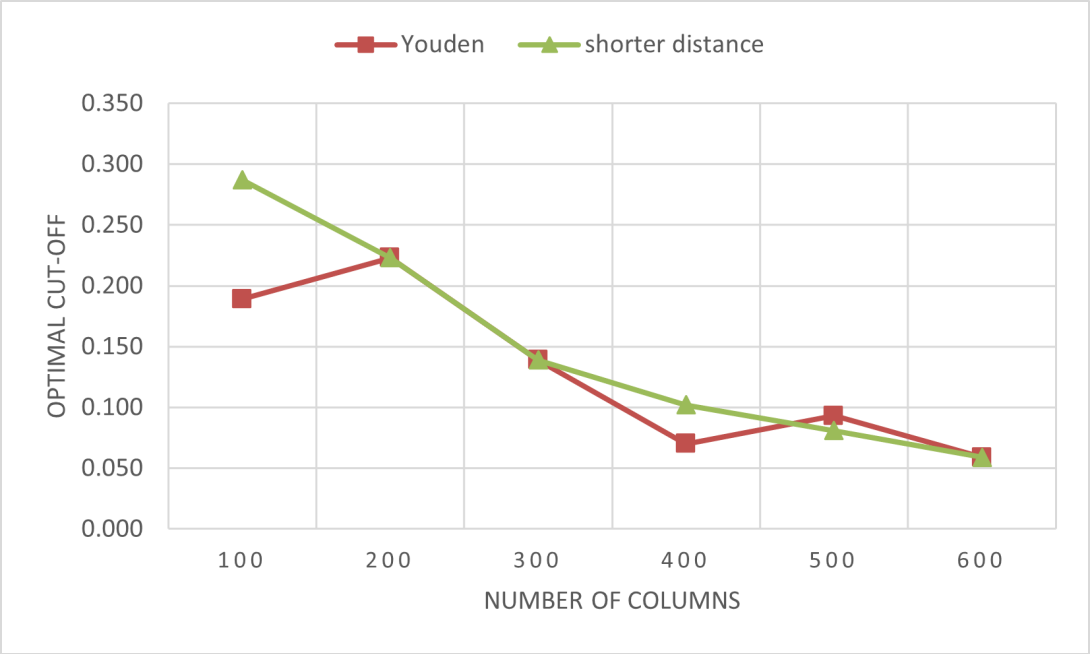


**Figure 4.13:** *Optimal cut-off point as the number of columns increases, from 100 to 600 features, 30% of anomalous rows, significance level of 0.05*

In summary, the results show that the model benefits significantly from a higher granularity of the data. As the number of columns (features) increases, the model becomes more reliable, robust and discriminative, a predictable behavior for anomaly detection systems based on Benford's Law.

### 4.2.2 Ratio of anomalous rows of 10% with significance level of 0.05

For this simulation we used the parameter $t_B$ of 10%, being the model configured with the following parameters:

- $m$, quantity of columns, varies between 100 and 9600 by steps of 500, i.e., $m = 100 + 500 \times k$ with $k = 0, 1, \cdots, 19$,
- $n = 2000$ (quantity of rows),
- $t_B = 0.1$ (ratio of anomalous rows),
- $t_m = 0.3$ (ratio of anomalies by anomalous row),
- $\alpha = 0.05$ (significance level).

As pointed out in the sensitivity analysis in relation to the number of cases, by decreasing the number of rows with anomalies, we are increasing the number of negative cases (reducing positives cases and maintaining the total number of cases), leading to an increase in false positives (assuming the same misclassification rate among negative

cases). The significance level was set at 0.05, that is, there is a 5% probability of rejecting the null hypothesis (that the data follow Benford's Law) when it is true, wich means that even if all the data are legitimate (not anomalous), the test will classify about 5% of cases as irregularities by pure statistical chance. Now, with only 10% of cases with anomalies, you enter a critical situation, the number of real anomalies (true positives) is almost the same size as the number of expected false positives, which makes the false positive rate soar proportionally. As the number of false positives increased, this will have a direct impact on precision as can be seen in Table 4.4

| Method | Medium Precision | Medium Recall | Medium F1-score | Medium Accuracy |
|---|---|---|---|---|
| Chi-square | 0.690 | 0.996 | 0.816 | 0.955 |
| MAD (medium absolute dev.) | 0.690 | 0.995 | 0.815 | 0.955 |
| Kolmogorov-Smirnov | 0.697 | 0.999 | 0.821 | 0.956 |
| Euclidean distance | 0.695 | 0.995 | 0.818 | 0.956 |
| Hellinger | 0.689 | 0.994 | 0.814 | 0.955 |
| Kullback-Leibler | 0.689 | 0.996 | 0.814 | 0.955 |
| Fisher combination | 0.401 | 0.999 | 0.572 | 0.851 |

**Table 4.4:** *Comparative table of the different methods as the number of features increases (10% of anomalous rows, significance level of 0.05)*

It can also be observed in the Table 4.4 that the *recall* remained very close to 1, despite the decrease in anomalous cases.

In Figure 4.14, is possible to verify that the different measures (except Fisher) have identical behavior between them.

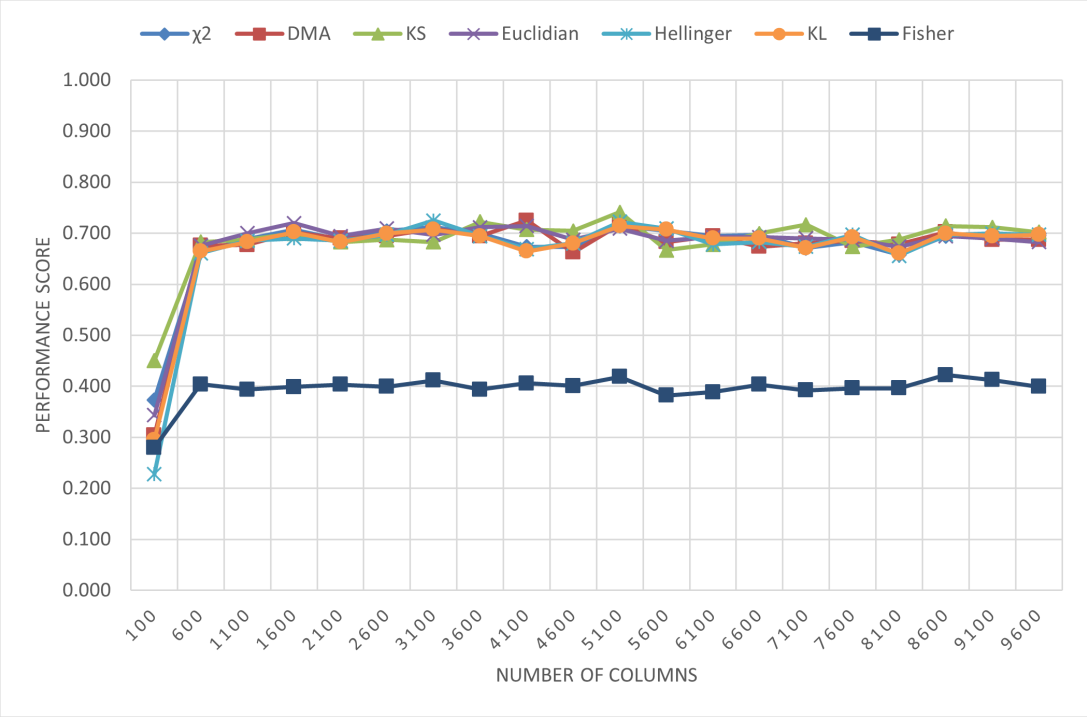


**Figure 4.14:** *Precision: Sensitivity analysis in relation to the number of features, 10% of anomalous rows, significance level of 0.05*

From 100 to 600 there is a very large jump, from an uncertain zone to the beginning

of stability, where from there forward, there does not seem to be significant changes. How will be the behavior of precision in this range? This analysis can be performed based on the evolution of precision shown in Figure 4.15.

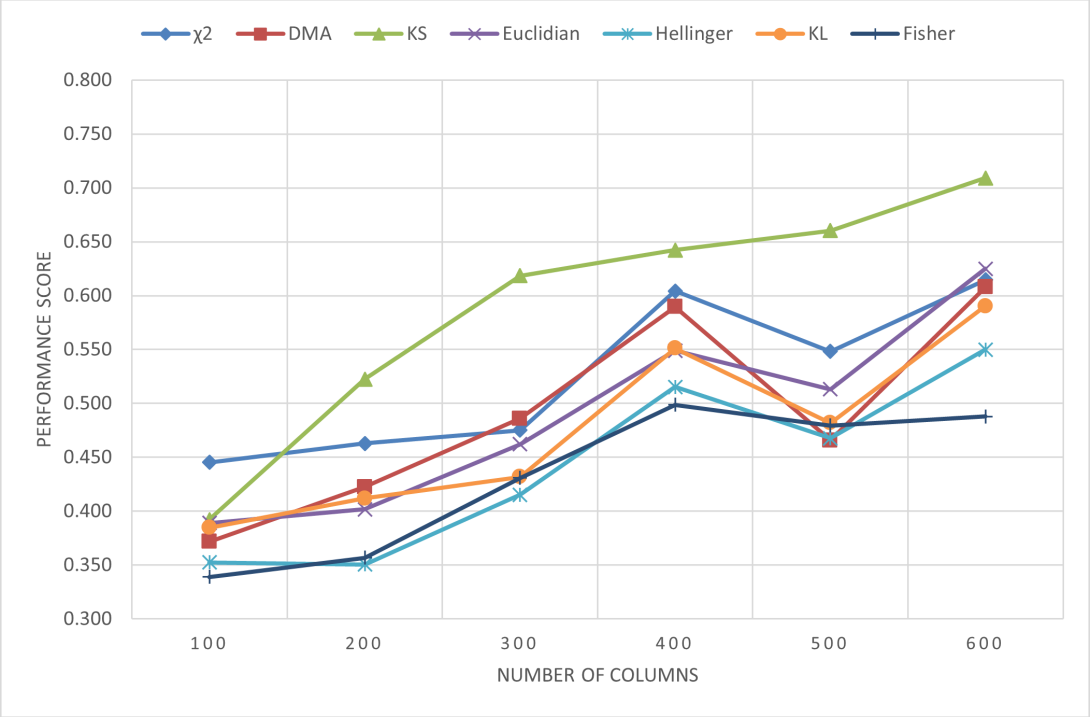


**Figure 4.15:** *Precision, 100 to 600 features in detail: Sensitivity analysis in relation to the number of features, 10% of anomalous rows, significance level of 0.05*

In this range, the KS test is the one with the highest curve at all points, growing consistently to exceed 0.70 (precision) with 600 columns. Hence, KS is the most robust test.

The $\chi^2$ remains relatively stable, without such significant gains. DMA, Euclidian, KL and Hellinger distances have similar performances, growing up to close to 0.60 with 600 columns. Fisher is the method with the weakest performance in the whole range, not exceeding 0.50. It is interesting the peak that exists around the 400 columns. All methods with the exception of KS and Fisher show a peak in precision at 400 columns, followed by a drop at 500 and further recovery at 600. It seems to suggest that gains are not always linear.

## 4.3 Sensitivity analysis in relation to the anomalous cases

In this section, the focus falls on anomalous cases, that is, data lines that, for any reason, do not follow Benford's Law, whose presence and intensity of change can significantly impact the performance of evaluation methods based on Benford's Law. Thus, we proceed to investigate different scenarios in which the proportion of abnormal cases varies, in order to verify whether the statistical tests applied maintain their discriminatory capacity for different degrees of non-compliance with Benford's Law. This approach

allows not only to evaluate the consistency of the results under alternative conditions, but also to identify possible limits of sensitivity of the methods used.

### 4.3.1 Ratios of anomalies per anomalous row of 10% with significance level of 0.05

Starting with the description of the parameters used:

- $m = 1000$ (quantity of columns),
- $n = 5000$ (quantity of rows),
- $t_B = 0.05 + 0.10 \times k$, with $k = 0, 1, \cdots, 9$ (ratio of anomalous rows),
- $t_m = 0.1$ (ratio of anomalies per anomalous row),
- $\alpha = 0.05$ (significance level).

Figure 4.16 shows de behavior of the model with chi-square test while the number of anomalous rows increases.

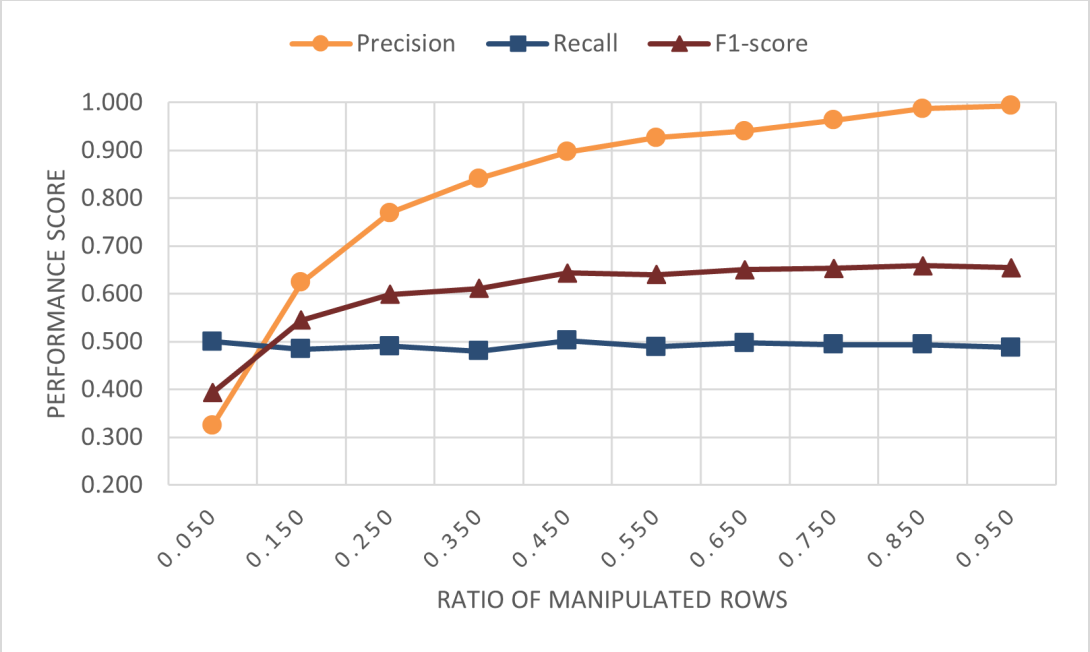


**Figure 4.16:** *Chi-square: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

Precision increases with more data anomalous rows. It consistently grows from 0.32 ($t_B = 5\%$) to 0.99 ($t_B = 95\%$). Since *recall* remains practically constant in all scenarios, the number of true positives increases proportionally with the increase in handled cases. As a result, false positives become less relevant in the precision calculation, leading to the observed growth.

Despite the increase in precision, *recall* did not improve significantly, remaining at $\approx 0.49 - 0.50$ in all ratios. This means that the model does not improve the ability to find all true positives (keeps the false negative rate similar). This may be a potential limitation if the goal is to maximize F1-*score*. However, as we will analyze in more

detail, the ratio of manipulations per manipulated row is fundamental to distinguish an anomalous case from one that is not, so with $t_B = 0.1$ the differences between these two cases are not statistically significant.

Initially $(5\% - 15\%)$, the F1-*score* rises from $0.39$ to $0.65$. From ratios of rows anomalous of $0.45$, it stabilizes around $0.64/0.66$. This shows that there is no significant gain in F1-*score* when increasing manipulation beyond $\approx 45 - 55\%$.

By increasing the ratio of anomalous rows, the number of positive instances increases. Thus, TN (true negatives) decreases drastically from 4489 to 235 mostly because there are simply fewer negatives in the data, so the maximum possible TN goes down. And FN (false negatives) increasing from 125 to 2432, which means the model is failing to detect anomalous rows.

The behavior obtained is due to the fact that classical statistical models are based on measuring distances between observed values and expected values under the null hypothesis $(H_0)$. The decision as to whether an observation is normal or anomalous is made based on this distance, according to a pre-established threshold. In this paradigm there is no learning, since the model does not fit based on labeled examples or iterative feedback, it is a fixed and non-adaptive approach. Consequently, this lack of learning has direct implications on performance metrics.

One implication is the constant *recall*. The model's ability to correctly detect anomalous instances (true positives) tends to remain relatively constant regardless of the proportion of anomalies in the datset. This is because the detection logic is invariable in relation to the distribution of data. However, when the proportion of abnormal rows is very small (i.e., very small number of positives), the *recall* value presents high variability due to the low sample base, stabilizing only when the number of anomalies becomes statistically representative.

Another implication is the increasing in precision as the amount of anomalies increases. As the percentage of anomalous rows in the datset increases, the number of true positives among examples classified as anomalous tends to increase. As a consequence, the proportion of false positives decreases relatively, which leads to an increase in precision.

This leads to a slightly increased F1-*score*. As the *recall* remains approximately constant and the precision increases, the F1-*score* (which is the harmonic mean between these two metrics) also increases. However, the growth of F1-*score* is less pronounced than that of precision, reflecting its harmonic nature and the stabilization of *recall*.

Figure 4.17 describes the evolution of precision, *recall* and F1-*score* behavior as the number of anomalous rows increases using the MAD.

As shown in Figure 4.17, the precision increases progressively, as in the chi-square test. It starts at $0.277$ ($5\%$ anomalous rows) and goes up to $0.99$ ($95\%$ anomalous rows). As in the chi-square, the model reduces false positives (FP from 258 to 14) as it increases the ratio of anomalous rows. Again, the more rows anomalous, the higher the precision.

The *recall* remains practically constant, starting at 0.396 ($5\%$) and oscillates between

$\approx 0.34 - 0.37$ for the remaining ratios. Unlike the chi-square, which kept the *recall* close to $0.49 - 0.50$, here the *recall* is consistently lower ($\approx 0.35$). The model fails even more in identifying true positives compared to the chi-square.

The F1-*score* stabilizes earlier and in lower value, rises from $0.32$ $(5\%)$ to $0.53$ $(85\% - 95\%)$. Although there is improvement, it reaches a lower value than with the chi-square. Initial improvement evident, but from $0.45$ there is no significant gains in F1-*score*.

The Kolmogorov-Smirnov performance analysis is based on Figure 4.18 that describes the evolution of precision, *recall* and F1-*score* as the number of rows anomalous increases using the KS test.

As Figure 4.18 illustrates, the precision increases with the ratio of anomalous rows, keeping the pattern. It starts at $0.36$ $(5\%)$ and goes up to $0.99$ $(95\%)$. As with the other methods, increasing the ratio of anomalous rows improves precision and reduces false positives. The behavior is consistent with other statistical tests in this respect.

Recall remains more stable and slightly higher than the previous methods. It varies between $0.54$ and $0.57$, higher than the mean absolute deviation ($\approx 0.35$) and similar or slightly above the chi-square ($\approx 0.50$). Even with an increase in the ratio of anomalous rows, *recall* does not degrade significantly. This suggests that the model based on Kolmogorov-Smirnov can better maintain the ability to identify true positives.

The F1-*score* reaches the highest value among all methods, F1-*score* starts at $0.43$ $(5\%)$ and reaches $0.71$ $(85\%)$. It is consistently higher than the absolute mean deviation ($\approx 0.53$) and slightly better than the chi-square ($\approx 0.66$). Thus, it reveals better overall balance between precision and *recall*.

The F1-*score* growth stabilizes later, continues to improve up to $0.85$ of anomalous rows, then stabilizes close to $0.71$. This shows that the Kolmogorov-Smirnov withstands better performance gains even with high data manipulation rate rows. Table 4.5 makes clear that KS has the highest values of precision, *recall* and F1-*score* among all measures studied.

Now about Euclidean distance method, Figure 4.19 shows the performance of the model using this method when increasing the ratio of anomalous rows.

As Figure 4.19 reveals, the precision increases with the anomalous ratio, consistent pattern with the other methods above. Starts at $0.26$ $(5\%)$ and goes up to $0.99$ $(95\%)$. As with the other methods, increasing the anomalous ratio reduces false positives. The expected behavior was obtained, as with the other methods.

The *recall* remains low and practically constant, it varies between 0.34 and 0.36, without major variations. It is the lowest *recall* value among all methods analyzed. This indicates that the Euclidean distance has less ability to detect true positives.

The F1-*score* remains lower throughout the range, starting at $0.30$ $(5\%)$ and growing only to $0.53$ $(95\%)$. Does not exceed the values achieved by Kolmogorov-Smirnov ($\approx 0.71$) or by chi-square ($\approx 0.66$). This indicates a lower balance between precision and *recall*.

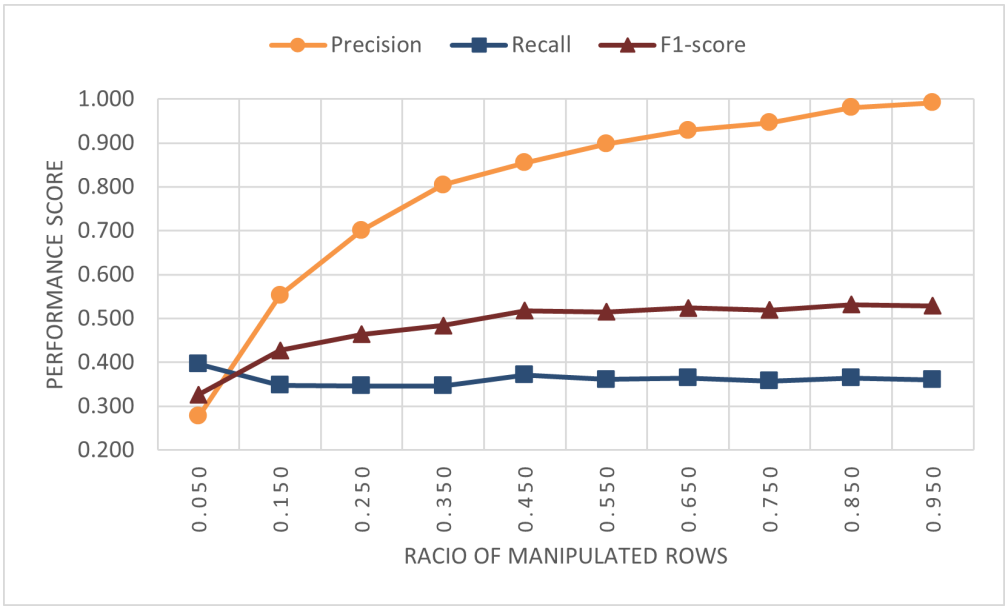


**Figure 4.17:** *MAD: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

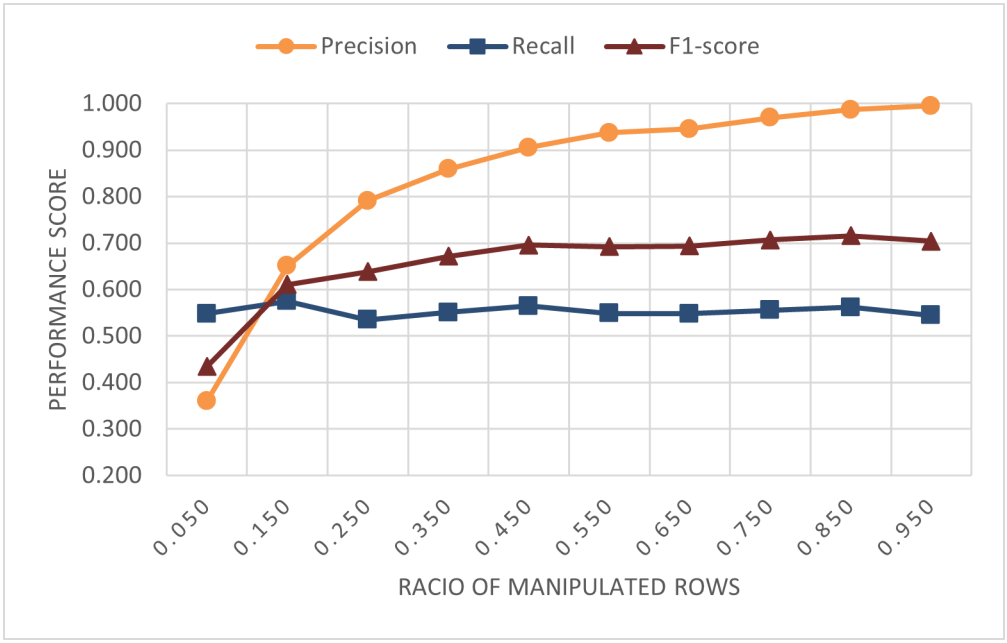


**Figure 4.18:** *KS: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

| Measure | $\chi^2$ (max) | MAD (max) | KS (max) | Eucl. (max) | Hell. (max) | KL (max) | Fisher (max) |
|---|---|---|---|---|---|---|---|
| **Precision** | 0.994 | 0.992 | 0.995 | 0.991 | 0.992 | 0.993 | 0.988 |
| **Recall** | 0.502 | 0.396 | 0.575 | 0.371 | 0.431 | 0.457 | 0.727 |
| **F1-score** | 0.658 | 0.531 | 0.716 | 0.527 | 0.586 | 0.614 | 0.825 |

**Table 4.5:** *KS: Summary of the results of the evolution of the ratio of anomalous rows, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

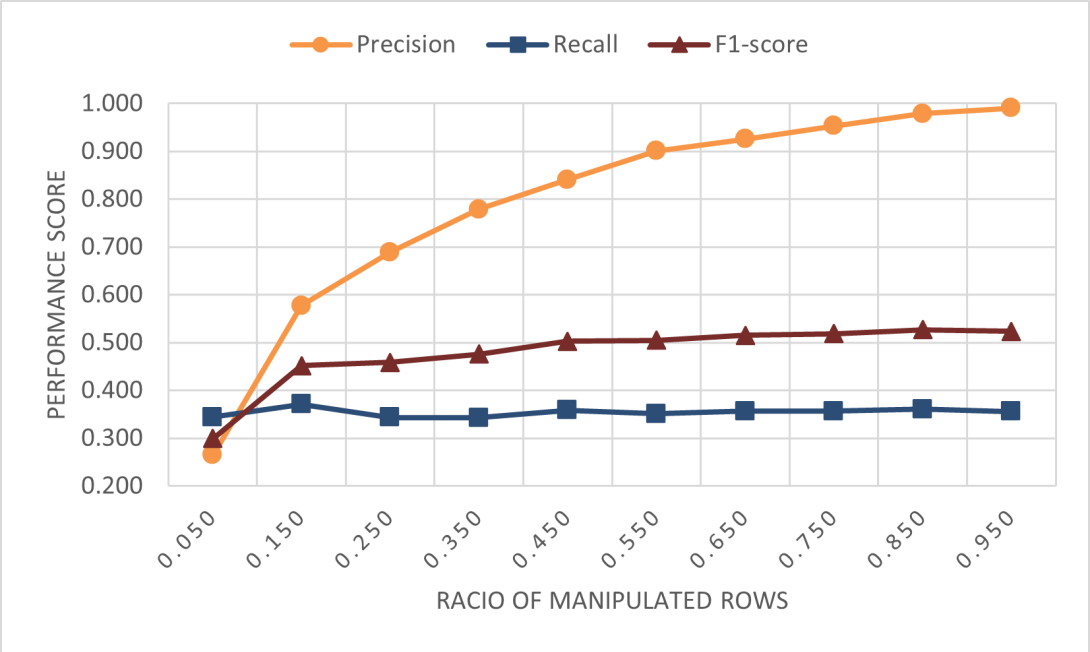


**Figure 4.19:** *Euclidean distance: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

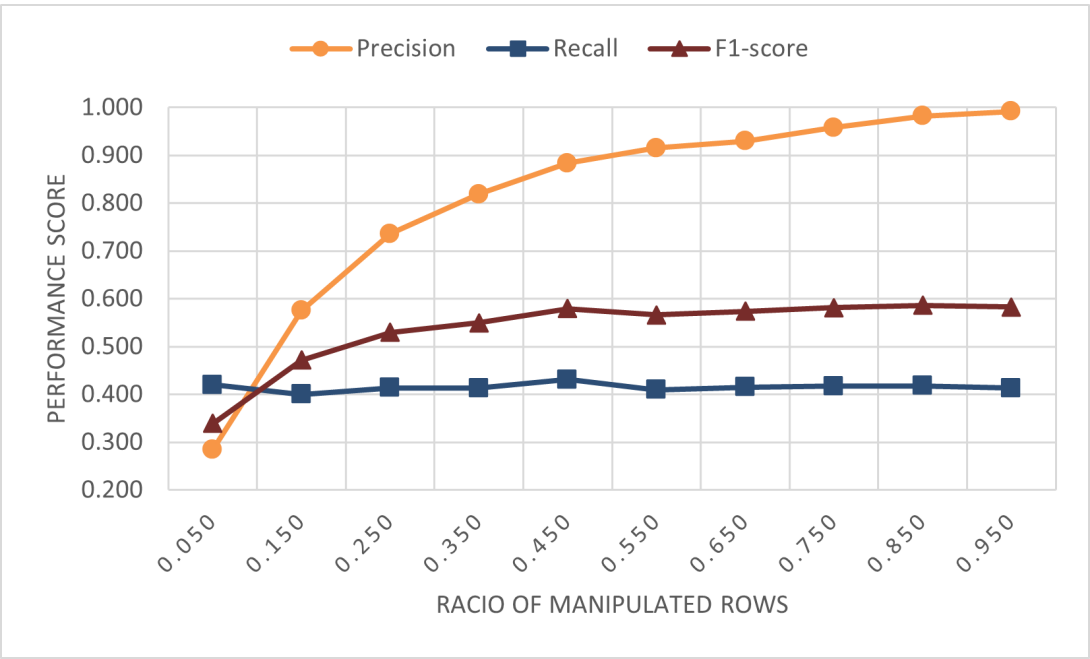


**Figure 4.20:** *Hellinger distance: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

With the Hellinger distance the precision increases consistently, starting at $0.28$ $(5\%)$ and reaching $0.99$ $(95\%)$. Progression follows the same pattern we saw in other methods, fewer false positives as the anomalous ratio increases. Therefore, it is the expected behavior (see Figure 4.20).

The *recall* remains stable, in the range of $0.40$ to $0.43$, with little variation over the ratios. It is better than the euclidean distance *recall* and mean absolute deviation, but

lower than the Kolmogorov-Smirnov *recall*. Although stable, it does not reach the best values.

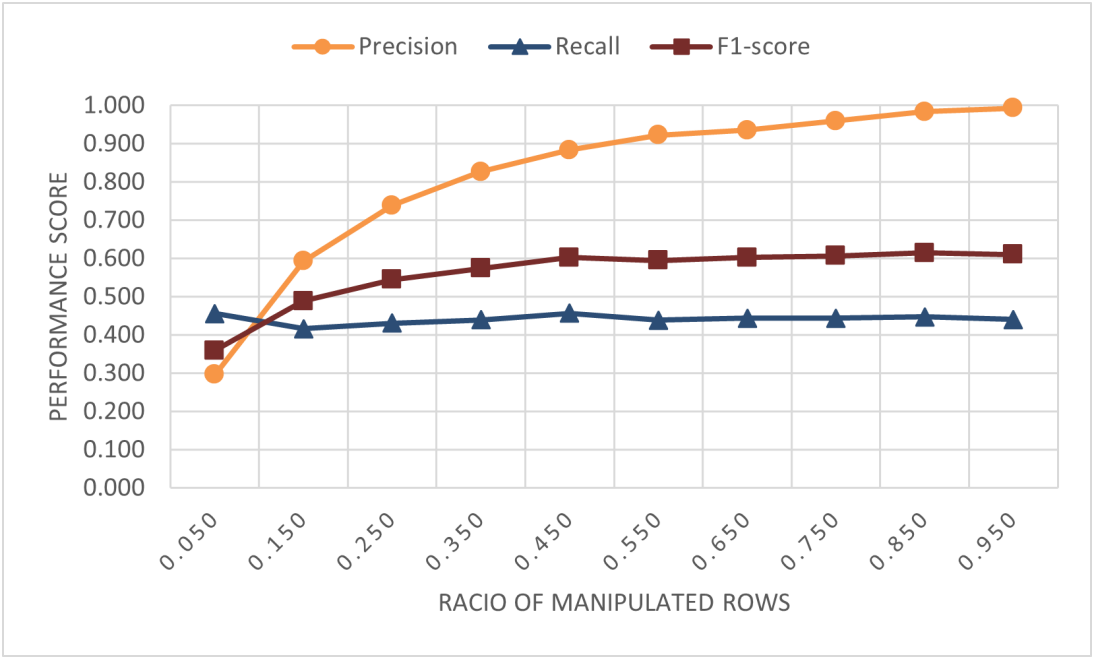


**Figure 4.21:** *KL: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

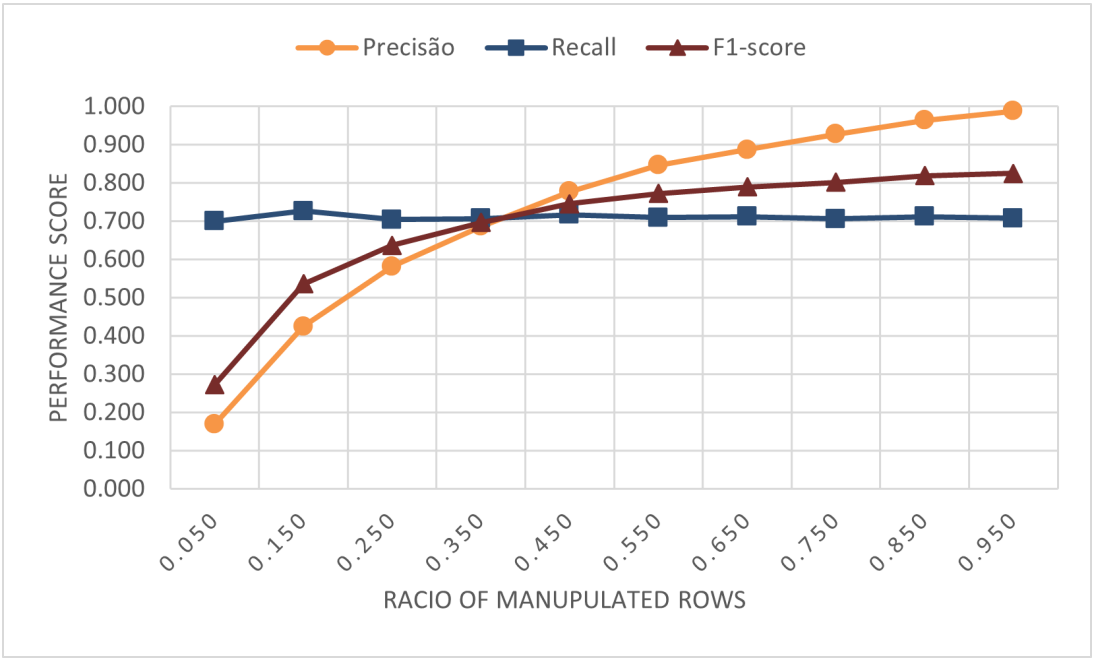


**Figure 4.22:** *Fisher: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

The F1-*score* progressively improves, starting at $0.34$ $(5\%)$ and rising to $\approx$ $0.59$ $(85\%)$, dropping slightly to $0.58$ $(95\%)$. It is better than the mean absolute deviation and euclidean distance but lower than the Kolmogorov-Smirnov and slightly below the

chi-square. This shows a slightly better balance between precision and *recall*.

Analyzing the Kullback-Leibler divergence, the precision starts at $0.296$ $(5\%)$ and goes up to $0.993$ $(95\%)$, see Figure 4.21. It follows the same general pattern observed, progressive increase in precision with more anomalous sets, or be expected and consistent with the other methods.

Recall ranges from $0.456$ $(5\%)$ to about $0.44$ $(95\%)$, with small fluctuations and remains above $0.43$ throughout the graph. It is superior to the *recall* of the euclidean distance, mean absolute deviation and Hellinger but inferior to Kolmogorov-Smirnov.

As for the F1-*score*, it grew from $0.36$ $(5\%)$ to a peak of $\approx 0.61$ $(85\%)$, then slightly decreased to $0.61$ $(95\%)$. It is better than the mean absolute deviation ($\approx 0.53$), euclidean distance ($\approx$ $0.53$) and Hellinger distance ($\approx$ $0.59$) but lower than the Kolmogorov-Smirnov ($\approx 0.71$) and slightly below the chi-square ($\approx 0.66$).

Finally, analyzing the results of the Fisher combination (Figure 4.22), the precision starts at $0.17$ $(5\%)$ and grows to $0.99$ $(95\%)$. It is lower in the first anomalous ratios compared with the other methods, but grows rapidly from $0.15$, which is expected because it follows the pattern of the other methods.

Recall remains high and stable ($\approx$ $0.70/0.72$) over all ratios (Figure 4.5). Fisher delivers the highest *recall* among all methods.

The F1-*score* increased from $0.27$ $(5\%)$ to $0.82$ $(95\%)$. Fisher outperformed all other methods in the F1-*score*. It stands out because instead of relying on a single test, it integrates the information from the other statistical tests at the same time and this makes him much better at detecting hidden anomalies even when each isolated method fails to perceive the complete pattern.

Therefore, the performance of the classification model reveals a consistent pattern influenced by the imbalance of the dataset, reflected in the disparity between the number of anomalous and not anomalous examples. It should be noted that all simulations were performed with a significance level of 0.05, which establishes a common criterion for the selection of relevant variables in each method.

As the ratio of anomalous rows increases from 0.05 to 0.95, there is a general trend of growth in precision in all methods, reaching values close to 0.99 in the higher ratios. However, the *recall* remains approximately constant. The ability of statistical tests to correctly identify an anomalous row (i.e., to generate true positives) depends only on the statistical discrepancy of this row from the expected behavior under $H_0$. This capability is, by definition, independent of the proportion of rows anomalous, which results in a *recall* practically invariant as that proportion varies.

The precision increases with increasing proportion of anomalies. The precision is defined as the ratio between true positives (TP) and the total of positive ratings (TP + FP). When the proportion of anomalous (positive) rows is increased, the absolute number of true positives (TP) also increases, assuming that the *recall* [TP/(TP+FN)] remains constant. In parallel, as the total number of negatives (TN+FP) decreases and the model maintains an approximately constant rate of false positives per negative [FP/(TN+FP)], the number of false positives (FP) tends to decrease. Thus, with

TP increasing and FP decreasing, the precision [TP/(TP+FP)] necessarily tends to increase. It is important to emphasize that this behavior results from the structure of the model, being therefore a different phenomenon from that observed in supervised methods optimized by machine learning.

It is important to distinguish the behavior of the statistical model used in this study from the behavior of supervised learning models. In machine learning models, especially when the objective function aims to maximize aggregate metrics such as precision, the presence of unbalanced classes can lead to classification bias. In these cases, the model can maximize the objective function by classifying the majority of observations in the dominant class, ignoring the minority class.

Among the methods analyzed, the combination of Fisher stood out as the most robust approach, achieving F1-*score* of approximately 0.82 and a *recall* of 0.72 in the highest ratio of anomalies (0.95), clearly surpassing the rest. Fisher was able to maintain a high level of precision without sacrificing both *recall*, unlike methods like Chi-square, Hellinger or Kullback-Leibler, which achieve high precision but remain with lower *recall* (typically below 0.50) and F1-scores more modest (not exceeding 0.66 in the best case). This better balance results from the ability of Fisher's combination to integrate statistical evidence from other tests, allowing more subtle signs of manipulation that isolated methods cannot detect.

After the initial application of the methods with the significance level of 0.05, the best cut-off point was determined. The results obtained, measured by the Youden index and by the smallest distance to the point $(0, 1)$ of the ROC curve, reveal relatively high values in all relationships of the anomalous rows, ranging between 0.172 and 0.261, as Figure 4.23 is showing. Thus, the trade-off between precision and *recall* improves with higher cut-off points relatively that used with $\alpha = 0.05$.

In summary, the model demonstrates high precision and relatively stable performance as the ratio of manipulates increases. Fisher's combination is the most promising technique in the evaluated set, but for a robust practical application it will be necessary to incorporate additional strategies that mitigate the effects of imbalance and enhance the detection capacity of anomalous cases.

All previous tests were based on a significance level of 0.05. However, the optimal cut-off point was calculated with two different methods, the Yoden index and the method of the lowest deviation to the point $(0, 1)$ of the ROC curve. The results obtained show an interesting evolution of the ideal cut-off point along the different ratios of anomalous rows. There is a general trend of increasing these values as the ratio of anomalous in the dataset grows, which indicates that with more data anomalous, the model can identify more clearly the optimal point that maximizes the balance between *recall* and precision. Despite this growth trend, there is a non-linear behavior, with some fluctuations in the average ratios (such as 0.45 and 0.65), suggesting that the ability of the model to find an ideal trade-off does not always evolve uniformly. This may be associated with the increased complexity of these intermediate scenarios, where the anomalies pattern can be more difficult to distinguish.

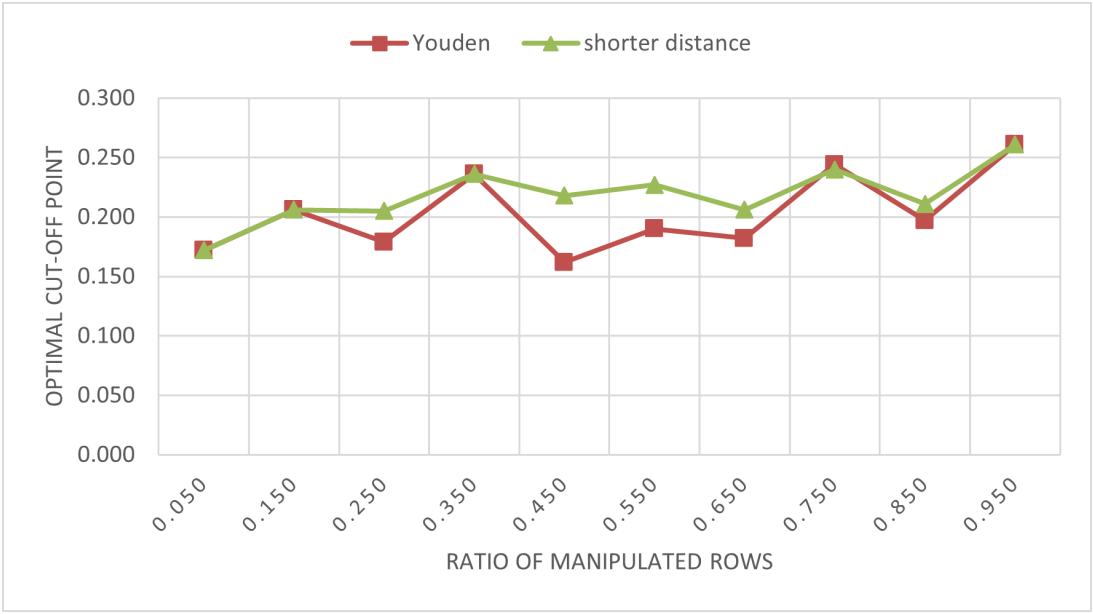


**Figure 4.23:** *Evolution of optimal cut-off point while ratio of anomalous rows increases, ratio of anomalies per anomalous row of 10%, significance level of 0.05*

Another relevant point is the consistency between the two methods (Youden and smaller distance), whose values were very close in all tested ratios. This convergence gives robustness to the conclusions and indicates that both methods provide similar recommendations for the optimal cut-off point in this context.

Finally, it is noteworthy that the highest values were observed in the extremities, particularly in the ratio of 0.95, where the Youden index reached 0.261. This suggests that the model finds the most robust optimal points when the ratio of anomalous rows is too low or too high, probably because the patterns become clearer in these extreme scenarios.

In summary, the model presents a positive evolution in the definition of the optimal cut-off point as it increases the ratio of anomalous data, but this evolution is not totally linear, reflecting the intrinsic complexity of the problem. Building on these findings and in order to refine the analysis, the same simulation was then performed but with a significance level of 0.2, closer to the cut-off point previously calculated.

### 4.3.2 Ratio of anomalous rows of 10% with significance level of 0.2

The parameters used in this simulation are the same as the previous one except for the level of significance, $\alpha$:

- $m = 1000$ (quantity of columns),
- $n = 5000$ (quantity of rows),
- $t_B = 0.05 + 0.10 \times k$, with $k = 0, 1, \cdots, 9$ (ratio of anomalous rows),
- $t_m = 0.1$ (ratio of anomalies per anomalous row),
- $\alpha = 0.2$ (significance level).

Figure 4.24 shows the results for the chi-square test.

The precision starts very low (0.122 in 5% of rows with anomalies) and grows almost linearly, reaching 0.98 in 95% of anomalous rows. In the previous scenario, precision performance is significantly better in almost all proportions of anomalies, being more stable and reaching high levels very early.

Regarding the *recall*, it remains relatively stable around 0.522 and 0.552 throughout the interval, without significant growth. In this scenario the *recall* is slightly higher than the *recall* obtained in the previous simulation, although the difference is not very large.

F1-*score* grows consistently, starting at 0.2 and reaching about 0.69 in the end, slightly higher than the previous simulation (0.65). Despite ending with slightly higher F1-*score* at the end of the interval, in the previous simulation it presents better balance in lower proportions.

Figure 4.25 shows the results for DMA, verifying an identical behavior to that verified for the chi-square.

As can be seen, the precision starts very low (0.12 in 5% of anomalous rows) but grows continuously, reaching 0.98 in 95%. As for the chi-square, precision, with a significance level of 0.05, behaves clearly better in the whole range, reaching high values much earlier.

If we look at the *recall*, we see that it was beneficial to increase the level of significance. This remains relatively stable around 0.517-0.556, with little variation over the whole interval. In Figure 4.17 the *recall* starts lower (0.396 in 0.05) and remains almost constant between 0.346-0.396 throughout the interval. With a significance level of 0.2, the *recall* is consistently higher, that is, it detects a larger fraction of the cases with anomalies.

With $\alpha = 0.2$, F1-*score* grows gradually from 0.194 to about 0.677, while with $\alpha = 0.05$, the F1-*score* starts at a higher value (0.326), grows up to 0.531 in 85% of rows with anomalies and ends at 0.528. In this simulation we obtained a higher F1-*score* at the end (greater balance between precision and *recall*), whereas in the previous scenario it is better at the beginning, but stabilizes at a lower level.

Figure 4.26 illustrates how KS test perform with $\alpha = 0.2$.

The precision starts low, 0.161, but grows continuously up to 0.987. However, with a significance level, $\alpha$, of 0.05 already starts much higher, 0.359, grows quickly and exceeds 0.90 in the proportion of 45% of anomalous rows, reaching 0.994. Precision results with $\alpha = 0.05$ are more promising in almost the entire range, reaching high values earlier.

Regarding the *recall*, the results are similar to those already analyzed in previous methods. For $\alpha = 0.2$, the *recall* remains stable and high (between 0.704 and 0.736) throughout the interval, while for $\alpha = 0.05$, despite remaining stable, it is lower (between 0.535 and 0.574). Therefore, for $\alpha = 0.2$, the *recall* is consistently higher, detecting a larger fraction of the anomalous cases.

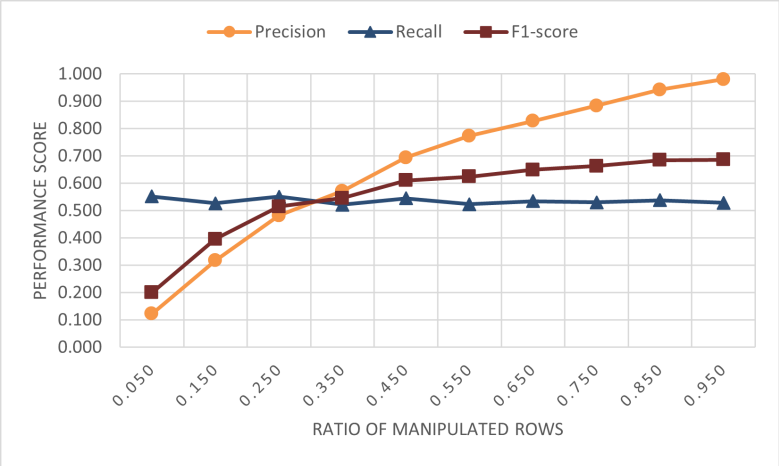


**Figure 4.24:** *Chi-square: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

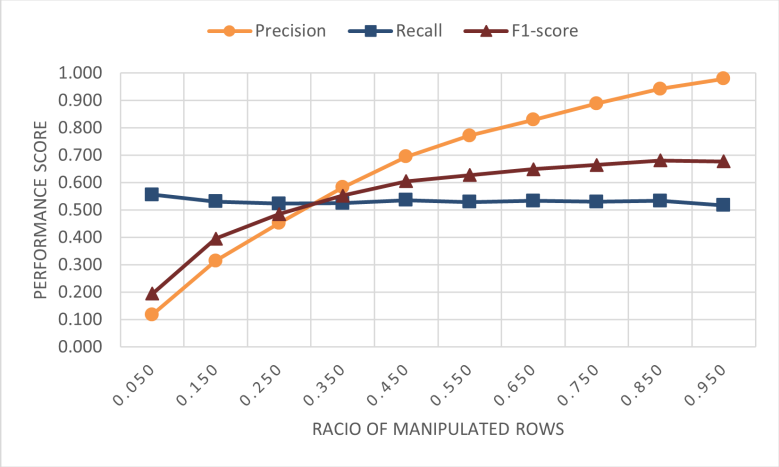


**Figure 4.25:** *MAD: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

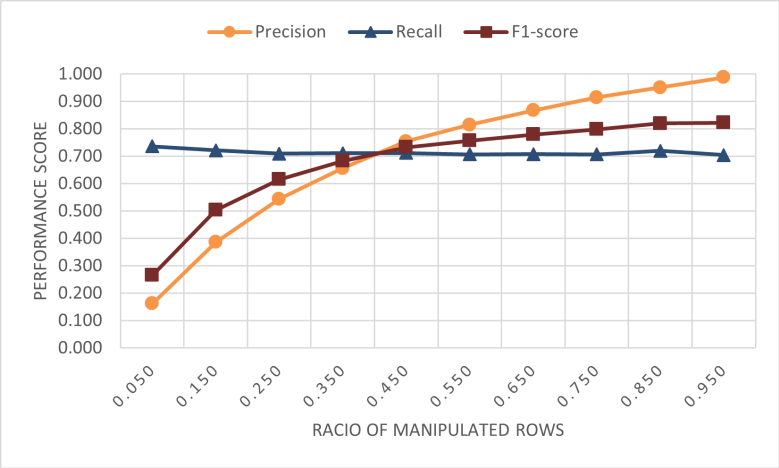


**Figure 4.26:** *KS: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

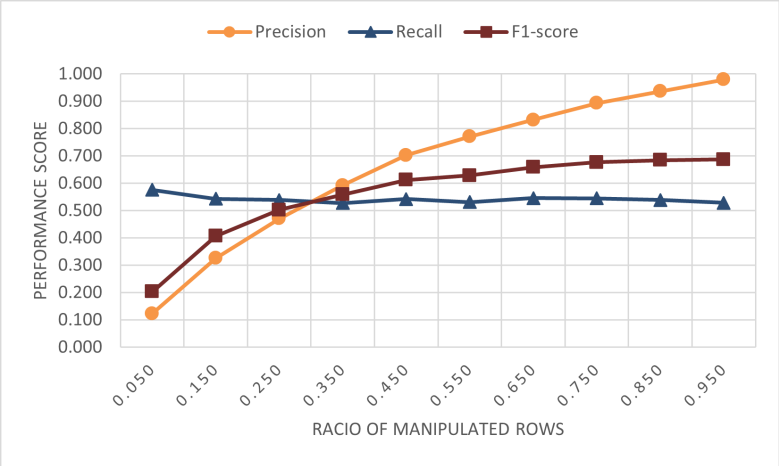


**Figure 4.27:** *Euclidean distance: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

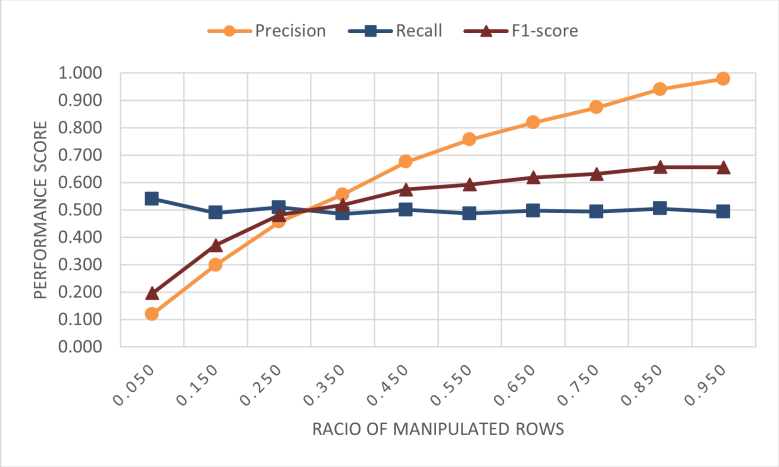


**Figure 4.28:** *Hellinger distance: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

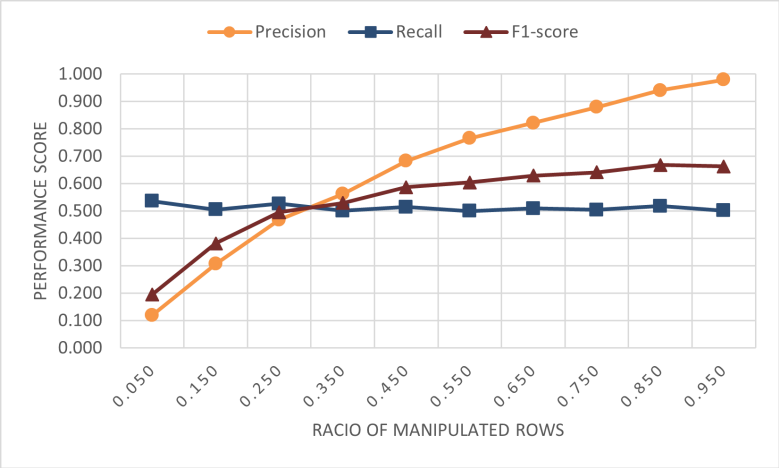


**Figure 4.29:** *KL: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 10%, significance level of 0.2*

F1-*score* evolves from 0.265 to 0.822, growing continuously and maintaining good balance. If we compare with Figure 4.18, the F1-*score* starts better, at 0.434, and grows up to about 0.716 in 85% of anomalous rows, but stabilizes without reaching the maximum values that can be obtained with $\alpha = 0.2$,. With this significance level value F1-*score* ends up overcoming the previous scenario in high levels of anomalies, reflecting better balance between precision and *recall*.

Now, analyzing the results with the Euclidean distance, Figure 4.27, we can conclude that the results do not differ much from the previous methods.

The precision starts very low, 0.123, but grows steadily up to 0.98. But again with $\alpha = 0.05$ the precision is clearly superior in the whole range, reaching high values earlier.

Analyzing the *recall*, it remains relatively stable around 0.527-0.576, with small oscillations. A much better performance compared to 0.343-0.371 with $\alpha = 0.05$. As for the previous methods with $\alpha = 0.2$ it is possible to identify a larger fraction of anomalies.

The F1-*score* grows continuously from 0.203 to 0.686, maintaining good evolution. However with $\alpha = 0.05$ it starts higher, at 0.299 and grows up to about 0.523, stabilizing at a lower level. As for the previous metrics, with a $\alpha$ of 0.2, the F1-*score* outperforms at higher levels of anomalies, reflecting better balance between precision and *recall*.

About Hellinger Distance we obtained similar results as shown in Figure.4.28.

The precision starts low at 0.120 and grows up to 0.978. With $\alpha = 0.05$ the values of precision already begin much higher, at 0.284, and exceed 0.90 in the proportion of 55% of anomalous rows and reaches 0.991.

The *recall* continues to have the same type of behavior as in previous measures. With $\alpha = 0.2$ it remains stable between 0.486 and 0.54, relatively high, compared to $\alpha = 0.05$ which is consistently lower (between 0.4 and 0.431).

The F1-*score*, for $\alpha = 0.2$ grows from 0.196 and reaches 0.657 in 85% of anomalous rows. With $\alpha = 0.05$ starts higher, at 0.339 and goes up to 0.586 in the ratio of 85% of anomalous rows and stabilizes at this level.

Figure 4.29 shows the results for KL divergence.

For $\alpha = 0.2$, the precision starts very low at 0.119 and grows to 0.979. For $\alpha = 0.05$ the precision already starts higher, at 0.284, and grows rapidly and exceeds 0.90 in the proportion of 55% of anomalous rows, reaching 0.979. Although in both scenarios the precision reaches 0.979, for $\alpha = 0.05$ the precision rises faster and stays at higher values.

As for *recall*, it remains stable and relatively high, between 0.50 and 0.54 over the interval, higher than that obtained for $\alpha = 0.05$ which did not exceed 0.457.

F1-*score* evolves from 0.195 to about 0.668 in the proportion of 85% of abnormal lines, growing continuously. In this scenario, as in the previous metrics, F1-*score* is higher, better balancing precision and *recall* in scenarios with more anomalies.

Therefore, generally for all metrics (except Fisher), increasing the significance level to a value closer to the ideal cut-off favors more *recall* and maintains continuous growth of the F1-*score*, which means that it detects a larger fraction of the anomalies, although

with more false positives at the beginning (low initial precision). It is a better option when the goal is to detect as many anomalies as possible (higher *recall*), however when the goal is to minimize false positives and increase reliability (higher precision) does not seem to be the best option.

Analyzing the Fisher combination for the two levels of significance, with $\alpha = 0.2$, the precision starts very low (0.109) and grows to 0.978. With $\alpha = 0.05$ starts higher (0.169) and exceeds 0.80 at 55% of anomalous lines, reaching 0.988. Thus, for $\alpha = 0.05$ it presents greater precision in the whole range, reaching high levels earlier.

With $\alpha = 0.2$ the *recall* remains stable between 0.645 and 0.684 throughout the interval. With $\alpha = 0.05$ the *recall* remains stable and higher, between 0.700 and 0.727. For the combination of Fisher, $\alpha = 0.05$ is also better in *recall*, identifying a larger fraction of the anomalous cases.

Moreover, for $\alpha = 0.2$ F1-*score* evolves from 0.187 to about 0.777, growing consistently, while for $\alpha = 0.05$, already starts higher, at 0.273, continuously grows and reaches 0.825. Therefore, $\alpha = 0.05$ also dominates in the F1-*score*, surpassing $\alpha = 0.2$ throughout the interval.

In short, unlike the other methods, where $\alpha = 0.2$ was stronger in *recall*, here the Fisher's Combination with $\alpha = 0.05$ is superior on all indicators.

This simulation was done with a ratio of $10\%$ of anomalies per row, a very low value of anomalies. Thus, a new simulation was carried out with a ratio of $40\%$ and the results are much more promising.

### 4.3.3 Ratios of anomalies per anomalous row of 40% with significance level of 0.05

In this simulation we changed $t_m$ to $0.4$. Below all the parameters used are defined:

- $m = 1000$ (quantity of columns),
- $n = 5000$ (quantity of rows),
- $t_B = 0.05 + 0.10 \times k$, with $k = 0, 1, \cdots, 9$ (ratio of anomalous rows),
- $t_m = 0.4$ (ratio of anomalies per anomalous row),
- $\alpha = 0.05$ (significance level).

In general, the model with $40\%$ of anomalies per anomalous row, presented excellent results, especially from the racio of anomalous rows of 0.25, with the metrics precision, *recall* and F1-*score* close to 1. The *recall* remained at 1 in practically all scenarios, showing that the model was highly effective in identifying all anomalous rows (minimizing false negatives). The differences between the methods were more evident in the precision and F1-*score* metrics. It happens the same as with a smaller proportion of anomalies per anomalous row, constant *recall* and increasing precision, for the same reasons. However, as the anomalous cases have more anomalies, it is easier to distinguish from the negatives, therefore, the model can classify better.

The main distinction between the methods was in precision and, consequently, in F1-*score*. The Kolmogorov-Smirnov method stood out as the most consistent (see Figure 4.30). Already in moderate ratio of anomalous rows (from 0.25), it gave F1-*score* higher than 0.93, reaching 0.9979 with precision of 0.9958 in the ratio of 0.95, the best result among all methods.

The mean absolute deviation also presented good performance as we can see in Figure 4.31, very close to the Kolmogorov-Smirnov, with increasing F1-scores and equally high, reaching 0.9977 in the largest ratio. The euclidean distance showed results practically identical to the mean absolute deviation, maintaining high performance and with minimal variations, which indicates that both methods have similar behavior in this scenario.

The Hellinger method also performed very well (Figure 4.32), although slightly less than those discussed earlier. Its precision values were slightly lower in the initial ratios, which reflected in slightly lower F1-scores. For example, 0.6553 in the ratio of 0.05, improving to 0.9975 in the ratio of 0.95. The Kullback-Leibler had similar behavior to Hellinger, with a marginal difference in favor of the second one in some cases.

The chi-square test (Figure 4.33) had a performance practically identical to the Kolmogorov-Smirnov and the mean absolute deviation at the highest ratios, which suggests that these tests based on distribution differences are particularly effective for this type of task.

In contrast, the Fisher combination, which aggregates the $p$-values of previous tests, had the worst performance among the evaluated methods (Figure 4.34). Despite maintaining *recall* equal to 1 in all cases, precision was significantly lower in the lowest ratios, reflecting many false positives. For example, with 0.05 of anomalous cases, the precision was only 0.2256, resulting in an F1-*score* of 0.368, quite below the other methods. Although performance improved in the higher ratios (with F1-*score* of 0.9943 in $t_B = 0.95$), this approach was less reliable, especially in scenarios with lower degree of anomalies.

In short, the methods that best balanced precision and *recall* were the Kolmogorov-Smirnov, mean absolute deviation, Euclidean distance and chi-square, all achieving excellent F1-scores above 0.99 at higher ratios. Fisher's combination, on the other hand, although conceptually promising, resulted in greater instability and performance decline, especially in the most challenging scenarios. In fact, it tends to classify more cases as anomalous than other methods and, as such, can be an advantage when the other methods present many false negatives, but in this scenario where all methods easily identify anomalous cases (few false negatives and, therefore, high recall) the Fisher method tends to create more false negatives (low precision).

Therefore, the classification model is robust and highly sensitive in the detection of anomalous rows, and methods based on distributions (such as Kolmogorov-Smirnov) provide a better balance between *recall* and precision. The use of a combination via Fisher did not bring significant gains and, in some cases, impaired the performance of the classifier.

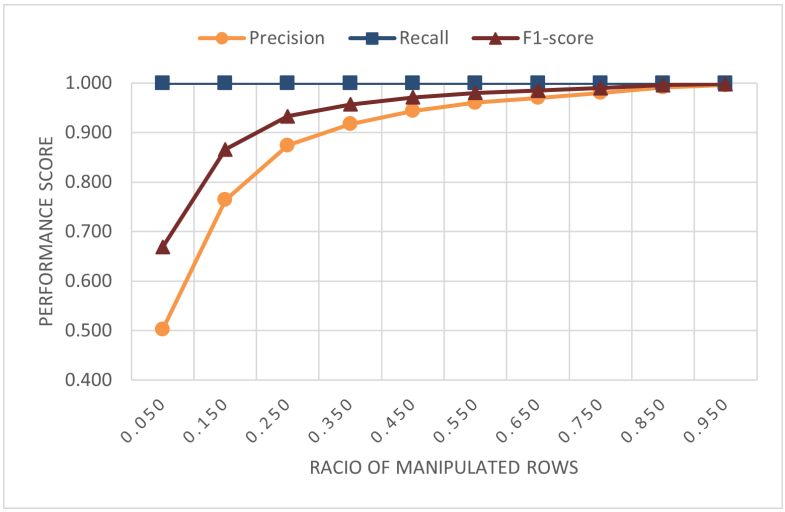


**Figure 4.30:** *KS: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 40%, significance level of 0.05*

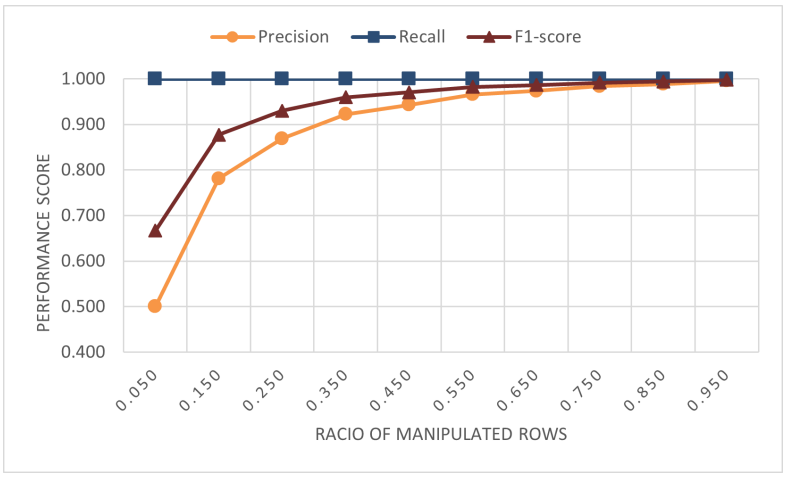


**Figure 4.31:** *MAD: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 40%, significance level of 0.05*

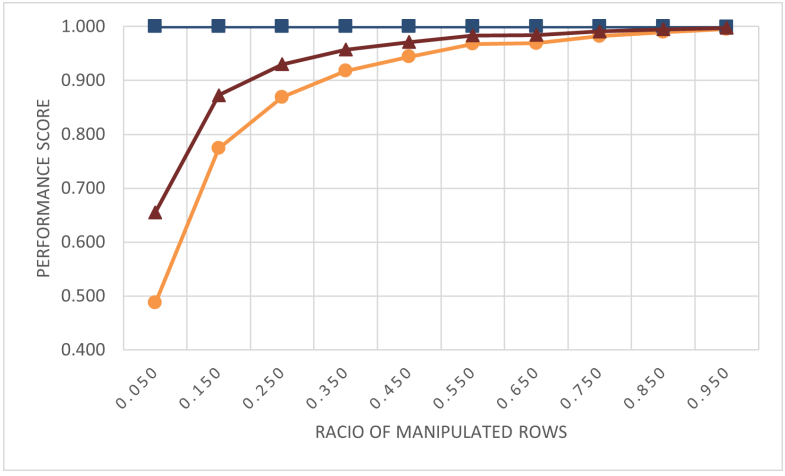


**Figure 4.32:** *Hellinger diatance: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 40%, significance level of 0.05*

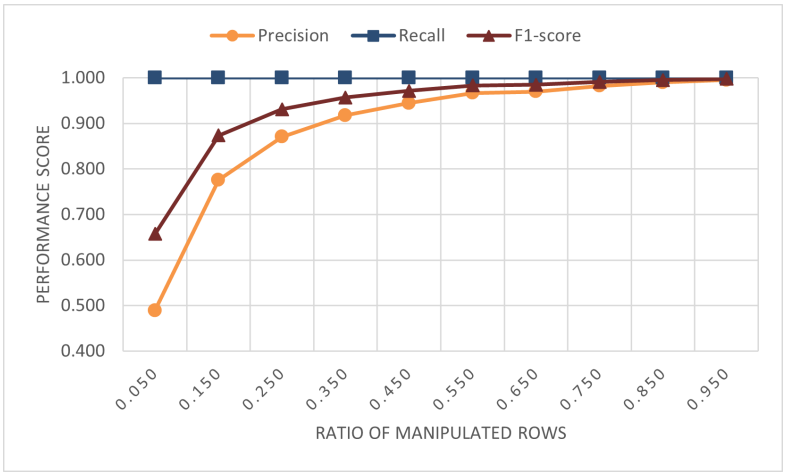


**Figure 4.33:** *Chi-square: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 40%, significance level of 0.05*

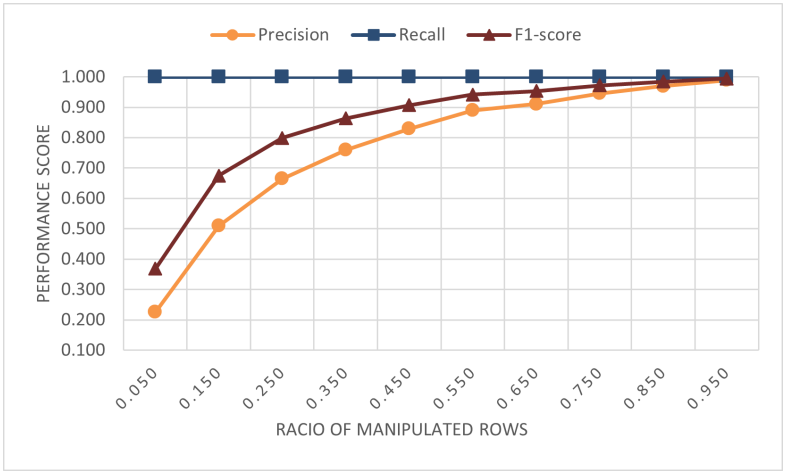


**Figure 4.34:** *Fisher: Sensitivity analysis in relation to the number of anomalous cases, ratio of anomalies per anomalous row of 40%, significance level of 0.05*

### 4.3.4 Ratios of anomalies per anomalous row of 20% with significance level of 0.05

In this simulation we tested an intermediate $t_m$ of $0.2$. The other parameters used were:

- $m = 1000$ (quantity of columns),
- $n = 5000$ (quantity of rows),
- $t_B = 0.05 + 0.10 \times k$, with $k = 0, 1, \cdots, 9$ (ratio of anomalous rows),
- $t_m = 0.2$ (ratio of anomalies per anomalous row),
- $\alpha = 0.05$ (significance level).

Starting by analyzing the *recall*, it was high from the beginning ($\approx 0.96$) and remains more or less constant. It was slightly lower than obtained with a ratio of 0.4 anomalies per row. It should be noted that, although the ratio of 0.2 is closer to 0.1, the results are closer to the ratio of 0.4 anomalies per row. Apparently, *recall* results improve significantly when the anomalies per row increases from 0.1 to 0.2.

Figure 4.35 shows the results of the KS test for the 3 ratios of anomalies per row used in the simulations.

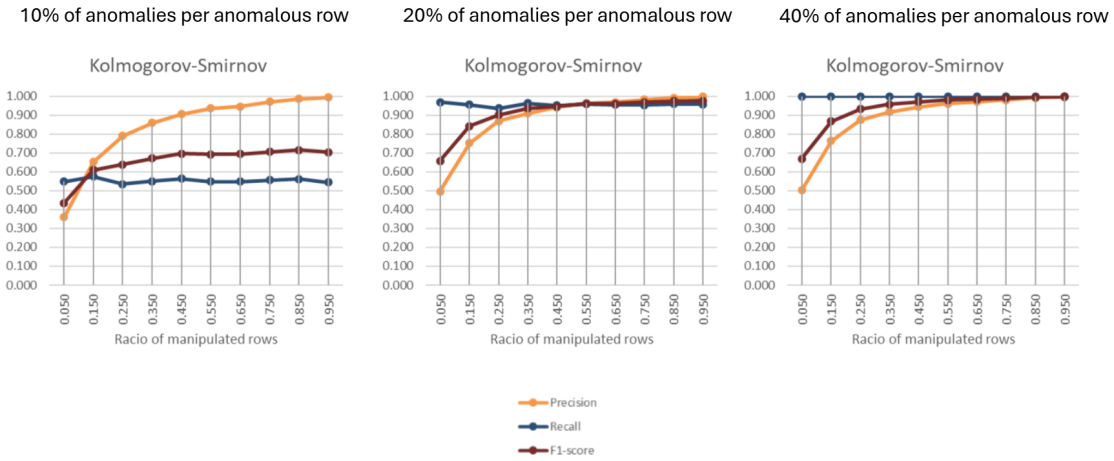


**Figure 4.35:** *Comparison of results for the KS test with 3 different proportions of anomalies per row*

With 10% of anomalies per row, the precision starts at 0.35 and ends at 1, with 20% and 40% begins at 0.5 and also reaches 1. Therefore the precision behavior is identical in the 3 proportions only starting at a much lower value for 10% of anomalies per row.

The results obtained for the other tests were identical to those obtained for the Kolmogorov-Smirnov. Once again, it is concluded that the positive increase in the $t_m$ ratio significantly improves the performance of the model. In fact, as expected, the higher the value of the ratio $t_m$, the greater the difference between the empirical distribution and Benford's law will tend to be and, therefore, the higher the values observed in dissimilarity measurements.

## 4.4 Sensitivity analysis in relation to the ratio of anomalies per anomalous case

### 4.4.1 Significance level of 0.05

For this analysis was used:

- $m = 1000$ (quantity of features)
- $n = 2000$ (quantity of cases)
- $t_B = 0.3$ (ratio of anomalous rows),
- $t_m$ in range $(0.05, 0.95)$ with steps of 0.05 (ratio of anomalies per anomalous row/-cases), i.e., $t_m = 0.05 \times k$, with $k = 1, \cdots, 19$.,
- $\alpha = 0.05$ (significance level).

The overall performance of the anomalies detection model improves progressively with increasing rate of changes per anomalous row. In scenarios with low anomalies, the model presents difficulties in identifying anomalous patterns consistently, resulting in low *recall* values and, consequently, low F1-*score* values. However, as the

anomalies becomes more pronounced, the model quickly achieves high performance, with close *recall* of 1.00 and robust F1-scores, while also maintaining satisfactory precision. The behavior of the cut-off points suggests an increasing *recall* of the model in more evident fraud contexts, which indicates good adaptability to different levels of data irregularity.

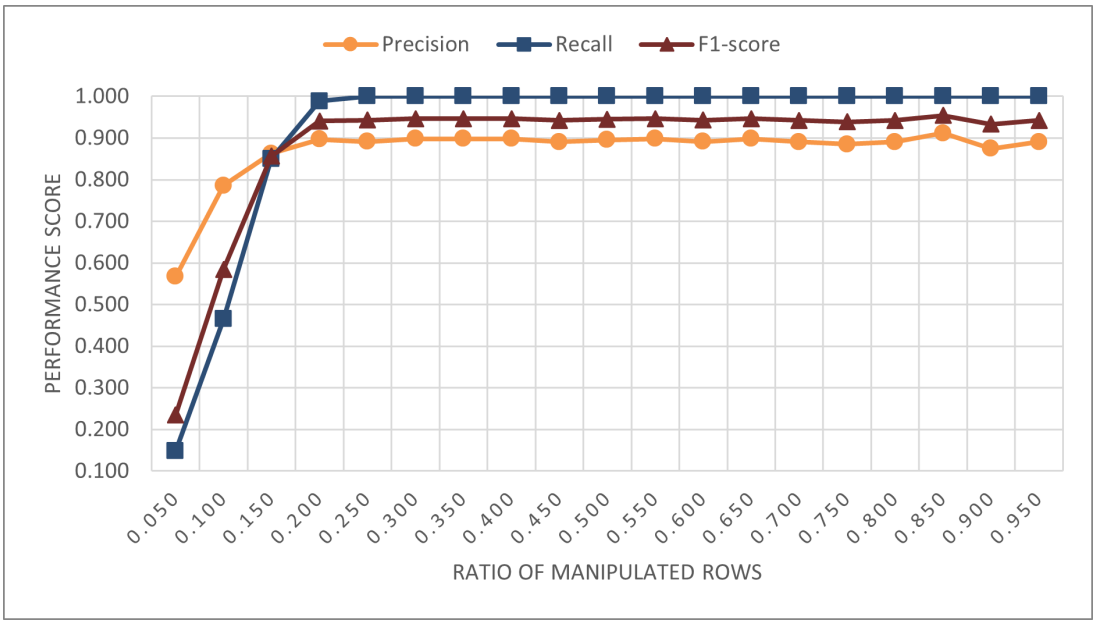


**Figure 4.36:** *Chi-square: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

Figure 4.36 reveals the model behavior with $\chi^2$. It has difficulty detecting anomalies when only $5\%$ or $10\%$ of the values in each anomalous row are changed. The *recall* is very low at $0.05$ anomalies per row, which indicates that the model can hardly identify positive cases. Precision is also poor, suggesting that even among the few cases flagged as abnormal, many are false positives.

When the proportion of anomalies per anomalous row is between $0.15$ and $0.20$, there is a sharp change in performance. From $0.15$, the model achieves an F1-*score* of $0.85$ and maintains excellent precision and *recall*. At $0.20$, the model already reaches $98.8\%$ *recall* and $89.7\%$ precision, showing itself to be much more reliable.

With irregularities up to $0.25$ the performance is consistently excellent. The *recall* is $1.00$ in almost all cases, the precision fluctuates between $87\%$ and $91\%$, which is very solid, and the F1-*score* stabilizes around $0.94/0.95$. The model becomes extremely robust when there are many changes within the anomalous rows.

Therefore, the chi-square model is sensitive to the intensity of anomalies, ineffective for subtle changes and failing to detect most anomalies (F1-*score* low at $0.05$ and $0.10$). It is highly effective from $15\%$ anomalies per row, making it almost perfect with $20\%$. The ideal cut by Youden's criterion quickly converges to a very low value $(0.001)$, reflecting that any small deviation is sufficient to indicate anomaly when the proportion of anomalies is high.

The behavior of the model with MAD (Figure 4.37) is very identical to the behavior with $\chi^2$. When the anomalies rate is $0.05$, the observed F1-*score* is only $0.20$, accompanied by a very reduced *recall* $(0.13)$, which highlights the considerable difficulty of the model in detecting anomalies when only a limited number of changes are made per row. By increasing the ratio to $0.10$, the F1-*score* rises to $0.45$, suggesting a progressively higher *recall*. At a rate of $0.15$, there is a substantial advance, with the F1-*score* reaching $0.78$ and a satisfactory balance between *recall* and precision, indicating that the model starts to operate more robustly from the moment we have at least $15\%$ of changes per row with anomalies.

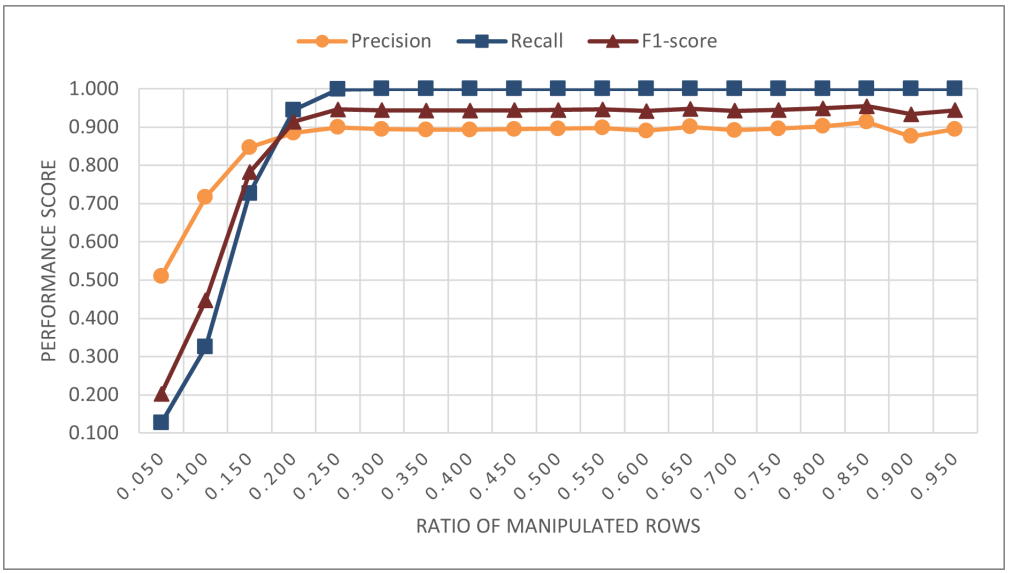


**Figure 4.37:** *MAD: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

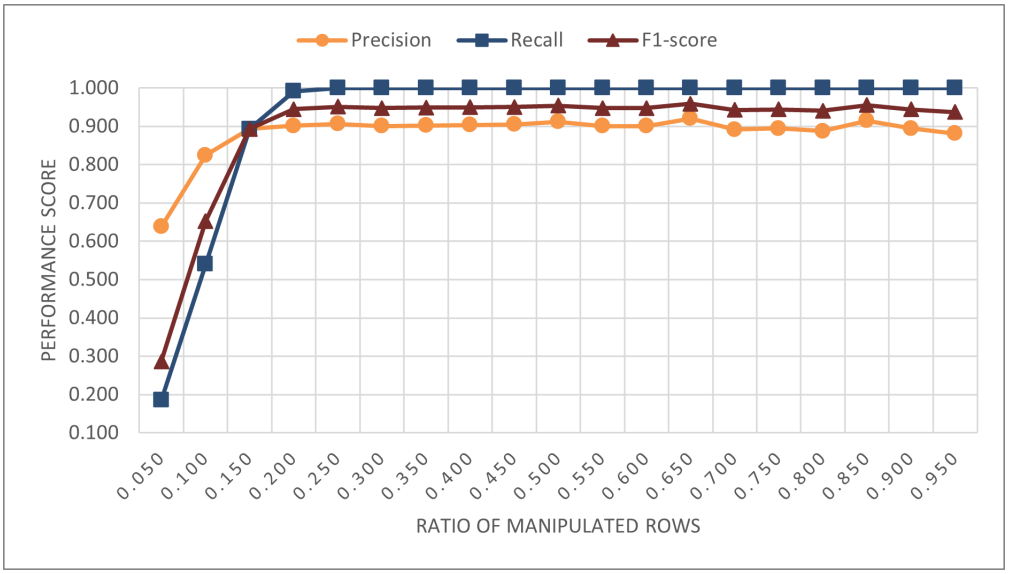


**Figure 4.38:** *KS: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

In the scenario where the change ratio reaches $0.20$, the F1-*score* exceeds $0.91$, with *recall* of $0.945$ and precision of $0.88$. For ratios between $0.25$ and $0.30$, the model achieves *recall* close to $100\%$, with precision around $0.90$ and F1-*score* ranging from $0.94$ to $0.95$. In this range, the performance of the model reaches a high level, being able to identify practically all anomalous rows with a reduced number of false positives.

Starting at $35\%$ of anomalies per row, the results tend to stabilize: *recall* remains at $1.0$, while precision remains between $0.89$ and $0.91$, with F1-*score* ranging between $0.94$ and $0.95$. At these levels, the model demonstrates high reliability to identify significant anomalies, even in the face of more expressive ratios of changes.

In general, there is a rapid evolution in the performance of the model as the anomalies rate per row increases. From $20\%$ of changes, there is already a high level of performance, while ratios equal to or greater than $35\%$ give the model robustness and better consistency.

With Kolmogorov-Smirnov test, when the ratio of anomalies is low, between $0.05$ and $0.15$, the model initially shows a modest performance (see Figure 4.38). With $5\%$ of anomalies per row, the precision reaches a reasonable value of $0.64$, but the *recall* is very low $(0.185)$, showing difficulties in capturing most anomalies, which results in F1-*score* of only 0.29. When the ratio rises to $10\%$, a significant improvement is observed, with *recall* of $0.54$ and F1-*score* of $0.65$. Already with $15\%$ of anomalies, the model achieves a balanced and optimal performance, with precision and *recall* both at $0.893$ and F1-*score* of $0.89$. It is concluded, therefore, that from $15\%$ of cells anomalous per row, the model begins to identify anomalies with good *recall* and specificity.

In the anomalies range between $0.20$ and $0.30$, the model shows a virtually perfect *recall* $(0.99)$ and F1-*score* of $0.94$ with $20\%$ of anomalies per row, maintaining an excellent balance between metrics. When the ratio increases to $25\%$ and $30\%$, performance stabilizes at very high levels, with precision around $0.90$, *recall* of 1 and F1-scores between 0.948 and 0.95. Thus, the model shows a better performance to identify anomalies when we have at least $20\%$ of anomalies per anomalous row.

With high anomalies ratios, equal to or greater than 0.35, the performance of the model remains consistent and robust. The *recall* remains 1.0, while the precision varies between 0.88 and 0.92, remaining high. The F1-*score* ranges between 0.94 and 0.96, signaling excellent performance. This shows that, in scenarios with many irregularities, the KS test offers stability, reliability and precision, being able to identify all anomalous rows without significant loss of precision.

In general, it is observed that the performance of the model improves rapidly as the ratio of anomalies per row increases. From $15\%$ of anomalous cells, the KS-based model already delivers robust results, and with $20\%$ or more anomalies, it achieves optimal performance with perfect *recall* and precision above $0.90$. The KS test therefore proves to be a highly sensitive and effective method for detecting anomalies as anomalies density grows.

For Euclidean distance (Figure 4.39), as other methods above, when the anomalies ratio is low, between $0.05$ and $0.15$, the model initially shows a rather limited perfor-

mance. With only 5% of anomalies per row, the F1-*score* is 0.20, and the *recall* is very low (0.12), which means that the model fails to detect most of the anomalous rows. When increasing to 10%, some improvement is observed, with F1-*score* of 0.46 and *recall* of 0.33. When the ratio of anomalies per row reaches 15%, the model finally achieves an acceptable performance, with F1-*score* of 0.80, precision of 0.86 and *recall* of 0.75. Thus, it is concluded that the effectiveness of the model only begins to be reliable from 15% of anomalies per row.

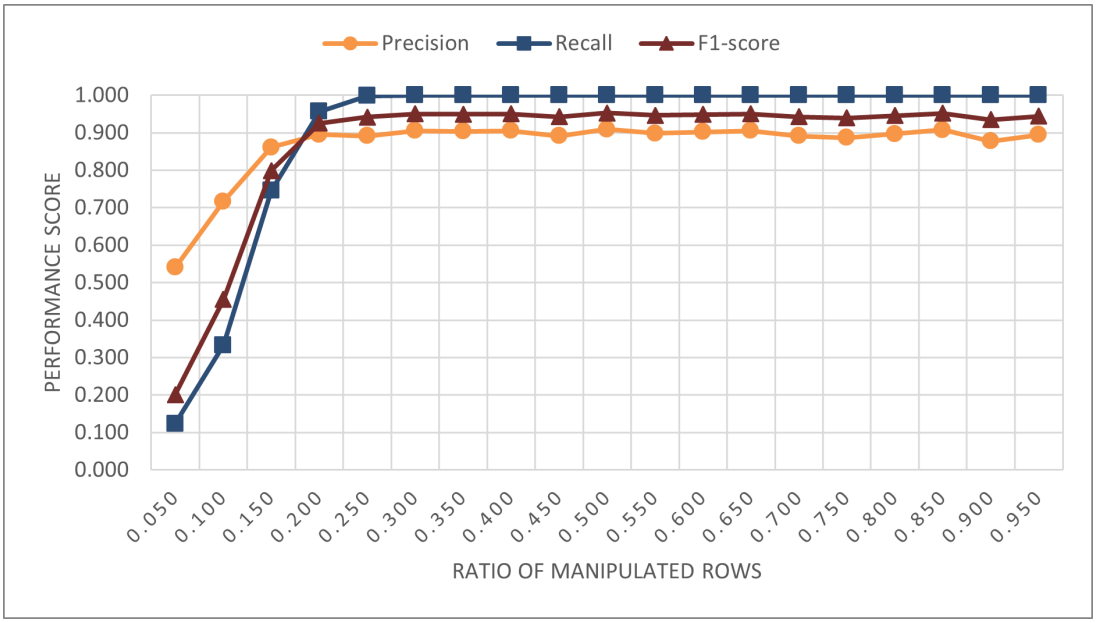


**Figure 4.39:** *Euclidean distance: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

In the moderate anomalies range, between 0.20 and 0.30, model performance becomes optimal. With 20% of anomalies, it reaches F1-*score* of 0.93 and *recall* of 0.96. At 25%, the performance is almost perfect, with F1-*score* of 0.94 and virtually total *recall* (0.998). When it reaches 30%, the model reaches its maximum performance with perfect *recall* (1.00) and F1-*score* of 0.95. Therefore, between 20% and 30%, the model has high *recall* and a very good precision.

With high anomalies ratios, equal to or greater than 35%, the model remains consistent and robust. The *recall* remains 1.0 in all situations, while the precision varies between 0.88 and 0.91, and the F1-*score* is between 0.93 and 0.95. This indicates that, with 35% or more of anomalies, the model detects all anomalies while maintaining a low number of false positives.

In general, the Euclidean distance model shows low *recall* in scenarios with weak anomalies. From 15%, it begins to display a relevant performance, and from 20%, it becomes highly effective, reaching F1-scores higher than 0.92, perfect *recall* and solid precision. The performance stabilizes at optimal levels for scenarios with more anomalies.

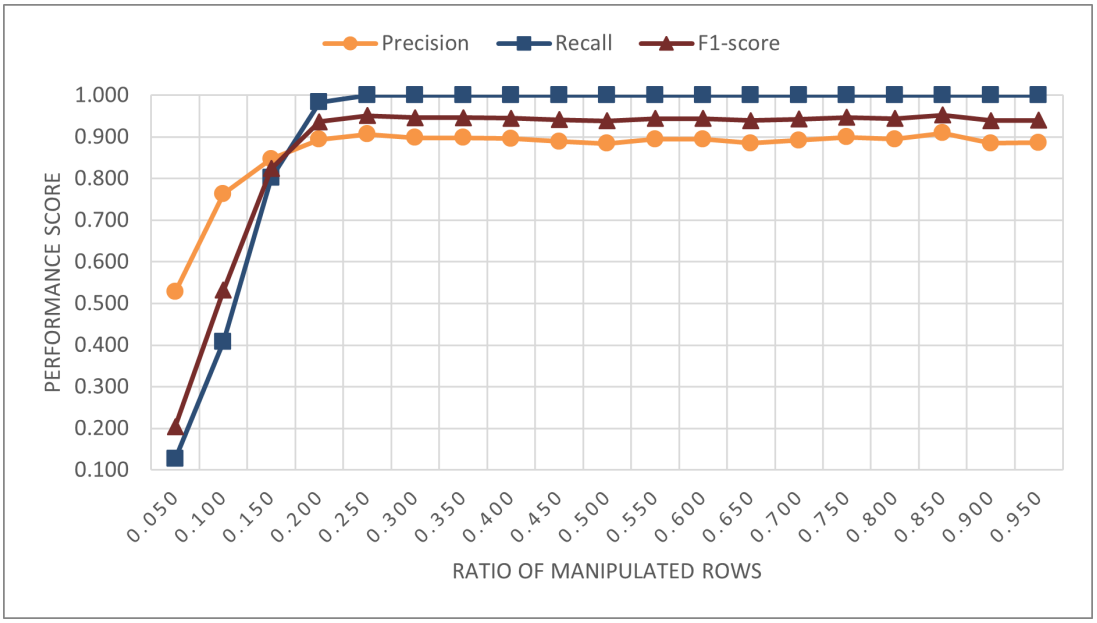


**Figure 4.40:** *Hellinger distance: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

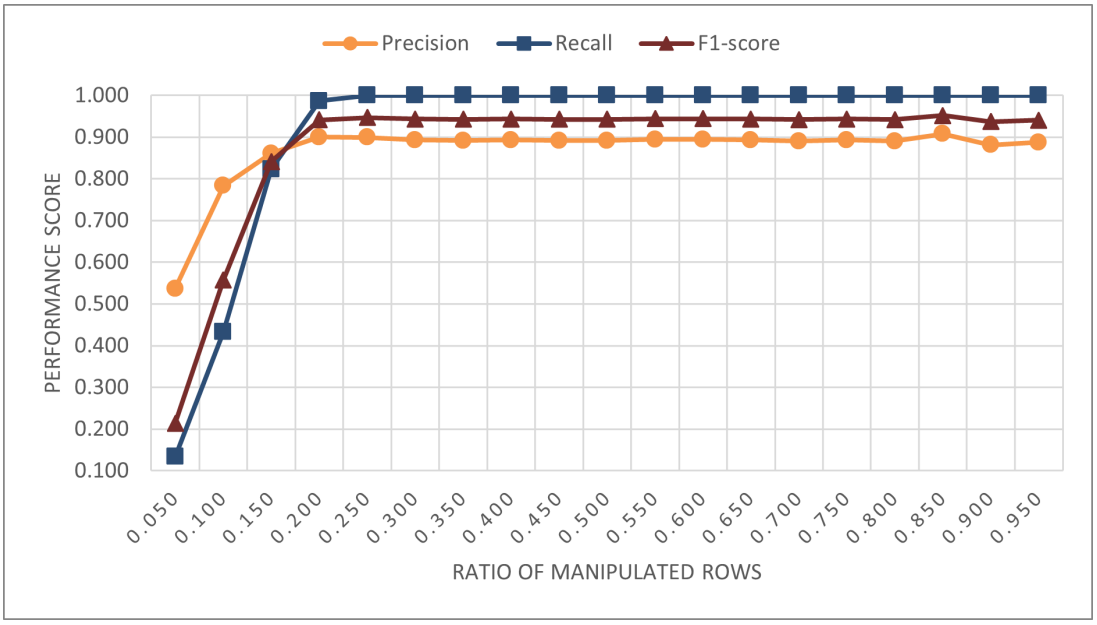


**Figure 4.41:** *KL: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

About Hellinger distance, the anomalies ratios between 0.05 and 0.15 show the following results: with 0.05, the performance is weak, with a F1-*score* of 0.20 and *recall* of 0.13 (see Figure 4.40). At 0.10, there is some gain, with F1-*score* of 0.53, showing an improvement in *recall* and precision. Already with 0.15, a good balance is observed, with F1-*score* of 0.82, precision of 0.85 and *recall* of 0.80. As in the Euclidean distance, performance only becomes robust from 15%.

In terms of anomalies between 0.20 and 0.30, with 0.20 the model achieves an ex-

cellent performance, with F1 of 0.94 and *recall* of 0.98. With 0.25, it almost reaches perfection, with F1 of 0.95 and *recall* of 1. With 0.30, it remains high, with F1 of 0.95 and *recall* of 1. From 20%, the model has excellent *recall* and good precision.

When the anomalies is above 0.35, the model maintains *recall* equal to 1.00, precision ranging between 0.88 and 0.91 and F1-*score* between 0.94 and 0.95 up to 0.95 of anomalies. The behavior of the model is very consistent with high anomalies per anomalous row ratios.

With the Kullback-Leibler divergence (Figure 4.41), anomalies ratios of 0.05, performance is very weak, with a F1-*score* of 0.21 and *recall* of 0.13. With 0.10, an improvement is observed, reaching a F1-*score* of 0.56 and precision of 0.78. Already at 0.15, the performance is very good, with a F1-*score* of 0.84 and a consistent balance between precision and *recall*. This behavior is similar to that observed with the distances of Hellinger and Euclidean.

In the anomalies range between 0.20 and 0.30, the results are even better, with 0.20, the F1-*score* reaches 0.94 and the *recall* is excellent, reaching 0.99. With 0.25, the F1-*score* rises to 0.95, with *recall* of 1.00, and with 0.30 remains at high levels, with F1-*score* of 0.94 and *recall* equal to 1.00. As in the other criteria, performance stabilizes from this range, with F1-*score* of 0.94.

With a high rate of anomalies per anomalous row, equal or greater than 0.35, the model maintains a very stable F1-*score*, ranging between 0.94 and 0.95, with *recall* always equal to 1 and precision ranging from 0.88 to 0.91. This pattern repeats what was observed in the other distances, perfect *recall* accompanied by a stable and high precision.

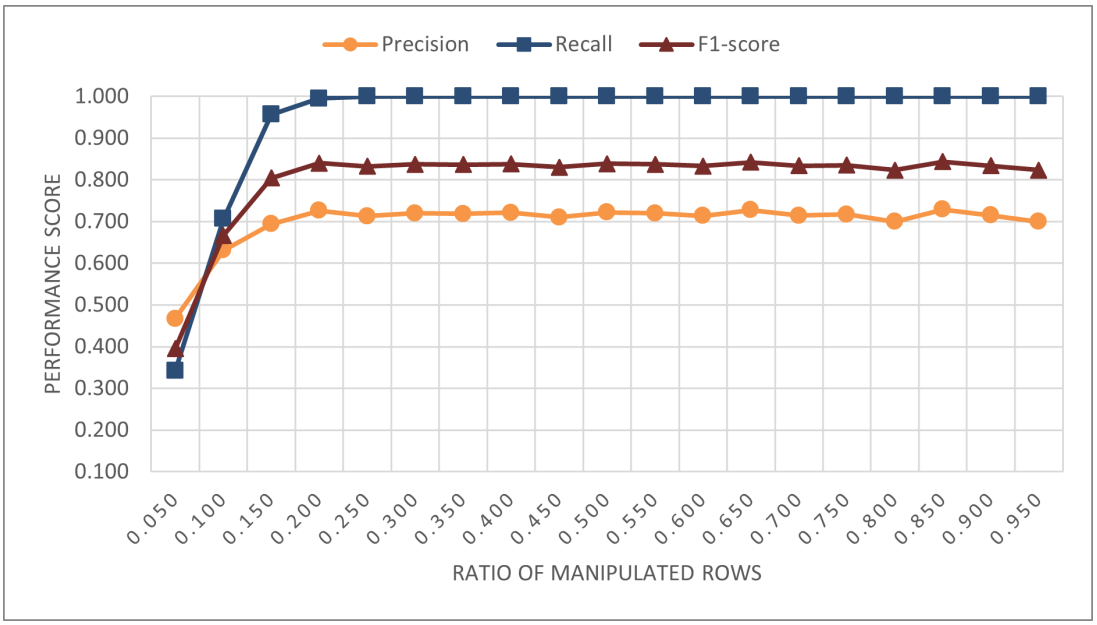


**Figure 4.42:** *Fisher: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

In the analysis of $p$-values combination by Fisher's method, low anomalies ratios

between 0.05 and 0.15 show interesting results, with 0.05, the F1-*score* is 0.39, already surpassing the other distances in this range (see Figure 4.42). With 0.10, the performance rises to F1-*score* of 0.67, clearly standing out from other approaches. At 0.15, the F1-*score* reaches 0.80, being slightly below the values obtained with the distances KL and Hellinger. In general, the Fisher method shows the best performance in the low to moderate anomalies ranges.

| Measure | Optimal cut-off | Best precision | Best recall | Best F1-score | Best Accuracy |
|---|---|---|---|---|---|
| Chi-square | 0.001 | 0.912 | 1.0 | 0.954 | 0.971 |
| Mean Absolute Deviation (MAD) | 0.001 | 0.913 | 1.0 | 0.955 | 0.972 |
| Kolmogorov-Smirnov | 0.001 | 0.920 | 1.0 | 0.958 | 0.974 |
| Euclidean | 0.001 | 0.909 | 1.0 | 0.952 | 0.970 |
| Hellinger | 0.001 | 0.909 | 1.0 | 0.952 | 0.970 |
| Kullback-Leibler | 0.002 | 0.908 | 1.0 | 0.952 | 0.970 |
| Fisher | 0.001 | 0.729 | 1.0 | 0.843 | 0.889 |

**Table 4.6:** *Comparative table of the different methods as the ratio of anomalies per anomalous case increases from 0.25 to 0.95 (30% of anomalous rows, significance level of 0.05)*

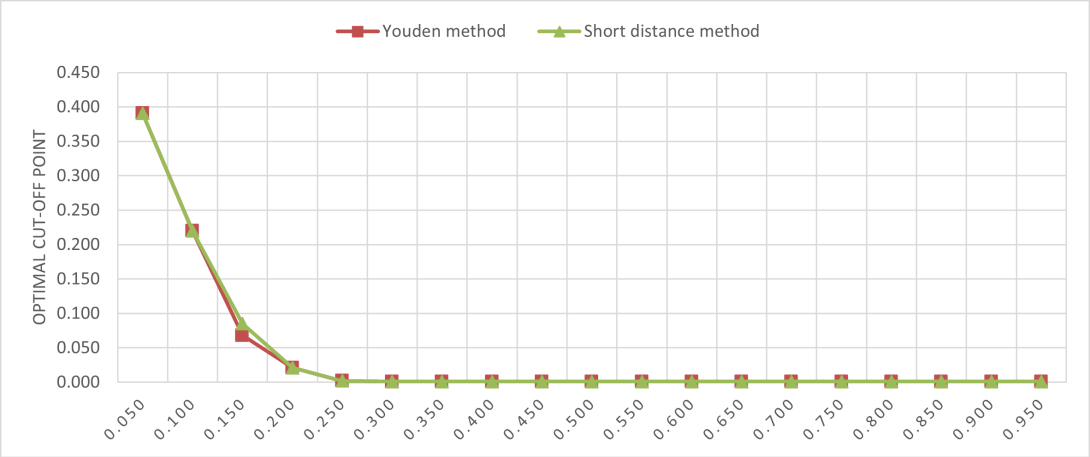


**Figure 4.43:** *Optimal cut-off calculation: Sensitivity analysis in relation to the ratio of anomalies per anomalous case, ratio of anomalous rows of 30%, significance level of 0.05*

In the anomalies between 0.20 and 0.30, the method achieves F1-*score* of 0.84 in 0.20, performance comparable to that of the distance KL, although slightly lower. Between 0.25 and 0.30, the F1-*score* stabilizes around 0.83 to 0.84. Although good, this performance is below that observed with Hellinger, KL and Euclidean, which reach about 0.95 in this range.

Already in the ratios of anomalies equal to or greater than 0.35, the F1-*score* remains practically constant, varying between 0.83 and 0.84, with *recall* always equal to 1. However, the precision is between 0.71 and 0.73, which is considerably lower compared to other distances. Thus, the Fisher method presents the worst performance when the rate of changes per row anomalous is high, while the other approaches reach F1-scores around 0.94 to 0.95.

In short, at the lowest anomalies ratios, all methods show poor performance in *recall* and F1-*score*, highlighting the difficulty inherent in detecting weak anomalies signals.

In this zone, all tests reveal less *recall* (see Table 4.6). Fisher's criterion, although also with modest results in this range, shows a slightly higher capacity for equilibrium.

As the ratio of anomalies increases (above $15\%$), all methods quickly converge to high *recall* values (close to $1.00$), being the main distinction observed in precision and therefore in the F1-*score*. Hellinger and KL distances, as well as the Euclidean, stand out for their high precision, maintaining values above $0.89$ for ratios greater than $25\%$, which leads to F1-scores greater than $0.94$. The Kolmogorov-Smirnov tests and mean absolute deviation closely follow this performance, with slight losses in precision at certain points.

Fisher's criterion, by combining the $p$-values of the six tests, works as a softener for inconsistent signals and as an amplifier of evidence when there is agreement between the criteria. Its performance becomes notoriously stable and competitive from $15\%$, with F1-scores between $0.83$ and $0.84$, slightly below the distances, but with an impressive consistency. The advantage of Fisher's criterion lies in its aggregator robustness, it is less sensitive to statistical noise from individual tests and more reliable in scenarios where anomalies is not uniformly detectable by all metrics.

Regarding the cut-off points, there is a clear convergence to extremely low values ($\approx 0.001$) in almost all methods from $25\%$ of anomalies (see Figure 4.43), indicating that the tests become highly sensitive to the presence of anomalies. Distance-based methods reach these cuts earlier, reflecting their greater reactivity to structural deviations.

### 4.4.2 Significance level of 0.001

The same sensitivity analysis was performed with a significance level of 0.001 as suggested by the Youden and minimum distance methods, revealing a high reliability profile, but with limitations in initial detection. As can be seen in the Figure 4.44, for the case of the Euclidean distance, for ratios greater than 25%, the model guarantees consistent precision up to 0.996, maximum *recall* to converge to 1.00, and excellent F1-*score* (very close to 1.00). This high reliability makes it the ideal method for contexts where false positives are inadmissible, such as forensic investigations that require high certainty or automated systems with low error tolerance. However, the limitations are evident, with low initial sensitivity, where for ratios below 15% performance is significantly limited, being effective only when there are substantial irregularities greater than 25%.

This behavior was identical in all methods, according to Fisher (see Figure 4.45). For this method the performance also started with lower values of precision and F1-*score* and improved significantly from 25% of irregularities. However, the precision did not exceed 0.900 and the F1-*score* did not exceed 0.948.

In conclusion, the model behaves better with levels of significance $\alpha = 0.001$ due to its stability at high ratios of irregularities, while for $\alpha = 0.05$ it requires additional care in the initial phase due to observed volatility. This characterization is fundamental for the informed choice of significance level according to the specific objectives of

the analysis and the tolerance to the risk of false positives of the detection system implemented.

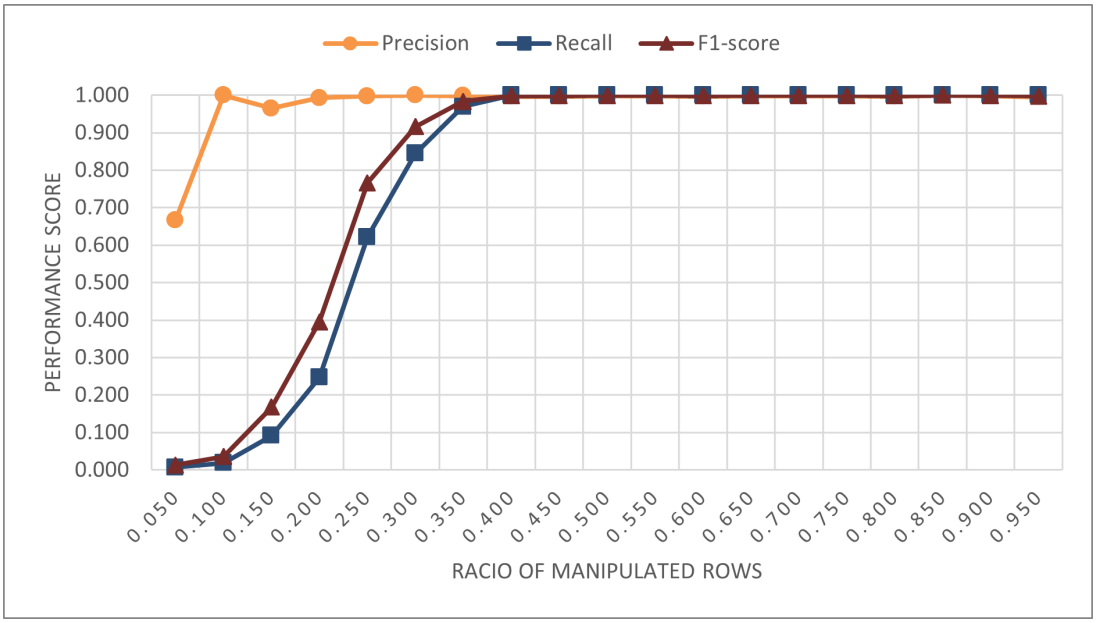


**Figure 4.44:** *Sensitivity analysis in relation to the ratio of anomalies per anomalous case (Euclidean distance with significance level of 0.001)*

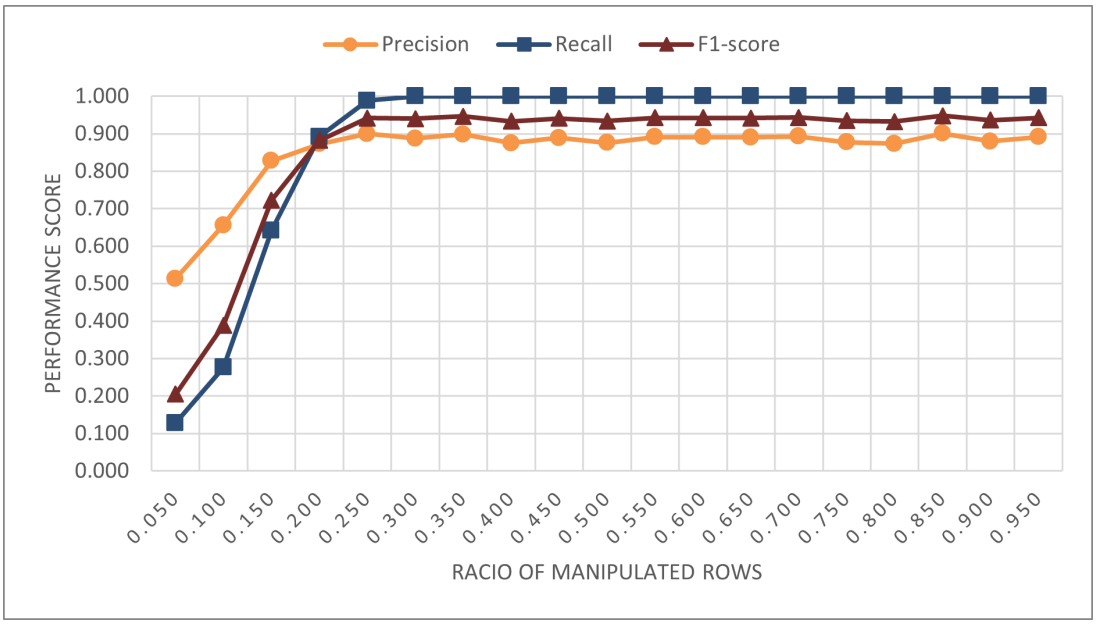


**Figure 4.45:** *Sensitivity analysis in relation to the ratio of anomalies per anomalous case (Fisher metod with significance level of 0.001)*

## 4.5 Discussion of the results

The sensitivity analysis allowed to identify the main factors that impact the effectiveness of statistical methods based on Benford's Law for the detection of anomalies. The

results show that the performance of these methods is strongly influenced by three fundamental variables: sample size, proportion of anomalous cases and data dimensionality.

Regarding the impact of sample size, there was a consistent improvement in the performance of all methods with the increase in the number of rows and columns. This trend becomes particularly evident from 600 observations, where the precision and F1-*score* metrics stabilize significantly. The chi-square test ($\chi^2$), traditionally employed in this type of analysis, revealed an adequate balance between precision and *recall* for samples of moderate or high consolidating itself as a safe option when trying to minimize both false positives and maximize correct detections. On the other hand, the mean absolute deviation, Hellinger distance and Kullback-Leibler divergence tests maintained a stable F1-*score* throughout the range of observations tested. In contrast, the Kolmogorov-Smirnov test and the Euclidean distance were more sensitive to sample size, showing high variability in precision, especially with less than 600 observations.

The dimensionality analysis revealed that increasing the number of columns substantially improves the performance of the different tests applied. With a reduced number of columns, all methods showed poor performance, resulting in low F1-scores due to the lower amount of information per observation. From 1100 columns, the results stabilize at a high level, often with F1-scores higher than 0.94, suggesting a maturity of the model from this point on.

The proportion of anomalies in the data emerged as a critical factor in the reliability of the methods. In scenarios with 40% of abnormal cases, all tests demonstrated excellent performance, with *recall* often equal to 1 and equally high accuracy, resulting in F1-scores close to the maximum value. This situation allows a clear separation between legitimate and anomalous patterns, reducing the impact of statistical noise. When the proportion of abnormal lines was reduced to 20%, there was a slight degradation in accuracy due to the relative increase in false positives, although *recall* remained high and F1-*score* remained at satisfactory levels.

The most challenging scenario occurs when only 10% of cases or less are anomalous, a situation in which accuracy suffers a significant drop, although *recall* remains high. This degradation is due to the fact that, with a significance level of 5%, the proportion of false positives becomes very close to the proportion of true positives, compromising the reliability of detections, i.e., the significance level generates as many false positives as true positives, and therefore the test is no longer reliable. Therefore, the system loses reliability because it is not clear whether a "detection" actually represents an irregularity or just statistical noise. This phenomenon is particularly relevant in databases that contain few truly anomalous cases, justifying the adoption of more stringent decision criteria or complementary approaches.

However, the use of more restrictive levels of significance, close to the values calculated by the Youden methods and minimum distance to the point $(0, 1)$, showed a substantial improvement in the performance of the model. This approach allows to

optimize the balance between sensitivity and specificity, resulting in a significant reduction of false positives without excessively compromising detection capacity.

The method of combining $p$-values via Fisher's criterion showed a different behavior from the other methods. Although it showed lower performance in high anomaly scenarios, it stood out positively in situations with more subtle anomalies (between 5% and 15%), obtaining F1-scores higher than the other methods. This feature derives from its aggregating nature, which allows smoothing the individual statistical noise and amplifying consistent signals between different tests. However, its excessive sensitivity to any statistical deviation makes it prone to generate false positives, so it is not recommended as a main decision criterion.

In summary, the effectiveness of statistical tests based on Benford's Law is strongly conditioned by the interaction between the number of abnormal cases, the proportion of changes per anomalous line and the sample size. In contexts with high prevalence of anomalies, the methods show robust performance and can be applied with confidence. However, in scenarios with discrete anomalies or low prevalence of altered cases, the tests become more vulnerable to statistical noise, justifying the adoption of complementary approaches, such as the joint use of several tests, cross-validation or adjustment of decision thresholds. The interpretation of results should consider the balance between *recall* and precision, the impact of sample size and chosen significance level.

# 5
# Conclusions and Future Work

This work successfully achieved its main goals through the development of a controlled synthetic data generator capable of simulating different standards of compliance and non-conformity with Benford's Law, and the proposal of an innovative record processing approach that enables direct identification of anomalous instances rather than limiting verification to the global dataset level. This granularity has proved to be a practical plus, allowing the precise location of potential frauds or errors. It was also sought to evaluate the performance of different measures of statistical divergence, namely the chi-square test, mean absolute deviation, Kolmogorov-Smirnov test, Euclidean distance, Hellinger distance, Kullback-Leibler divergence and combination of $p$-values by the Fisher method, analyzing the effectiveness of the methods through precision metrics, *recall*, F1-*score* and confusion matrix. Another central objective was to investigate more efficient, fast and interpretable alternatives to traditional machine learning models, exploring statistical methods based on Benford's Law and the possibility of combining them with machine learning techniques. This integration paves the way for hybrid models, more flexible and accurate in anomaly detection. Finally, the project sought to contribute to the creation of a practical, accurate and reliable tool for auditors, accountants and researchers, with potential application in multiple domains, reinforcing the ability to detect any type of anomaly and support decision-making.

The project makes three fundamental contributions to the field. First, we highlight the development of a benchmarking tool based on the generation of synthetic data with well-defined characteristics (including Benford distributions, uniform distributions, gaussian noise, uniform noise and outliers). This tool is a valuable resource for the scientific community, allowing to test, compare and validate different methods of anomaly detection, thus promoting replicability and the advancement of new solutions in an open source format accessible to researchers (https://github/PatriciaMartinho/Anomaly-Detection-in-Numerical-Data). Second, a deviation-sensitive classification algorithm was designed to identify not only the presence of anomalies but also their exact location in the data lines, overcoming limitations from previous studies. This approach also includes the exploration of the theoretical limits of detection based on

Benford's Law, such as the minimum sample size required, and the analysis of the behavior of different dissimilarity measures in the capture of deviations. Third, the work contributes to the creation of a more accurate, scalable and efficient model, able to analyze large volumes of data in real time and to be integrated into existing audit systems. This statistical-based model, which can be combined with machine learning methods, allows it to be used in hybrid solutions, flexible and adaptable to different application areas such as financial auditing, economics or social sciences. In order to support the development of such a model, the analysis also focused on understanding the key factors that influence the effectiveness of statistical methods based on Benford's Law.

The comprehensive analysis revealed key factors that determine the effectiveness of statistical methods based on Benford's Law for anomaly detection. The results show that the performance strongly depends on the sample size, the proportion of abnormal cases, the proportion of anomalies per anomalous case (row) and the dimensionality of the data. It was found that, from about 600 observations and 1100 columns, the performance metrics stabilize at high levels, ensuring greater robustness. Among the evaluated methods, the chi-square test stood out for its consistency, while the Kolmogorov-Smirnov test and the euclidean distance were those that showed greater sensitivity to sample size. The proportion of anomalies was decisive in scenarios with high prevalence (40%), the methods presented almost perfect performance. In contexts with low prevalence (10%), reliability was reduced due to the relative increase of false positives. This limitation highlights the need to adjust decision criteria or adopt complementary approaches in databases with few anomalies. Finally, the Fisher $p$-value combination method performed better in scenarios with a low proportion of anomalies, although its high sensitivity makes it susceptible to generating false positives, limiting its use as a main criterion.

In summary, the methods based on Benford's Law are effective and reliable in contexts of high prevalence of anomalies and sufficiently large samples, but require caution and possibly hybrid strategies in scenarios with more reduced data size or low incidence of irregularities.

The results of this project have significant practical applications in several contexts that depend on the analysis of large volumes of data. In financial audits, methods based on Benford's Law allow the identification of anomalous patterns in transactions and accounting reports, helping to detect fraud or errors in near real time. In the economic sphere, governments and institutions can use these methods to validate the consistency of statistical data such as employment indicators, inflation or tax revenues, ensuring more reliable decisions. In social sciences and academic research, the application of techniques allows identifying inconsistent responses or possible manipulations in large-scale surveys, increasing the credibility of conclusions. In addition, the developed model is scalable and can be integrated into hybrid solutions with machine learning algorithms, allowing the creation of adaptive systems that continuously monitor data and generate automatic anomaly alerts. Thus, the proposed approach con-

tributes to more informed decisions based on reliable data, whether in audits, public policies, market analysis or scientific research.

Nevertheless, despite these encouraging results and applicability, it is important to acknowledge certain limitations of the present study that may affect the generalization and applicability of the findings and should be considered for future research applications.

The study employed a limited number of proportions of cases handled and used a fixed number of variables (m = 1000) and cases (n = 5000). In real contexts, the data may present different proportions or more complex structures, which can affect the generalization of results. Although the data generator allows to simulate different types of anomalies, in the study was used only uniform distribution. In addition, the abnormal cases all had a fixed proportion of anomalies. In practice, the anomalies can vary in quantity and pattern and may influence the sensitivity of the statistical tests used.

The analysis concentrated on specific statistical tests using predominantly a 0.05 significance level, and variations in this parameter could substantially influence false positive or false negative rates. While the simulations carried out provide valuable controlled environments for variable exploration, they cannot capture all complexities present in real data, including non-linear correlations, noise variations, or unexpected patterns. On the other hand, the study was conducted with moderate-sized datasets, and performance applicability may vary significantly with larger or more complex databases, an aspect not evaluated in this work.

For future work, it is recommended to perform several replications of each scenario and to explore a greater variety of data manipulation scenarios, including the use of a significance level closest to the calculated by Youden method, different proportions of altered cases and heterogeneous patterns of manipulation, in order to better reflect the diversity found in real databases.

It would also be interesting to test the behavior for the other digits, besides the first. The inclusion of additional digits could allow, for example, to detect more subtle manipulations, in which the first digit is adjusted so as to appear compliant, but the following digits reveal discrepancies. In addition, in a context of classification models, the combined use of multi-digit distributions could serve as a broad set of variables (features), potentially improving the predictive capacity and robustness of the model, although in this case we are dividing the possible outcomes into a larger number of categories, which certainly implies the need for a larger number of features to properly identify anomalous cases.

The use of subsequent digits also raises challenges, namely the increase in dimensionality, the greater proximity of expected distributions with uniformity and, consequently, the increased risk of noise introduction. Nevertheless, research on the impact of multi-digit integration on the performance of classifiers based on Benford's Law constitutes a relevant line of investigation.

It would also be important to apply complementary methods that do not depend

exclusively on Benford's Law, allowing to evaluate the effectiveness of detection in datasets that do not naturally follow this distribution. Another relevant line of research is to test different levels of significance and include a wider range of statistical tests and divergence metrics, in order to understand how these choices influence the sensitivity and specificity of methods. In addition, it is recommended to validate the results obtained in simulations by applying the tests to real databases, ensuring that the conclusions are maintained in practical contexts with noise, complex correlations and unexpected patterns. Finally, given the growing size of modern datasets, it is advisable to study the scalability and computational efficiency of methods, optimizing algorithms when necessary, to ensure that the approach remains viable in large-scale scenarios. These research directions will further strengthen the practical applicability and theoretical foundation of Benford's Law-based anomaly detection methods. Nonetheless, the preliminar simulation study conducted in this project demonstrates that anomaly detection in numerical data can be effectively achieved using Benford's Law in combination with statistical hypothesis testing. Moreover, the choice of divergence metric and cut-off point should be tailored to the specific characteristics of the data under evaluation to ensure an appropriate balance between sensitivity and specificity.

*This page intentionally left blank.*

# Bibliographical references


[1] J. A. Alvarez-Jareño and J. M. Pavia. "Using machine learning for financial fraud detection in the accounts of companies investigated for money laundering". In: *Forensic Science International* (2017). DOI: https://10.1016/j.forsciint.2017.11.008.

[2] L. H. Aros et al. "Financial fraud detection through the application ofmachine learning techniques: a literature review". In: *Humanities and Social Sciences Communications* 11.1130 (2024). DOI: https://doi.org/10.1057/s41599-024-03606-0.

[3] L. Arshadi and A. H. Jahangir. "Benford's law behavior of Internet traffic". In: *Journal of Network and Computer Applications* 40 (2014), pp. 94–205. DOI: https://doi.org/10.1016/j.jnca.2013.09.007.

[4] E. Badal-Valero, J. A. Alvarez-Jareno, and J. M. Pavía. "Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish court case". In: *Forensic Science International* 282 (2017), pp. 24–34. DOI: https://doi.org/10.1016/j.forsciint.2017.11.008.

[5] G. Barroso. *All models are wrong, but some are useful*. Ed. by AdMoRe-updates. 2018. URL: https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/.

[6] F. Benford. "The Law of Anomalous Numbers". In: *American Philosophical Society* (1938). URL: http://www.jstor.org/stable/984802.

[7] A. Berger and T. P. Hill. "A basic theory of Benford's Law". In: *Probability Surveys* (2011). DOI: 10.1214/11-PS175.

[8] A. Berger and T. P. Hill. "The mathematics of Benford's law: a primer". In: *Statistical Methods and Applications* (2020). DOI: https://doi.org/10.48550/arXiv.1909.07527.

[9] M. F. Brilhante et al. "Meta-analysis of Genuine and Fake p-Values". In: *Journal of Statistical Theory and Practice*. 19th ser. (2025), p. 29. DOI: https://doi.org/10.1007/s42519-025-00445-3.

[10] M. F. Brilhante et al. “Two P or Not Two P: Mendel Random Variables in Combining Fake and Genuine p-Values”. In: *AppliedMath*. 4th ser. (2024), pp. 1128–1142. DOI: https://doi.org/10.3390/appliedmath4030060.

[11] J. Caballero et al. “Benford’s Law in histology”. In: *Journal of Pathology Informatics* 18, 2-s2.0-105010485022 (2025). ISSN: 22295089. DOI: 10.1016/j.jpi.2025.100458.

[12] B. Chakrabarty, P. C. Moulton, and L. Pugachev. “Catch me if you can: In search of accuracy, scope, and ease of fraud prediction”. In: *Review of Accounting Studies* 30, 2-s2.0-85204557302 (2 2025). Ed. by Springer, pp. 1268–1308. ISSN: 13806653. DOI: 10.1007/s11142-024-09854-4.

[13] J. Collins. “Using Excel and Benford’s Law to detect fraud”. In: *Journal of Accountacy* (2017). URL: https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud/.

[14] A. Constantinides et al. “Applying Benford’s Law as an Efficient and Low-cost Solutionfor Verifying the Authenticity of Users’ Video Streams inLearning Management Systems”. In: *ACM Digital library* (2022). DOI: https://dl.acm.org/doi/10.1145/3486622.3493993.

[15] J. Debener, V. Heinke, and J. Kriebel. “Detecting insurance fraud using supervisedand unsupervised machine learning”. In: *Journal of Risk and Insurance* 90 (3 2023), pp. 743–768. DOI: DOI:10.1111/jori.12427.

[16] J. Deckert, M. Myagkov, and P. C. Ordeshook. “Benford’s Law and the Detection of Election Fraud”. In: *Cambridge University Press*. Political Analysis 19 (2017), pp. 245–268. DOI: 10.1093/pan/mpr014.

[17] A. Diekmann. “Not the First Digit! Using Benford’s Law to Detect Fraudulent Scientific Data”. In: *Journal of Applied Statistics* 34.3 (2007), pp. 321–329. DOI: 10.1080/02664760601004940.

[18] P. D. Drake and M. J.Nigrini. “Computer assisted analytical proceduresusing Benford’s Law”. In: *Journal of Accounting Education* 18 (2 2000), pp. 127–146. DOI: https://doi.org/10.1016/S0748-5751(00)00008-7.

[19] C. Durtschi, W. Hillison, and C. Pacini. “The Effective Use of Benford’s Law to Assist in Detecting Fraud in Accounting Data”. In: *Journal of Forensic Accounting* V (2004), pp. 17–34. ISSN: 1524-5586. URL: https://www.researchgate.net/publication/241401706_The_Effective_Use_of_Benford's_Law_to_Assist_in_Detecting_Fraud_in_Accounting_Data.

[20] J. Ensminger and J. Leder-Luis. “Detecting Corruption: Evidence from a World Bank project in Kenya”. In: *World Development* 188, 2-s2.0-85211130460 (2025). ISSN: 0305750X. DOI: 10.1016/j.worlddev.2024.106858.

[21] P. Fernandes and M. Antunes. "A Benford's Law Based method to Detect Manipulated Digital Photos". In: (2021). URL: https://www.dcc.fc.up.pt/~mantunes/papers/recpad2022.pdf.

[22] P. Fernandes and M. Antunes. "Benford's law applied to digital forensic analysis". In: *Forensic Science International: Digital Investigation* 45 (2023). DOI: https://doi.org/10.1016/j.fsidi.2023.301515.

[23] P. Fernandes, S. Ó Ciardhuáin, and M. Antunes. "Distance-based feature selection using Benford's law for malware detection". In: *Computers & Security*. 104625th ser. 158 (2025). DOI: https://doi.org/10.1016/j.cose.2025.104625.

[24] P. Fernandes, S. Ó Ciardhuáin, and M. Antunes. "Enhancing IoMT Security by Using Benford's Law and Distance Functions". In: *Pattern Recognition and Image Analysis. IbPRIA 2025*. Ed. by N. Gonçalves, H.P. Oliveira, and J.A. Sánchez. Vol. 15937. Lecture Notes in Computer Science. Cham: Springer, 2025, pp. 54–67. DOI: 10.1007/978-3-031-99565-1_5.

[25] P. Fernandes, S. Ó Ciardhuáin, and M. Antunes. "Unveiling Malicious Network Flows Using Benford's Law". In: *mathematics* (2024). DOI: https://doi.org/10.3390/math12152299.

[26] R. M. Fewster. "A Simple Explanation of Benford's Law." In: *The American Statistician* 63.1 (2009). DOI: 10.1198/tast.2009.0005.

[27] R. A. Fisher. *Statistical Method for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[28] N. Gauvrit and J. Delahaye. "Pourquoi la loi de Benford n'est pas mystérieuse". In: *Centre d'analyse et de mathématique sociales de l'EHESS* (June 30, 2008), pp. 7–15. ISSN: 1950-6821. DOI: 10.4000/msh.10363. URL: http://journals.openedition.org/msh/10363.

[29] J. Grodner and A. E. Rubin. "Benford's Law: applications to chondrules and refractory inclusions". In: *Discover Space* 129 (2025), p. 2. DOI: 10.1007/s11038-025-09561-3.

[30] T. P. Hill. "A Statistical Derivation of the Significant-Digit Law". In: *Statistical science* 10.4 (1995), pp. 354–363. DOI: 10.1214/ss/1177009869.

[31] A. Hyseni and J. Petráš. "Big Data in Electric Power Engineering". In: *Research and Developments in Electrical Power Engineering. ELEKTROENERGETIKA 2024*. Ed. by Ľ. Beňa et al. Vol. 1446. Lecture Notes in Electrical Engineering. Cham: Springer, 2025. DOI: 10.1007/978-3-031-97333-8_34.

[32] G. Judge and L. Schechter. "Detecting Problems in Survey Data Using Benford's Law". In: *Journal of Human Resources* 44.1 (2009), pp. 1–24. URL: https://www.jstor.org/stable/20648886.

[33] E. Kessel. "Benford's Law: Potential Applications for Insider Threat Detection". In: *Software Engineering Institute* (Dec. 17, 2020). URL: https://insights.sei.cmu.edu/blog/benfords-law-potential-applications-insider-threat-detection/.

[34] J. Kobiela and P. Dzierwa. "Application of Benford's Law to the Identification of Non-authentic Digital Images". In: *Advances in Mobile Computing and Multimedia Intelligence. MoMM 2024*. Ed. by P. Delir Haghighi et al. Vol. 15341. Lecture Notes in Computer Science. Cham: Springer, 2025, pp. 115–129. ISBN: 978-3-031-78048-6. DOI: 10.1007/978-3-031-78049-3_12.

[35] A. E. Kossovsky. *A Comprehensive Summary of the Benford's Law Phenomenon. On the Unequal Spread of Digits within Scientific and Typical Data*. 2nd. World Scientific, 2025, p. 284. DOI: https://doi.org/10.1142/14054.

[36] H. Li et al. "Research on improving the quality of groundwater self-monitoring via Blockchain technology". In: *Environmental Impact Assessment Review* 112, 2-s2.0-85214574519 (2025). ISSN: 01959255. DOI: 10.1016/j.eiar.2025.107811.

[37] D. C. Montgomery. *Introduction to Statistical Quality Control*. Ed. by Wiley. 8th ed. 2020.

[38] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers*. 7th. New York: John Wiley & Sons, 2018.

[39] B. Murteira et al. *Introdução à Estatística*. Ed. by Escolar Editora. 4th. 2023.

[40] S. Newcomb. "Note on the Frequency of Use of the Different Digits in Natural Numbers". In: *The Johns Hopkins University Press* 4.1 (1881), pp. 39–40. DOI: https://doi.org/10.2307/2369148. URL: https://www.jstor.org/stable/2369148.

[41] M. J. Nigrini. "A taxpayer compliance application of Benford's law". In: *Journal of the American Taxation Association* 18.1 (1996), pp. 72–91. ISSN: 0198-9073.

[42] M. J. Nigrini. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Hoboken, NJ, USA: JohnWiley & Sons, 2012.

[43] M. J. Nigrini. "Using Benford's Law to reveal journal entry irregularities". In: *Journal fo Accountacy* (2022). URL: https://www.journalofaccountancy.com/issues/2022/sep/using-benfords-law-reveal-journal-entry-irregularities/.

[44] L. Pardo. *Statistical Inference Based on Divergence Measures*. 1st Edition. New York: Chapman and Hall/CRC, Nov. 12, 2018. ISBN: 9780367578015. DOI: https://doi.org/10.1201/9781420034813.

[45] D. Pestana and S. F. Velosa. *Introdução à probabilidade e à estatística*. 4th. Vol. I. Fundação Calouste Gulbenkian, 2010.

[46] V. Raju et al. "Supervised Learning Models for Enhancing Financial Fraud Detection Systems". In: *Journal of Information Systems Engineering and Management* 10.17 (2025). DOI: https://doi.org/10.52783/jisem.v10i17s.2786.

[47] R. Santos et al. "Accuracy Measures for Binary Classification Based on a Quantitative Variable". In: *REVSTAT-Statistical Journal*. 17(2) (2019), pp. 223–244. DOI: https://doi.org/10.57805/revstat.v17i2.266.

[48] P. E. Sastroredjo. "Benford's Laws Analysis on Tax Irregularities inBanking and Investment Activities: The Case of theFTSE All-Share Index". In: *Review of Integrative Business and Economics Research* 14, 2-s2.0-85219022795 (2 2025), pp. 674–686. ISSN: 24146722.

[49] Shalini et al. "Data Quality Assessment Using Benford's Law and Excel". In: *IEEE Xplore* (2023). DOI: 10.1109/ICAECT57570.2023.10118030.

[50] B. Šinik and A. Tošić. "Testing life-cycle assessment data quality with Benford's law reveals geographic variation". In: *Ecological Informatics* 90, 2-s2.0-105008090181 (2025). ISSN: 15749541. DOI: 10.1016/j.ecoinf.2025.103227.

[51] S. Suboh and I. A. Aziz. "Discovering Patterns and Deviations in Data: Comparision of Anomaly Dtection Procedure in Regression". In: *Advances and Applications in Statistics* 91.9 (Sept. 2024), pp. 1195–1215. DOI: https://doi.org/10.17654/0972361724063.

[52] A. Taushanov. "Seismic analysis based on Newcomb–Benford Law deviation estimation". In: *Progress in Engineering Science* 2, 2-s2.0-105000697186 (1 2025). DOI: 10.1016/j.pes.2025.100051.

[53] S. Tirunagari et al. "Using Benford's law to detect anomalies in electroencephalogram: An application to detecting alzheimer's disease". In: *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Manchester: IEEE, 2017, pp. 1–6. DOI: 10.1109/CIBCB.2017.8058547.

[54] J. Vičič and A. Tošić. "Application of Benford's Law on Cryptocurrencies". In: *Journal of Theoretical and Applied Electronic Commerce Research* (2022). DOI: 10.3390/jtaer17010016.

[55] Z. Wang, G. Xu, and M. Ren. "Can attention detect AI-generated text? A novel Benford's law-based approach". In: *Information Processing and Management* 62, 2-s2.0-105000247439 (4 2025). ISSN: 03064573. DOI: 10.1016/j.ipm.2025.104139.

[56] W. A. Woodward, B. P. Sadler, and S. Robertson. *Time Series for Data Science. Analysis and Forecasting*. 1st. New York: Chapman and Hall/CRC, Aug. 1, 2022. DOI: https://doi.org/10.1201/9781003089070.

[57] J. Zhang, Z. Wang, and X. Tang. "How can governments mitigate statistical data manipulation? Evidence from China's enterprises' direct report reform". In: *Eco-*

*nomics Letters* 250, 2-s2.0-105000635702 (2025). ISSN: 01651765. DOI: 10.1016/j.econlet.2025.112293.

*This page intentionally left blank.*