ACROSS: Accessible Country-Relevant Outlier Selection System

Cantay Caliskan
University of Rochester

Oskar Lelko
University of Rochester

Benjamin Noe
Georgia Institute of Technology

Keitaro Ogawa
Columbia University

December 28, 2024

Abstract

This study examines the disparity between traditional outlier detection methods used in political science and the advancements in contemporary data science techniques. Expanding on foundational work, such as Cook's Distance (1977), this paper evaluates the performance of 33 advanced outlier detection methods in the context of cross-country socioeconomic and political datasets. A systematically constructed labeled test set, covering 14 countries identified as empirical outliers across decades of political science research, serves as the basis for this analysis. The findings reveal that Cook's Distance, a traditional method widely used in political science, is significantly outperformed by at least 90 combinations of modern algorithms and data normalization strategies. Among these, the Cluster-Based Local Outlier Factor demonstrates an 87% improvement over Cook's Distance. Additionally, the novel ensemble model introduced in this study, ACROSS (Accessible Country-Relevant Outlier Selection System), achieves a remarkable 251% improvement in recall and a notable enhancement in the specificity-sensitive F1 metric compared to Cook's Distance. The sensitivity analysis was conducted using artificially created two-dimensional datasets designed to represent normal and skewed distributions, allowing for a controlled assessment of algorithm performance. On datasets following a normal distribution, ACROSS achieved a mean recall of 0.935 and an F1 score of 0.781, significantly outperforming competing methods. For skewed distributions, ACROSS maintained robust performance, achieving a mean recall of 0.912 and a mean F1 score of 0.745. These results underscore ACROSS's adaptability to diverse data patterns, including those with extreme variations in numerical ranges. By achieving recall improvements exceeding 10% and F1 score enhancements surpassing 12% compared to competing methods in controlled experiments, ACROSS demonstrates remarkable generalizability and effectiveness. These results position ACROSS as a versatile and reliable tool for outlier detection, offering significant contributions to enhancing analytical accuracy and methodological rigor in political science research.

Keywords: Outlier Detection, Political Methodology, Ensemble Models

1 Introduction

> If a potentially critical observation has been detected using one or more of the above measures then the examination of the effects of deleting the observation seems a natural next step. However, the problem of determining which point(s) to delete can be very perplexing.
>
> *R. Dennis Cook, the inventor of Cook's Distance, on the need for algorithmic outlier detection [18, p. 15].*

Outlier detection is a critical aspect of data analysis across various fields, including political science, economics, and social sciences. While much of the traditional focus has been on methods designed to identify significant observations within a linear regression context, such as Cook's Distance, the issue of outliers extends well beyond this scope. Outliers can arise in a multitude of contexts, requiring diverse approaches for both predictive and inferential analysis.

This paper primarily focuses on outlier detection methods within predictive frameworks, which have seen significant advancements in recent years. While our study emphasizes predictive techniques, particularly through the development of an ensemble method called ACROSS, we also recognize the broader implications of outliers on statistical inference. The inconsistent treatment of outliers across different detection methods can lead to inferential errors, which may significantly impact the validity and reliability of research findings. To address these concerns, we explore the effectiveness of various outlier detection methods in the context of cross-country socioeconomic and political indicators. By building on traditional approaches like Cook's Distance and integrating modern ensemble techniques, such as ACROSS, this study aims to enhance the accuracy and reliability of outlier detection. In doing so, we provide a comprehensive analysis that bridges the gap between traditional methodologies and modern advancements, while also considering the potential for varied inferential outcomes depending on the method employed.

This paper is organized to methodically build a foundation for the introduction of the ACROSS model, an ensemble approach to outlier detection in cross-country political and socioeconomic datasets. Section 1 begins with a thorough examination of the challenges inherent in traditional outlier detection methods within the context of political science, grounding the study in established literature. Section 2 delves into the limitations of existing models, particularly their inadequacies in handling high-dimensional and non-linear data relationships, thus setting the stage for the necessity of an advanced method. The core of the paper is presented in Section 3, where we introduce and detail the ACROSS model, discussing its design, application, and superiority over earlier methods. The latter sections of the paper focus on the empirical results and implications, where we rigorously test the ensemble model across various scenarios and compare its performance to traditional approaches. Finally, the conclusion synthesizes the findings and highlights the broader contributions of this research to the field of political science, suggesting avenues for future inquiry.

1.1 Outlier Detection for Political Science

Countries remain a central focus of political science research, as they represent the fundamental political structures that govern the lives of individuals. Concurrently, the standard tools of data science and analysis are continuously advancing, enabling analysts

and academic researchers to extract knowledge from data with greater precision. Among these tools is outlier detection, also referred to as anomaly detection—a category of algorithms designed to identify data points that significantly deviate from the broader dataset. Outlier detection has been widely applied across various industries, including healthcare and cybersecurity [51, 11]. However, despite its extensive use in quantitative economics [25], there has been limited systematic exploration of novel outlier detection methods within the political science literature. This study seeks to address this gap by conducting a comprehensive comparison of outlier detection methods in the context of political science, thereby illustrating their potential to significantly enhance research within the field.

In the study of comparative politics, outliers often represent countries that deviate significantly in variables such as human development, economic growth, or conflict, relative to factors like regime type or variety of capitalism. Awareness of outliers is vital to political science research for two primary reasons: first, outliers challenge the boundaries of existing theories, thereby fostering theoretical refinement; second, they serve as valuable case studies that merit in-depth, standalone analysis [32].

On the discriminative side, the removal of outliers is essential for ensuring the accuracy of conclusions drawn from data, as it helps to verify adherence to both methodological and empirical model constraints. Methodological considerations may include evaluating the suitability of a linear model [49], while empirical considerations might involve assessing the robustness of results across diverse political contexts. The importance of this process is clearly illustrated by [14] and [13], who show that careful removal of outliers in replication data can invalidate—or even reverse—the findings of published political science studies. As [15, p. 153] asserts:

> When researchers ignore abnormal observations, especially with respect to dependent variables, their empirical results can be misleading. Unfortunately, the fact remains underappreciated in studies of political science.

Outlier detection serves purposes beyond merely eliminating anomalous data points; it also plays a crucial role in identifying data subsets that are noteworthy precisely because they deviate from the norm. Welles (2014) compellingly argues that, in the context of social groups, outliers do not merely represent deviations but rather "the plentiful representation of minorities." In the era of big data, outlier detection allows us to address the historical omission of "women, minorities, and statistical outliers" from the scientific record [52]. This perspective is further reinforced by Noble (2018), who underscores how biases in data collection and analysis can marginalize specific groups, thereby highlighting the importance of inclusive methodologies in social research [41]. Such considerations are particularly relevant in social and political science, where outlying data may reflect counter-majority trends or represent marginalized communities.

In general, employing robust algorithms to remove outliers from multivariate data allows researchers to prevent non-representative responses from introducing bias into their models. Outlier detection also directs attention to data that might otherwise be overlooked, thereby facilitating more focused and equitable research. Innovative applications of outlier detection algorithms to computational social science have already generated knowledge related to politics: they can identify significant political events from Twitter activity [54], determine gerrymandering from electoral maps [43], and prove that so-called surgical drone strikes have an impact on the civilian population [16]. An outlying country like Tanzania may give insight into how it has maintained peace despite high ethnic

fractionalization [38], while the outlying case of Vietnam can demonstrate success in providing education despite resource constraints [35].

While much of this argument is applicable across the broader social sciences, the technical aspects of this paper necessitate a more clearly defined problem scope. Consequently, this study will focus specifically on multivariate datasets of cross-country socioeconomic and political indicators, given their prevalent use in political science and international relations research.

1.2 The Problem of Algorithm Choice

As early as the 1980s, regression diagnostics, such as Cook's Distance, were employed to identify outliers in political science research [12]. While these methods have proven to be reliable and have remained standard in political data analysis over several decades [34, 36, 47], they are not without limitations, particularly in their effectiveness when dealing with high-dimensional data. Moreover, measures like Cook's Distance were originally designed for use with least squares models and may no longer be suitable for more complex models, such as support vector machines or deep neural networks, which are gaining popularity in the field due to the rise of computational social science [56, p. 14]. Neural networks, in particular, are increasingly utilized for tasks such as sentiment analysis of political discourse on social media, predictive modeling of political outcomes, and uncovering complex patterns in large-scale political data [56, p. 3].

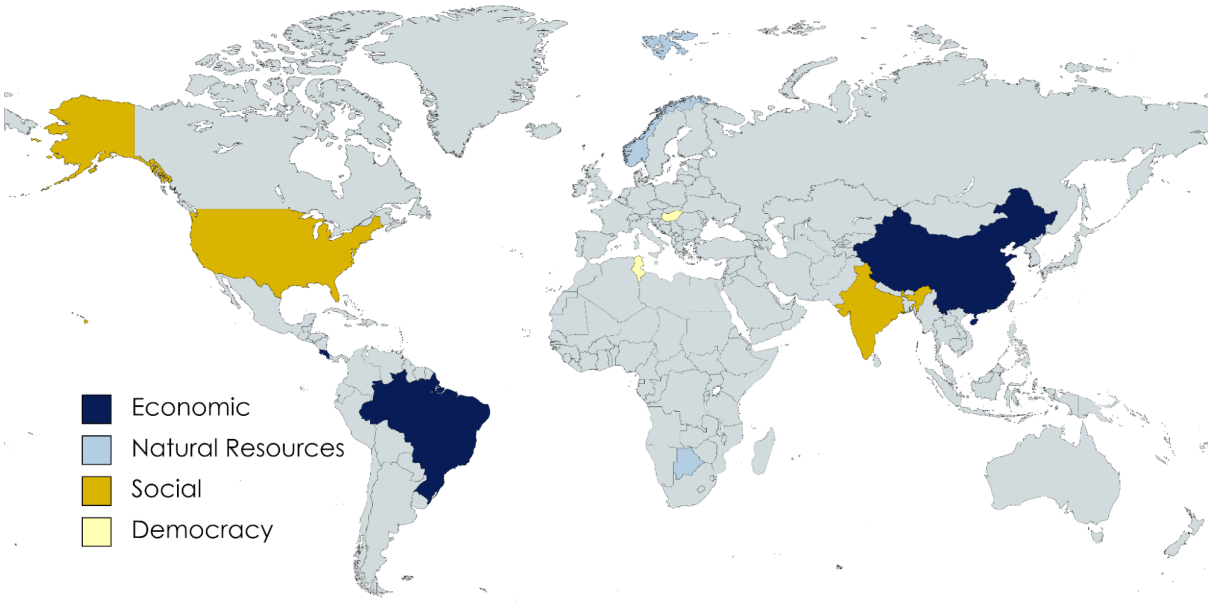
Although the performance of various novel algorithms has been extensively surveyed and compared in recent studies [3, 51, 24], such analyses are notably absent in the domain of political science. While these algorithms have demonstrated promise in specialized fields like fraud and intrusion detection—where outlier detection techniques originally emerged [19]—it is crucial to assess their applicability to political science research. Data in political science, particularly in cross-country comparisons, presents unique challenges. These include a limited sample size coupled with a large number of variables, which can introduce significant bias and variance: a dataset of 195 countries might encompass hundreds of socioeconomic and political indicators. Additionally, the challenge of data labeling is significant; different countries may be considered outliers based on varying variables, yet no unified dataset exists that treats countries as outliers, given the exponential number of possible country-variable combinations. Unlike the clear definition of an outlier in signal processing—such as abnormal noise—defining what constitutes an outlier in the context of nations is far more complex.

In response to these challenges, this study introduces ACROSS (Accessible Country-Relevant Outlier Selection System), an ensemble model designed to offer a more tailored approach to outlier detection in international socioeconomic and political indicators.

2 Empirical Outliers In the Literature

As we move from a theoretical discussion of outlier detection to examining empirical examples, it is crucial to connect the importance of identifying outliers in political science with concrete cases that have been studied across different contexts. By exploring how outliers have been identified and analyzed in various sociopolitical settings, we gain a clearer understanding of the real-world impact of these anomalies on theoretical frameworks and policy analysis. This examination will highlight the practical implications of

Figure 1: Examples of outliers in terms of various sociopolitical variables



outlier detection, illustrating why a more sophisticated approach, such as the one proposed in this paper, is necessary for handling the complexities of multivariate datasets in political science.

Individual country outliers are typically identified through data analysis and substantiated by theoretical frameworks. In cross-country socioeconomic and political datasets, researchers generally select explanatory and outcome variables, then examine how the outcome variable varies across nations within a specific region.[1] Using this approach, we explore several examples of empirical outliers from the literature, organized by decade.

Effective outlier detection in cross-country socioeconomic and political research requires a comprehensive approach that considers anomalies in both explanatory and outcome variables. Focusing solely on one type of variable can lead to incomplete or skewed analyses, as outliers in either category can have substantial effects on the results. Therefore, any rigorous analysis must be designed to identify and address outliers across all relevant variables. ACROSS exemplifies this approach by not limiting its detection capabilities to outliers associated with dependent variables; instead, it systematically analyzes the entire dataset to identify outliers in both independent and dependent variables. This ensures a more accurate representation of the data and a deeper understanding of the underlying patterns, ultimately leading to more robust and reliable conclusions in cross-country comparative studies.

2.0.1 Economic

In the 1990s, Brazil, China, and Costa Rica emerged as notable outliers, diverging from other countries in their regions and with similar government types, particularly in the areas of economic performance and military spending. First, Brazil, despite its democratic governance, increased its military expenditure—a pattern atypical for a democracy. This

[1]It is important to note that ACROSS is designed to detect outliers in both explanatory and outcome variables, ensuring a comprehensive analysis. This distinction is crucial because outliers can occur in either dependent or independent variables, and their effects on the analysis can differ significantly depending on the type.

behavior was driven by concerns over far-left groups potentially overthrowing the democratic government [55]. Next, China stood out as perhaps the most significant outlier of the decade. Despite its authoritarian regime, China experienced unprecedented economic growth and attracted substantial foreign direct investment (FDI). While democratic nations generally attract higher FDI due to greater stability and economic freedom, China's FDI inflows surpassed many democracies, possibly due to its guaranteed property rights [15]. Finally, Costa Rica deviated from its Latin American neighbors through its stable economy: while Brazil militarized, Costa Rica's decision to abolish its military enabled significant investments in social programs, resulting in notably high-quality education and healthcare. Additionally, "economic openness" further contributed to the country's economic success [46].

2.0.2 Natural Resources

Norway and Botswana each presented unique cases related to natural resources during the 1990s and 2000s. Unlike many other resource-rich countries, Norway's abundance of natural resources did not result in liberal immigration policies. Norway's pluralist democracy distinguishes it from many oil-rich nations, where immigration policies are often more liberal to attract foreign workers. In Norway, however, immigration policies are shaped by compromises between pro- and anti-immigration groups [13]. Conversely, Botswana, a developing nation also rich in natural resources, exhibited an unexpected pattern of economic development. Contrary to the common expectation that resource-rich developing countries will experience high levels of inequality and authoritarianism, Botswana has maintained a democratic system, bolstered by anti-corruption policies and a transparent public sector. This makes Botswana an outlier among resource-rich, developing nations [28].

2.0.3 Social

During the 2000s, the United States exemplified an outlier in several socioeconomic indicators. Despite high per capita healthcare spending, the country did not achieve commensurate improvements in health outcomes, particularly in terms of disability-adjusted life years from 2004 to 2011 [40]. Additionally, the elevated GDP per capita did not translate into stronger union representation or a more influential left-wing political presence [31]. India also emerged as an outlier during this period. Although the country experienced an increase in female literacy rates between 2000 and 2015, this did not lead to a corresponding rise in female workforce participation, largely due to "prevailing social norms and patriarchy" [22, p. 41].

2.0.4 Democracy

Given the diverse paths of democratic development around the world, examining outliers in democracy research offers valuable insights into the complexities of political change. The 2010s saw the emergence of new outliers in terms of regional democracy, most notably Tunisia and Hungary. Tunisia's transition to democracy was an exception, particularly when contrasted with the predominantly authoritarian landscape of the Middle East and North Africa. Tunisia has been an outlier due to several factors, including a large middle class, a well-educated population, a relatively homogenous society, and strong ties to Europe—factors that have enabled the country to maintain more democratic institutions

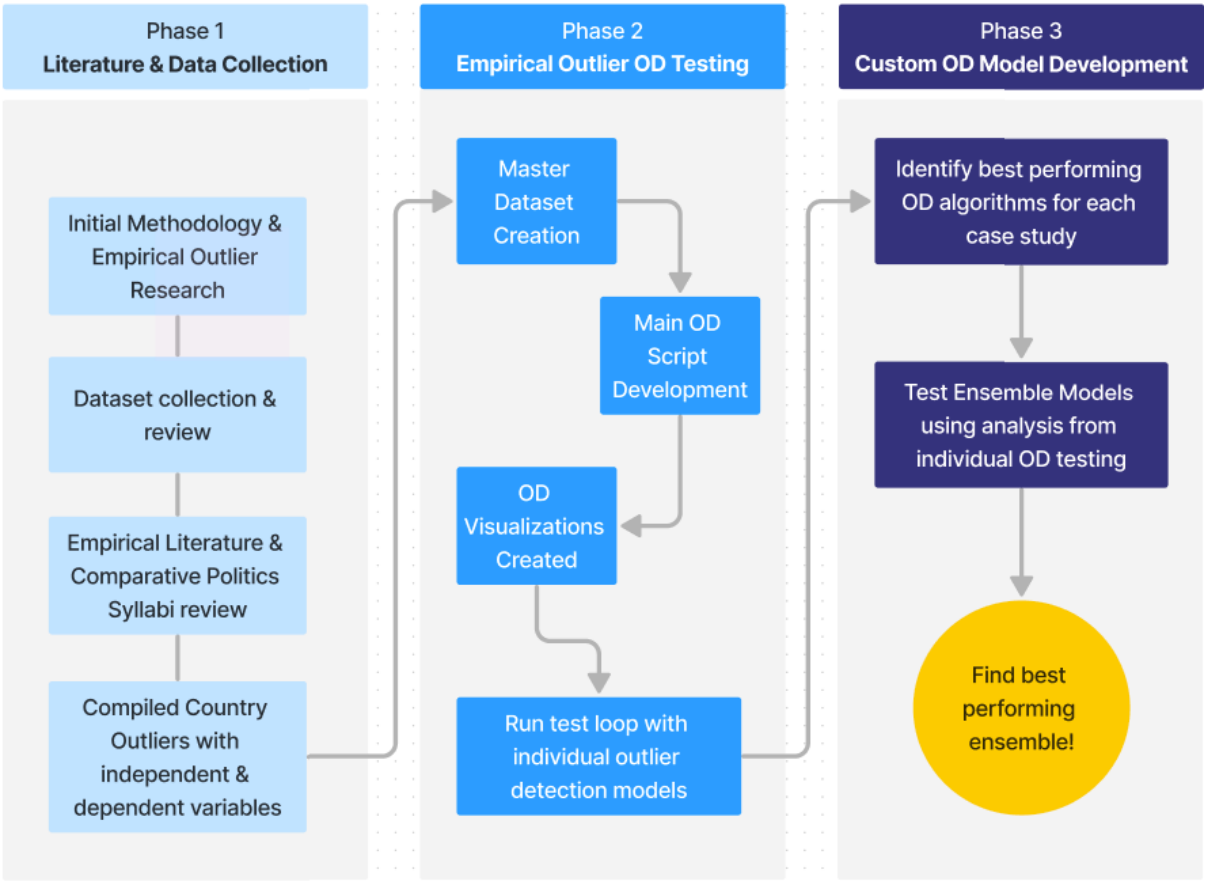
than those of other MENA nations [4]. In contrast, Hungary, a member of the European Union, exhibited a significant retreat from democratic principles, challenging the democratic norms typically associated with EU membership. Hungary has faced criticism for its increasing government control over the judiciary, media, and civil society—practices that set it apart from other European nations [7, 60, 10].

3 Methods

We begin by identifying the prevalent themes of outlier detection in the political science literature and develop a custom-labeled set of representative outliers based on predefined criteria. Next, we compile a master dataset from a broad spectrum of reliable socioeconomic indicators and select a collection of 163 candidate outlier detection methods. Each method is then tested against the labeled outliers, and their average performance is calculated. Finally, we present the results in a standardized format and use them to develop ACROSS (Accessible Country-Relevant Outlier Selection System), a custom outlier detection method specifically tailored to the field of political science. This method is based on a voting ensemble of the most generalizable algorithms.

The following subsections provide a detailed discussion of each technical step.

Figure 2: Our research methodology



3.1 Candidate Algorithms

We leverage PyOD [57], a popular open-source Python library for outlier detection, to test a wide range of unsupervised outlier detection algorithms across six paradigms: linear models, probabilistic models, proximity-based models, ensemble models, neural networks, and graph-based models. We additionally anticipate that many algorithms will vary in their performance based on the type of data normalization used. Therefore, for each PyOD algorithm, we additionally test four standard data normalization methods on our data matrix: min-max scaling, robust scaling with interquartile range, scaling with unit variance, and normalization to unit norm[2].[3] We use the default hyperparameters of each algorithm for effective comparison with other benchmarks.

The use of PyOD allows us to capture outlier detection development across decades, with its methods ranging from 1977 to 2023, as well as access to a unified API, enabling us to run thousands of experiments. Additionally, the community behind PyOD has recently published a comprehensive benchmark comparing the performance of PyOD methods on a range of tasks that excludes political science, allowing us to compare our results with those achieved on labeled datasets from other disciplines [24].

PyOD is an ideal choice for analyzing and implementing outlier detection algorithms due to its extensive library, which spans methods developed from 1977 to 2023. This breadth allows researchers to access a wide range of algorithms, from classic approaches to cutting-edge techniques, all within a unified API. The library's versatility enables us to conduct thousands of experiments efficiently, providing a robust platform for comparing different outlier detection methods. PyOD covers a substantial majority—estimated around 70-80%—of the most relevant and widely used outlier detection techniques, making it one of the most comprehensive tools available for this purpose. Additionally, the PyOD community has recently published a comprehensive benchmark that evaluates the performance of PyOD methods across various tasks, offering valuable insights into how these methods perform in diverse contexts. This benchmark serves as a critical point of reference, allowing us to situate our findings within a broader framework of outlier detection research [24].

After excluding outlier detection methods that exceed the time limit on any of the target tasks[4], we retained 33 methods and 163 method-normalization pairs for testing on each country. The complete set of algorithms, along with our full results, is provided in the Appendix.

3.2 Identifying Outliers

To compare the algorithms, we evaluated them on a custom test set of 14 countries listed in Table 1, with each country representing a distinct outlier detection task. These countries were empirically identified as outliers based on a range of indicators, including

[2]The four methods use the sklearn API's minmax_scale, robust_scale, scale, and normalize, respectively [42].

[3]Normalization is crucial for detecting outliers in social and political data because these datasets often contain variables measured on different scales, such as GDP per capita and literacy rates, which can skew the analysis if not adjusted. By normalizing the data, we ensure that each variable contributes equally to the outlier detection process, preventing variables with larger ranges from disproportionately influencing the results.

[4]We set a time limit of 2 minutes per detection task (sample size 81-5092 rows, 1-3 dimensions) on an M1 MacBook Pro with 16 GB of memory.

regime type, economic performance, health outcomes, and ethnic fractionalization, among others. Several of these cases are further explored in the Literature Review section.

Although outlier detection is an essential part of political science methodology, there is no available dataset of countries that are labeled as empirically "anomalous." Intuitively, nearly any country can be considered an outlier with respect to some specific set of variables. This stems from the nature of our problem: there is a limited number of countries, but a theoretically unbounded set of variables that could be devised to describe the relations between millions of people. Therefore, in crafting our test set of outlying countries, we rely on the following proxy: a country can be considered an outlier, with respect to a specific set of variables and in a given timespan, if credible political science research has been published that analyzes that country in the framework of a most similar system's design as an exceptional case study.[5] Our approach is related to Welles's argument that outliers in social science can be considered phenomena worthy of specific research [52].

To systematically identify these outlying countries, we examined PhD-level comparative politics course syllabi from top-ranked U.S. political science departments[6] from 2010-2024, and counted the frequency with which each country was mentioned[7]. We used this frequency as a proxy for the degree of interest political scientists have in various countries and their representativeness in cross-country comparisons. For the most frequently mentioned countries, we reviewed articles published in prominent political science journals, coding the relevant indicators in which these countries are considered outliers, as well as the time periods analyzed. The ACROSS model was particularly effective in identifying these empirically outlying countries, highlighting its utility in political science research. The result is a custom test set of 14 empirical outliers based on a diverse set of 20 socioeconomic and political variables.

[5]Most Similar Systems Design (MSSD) is a method in political science where similar cases are compared to identify the variable responsible for differing outcomes (Lijphart, 1971; Przeworski & Teune, 1970). This approach helps isolate causal factors by holding other variables constant.

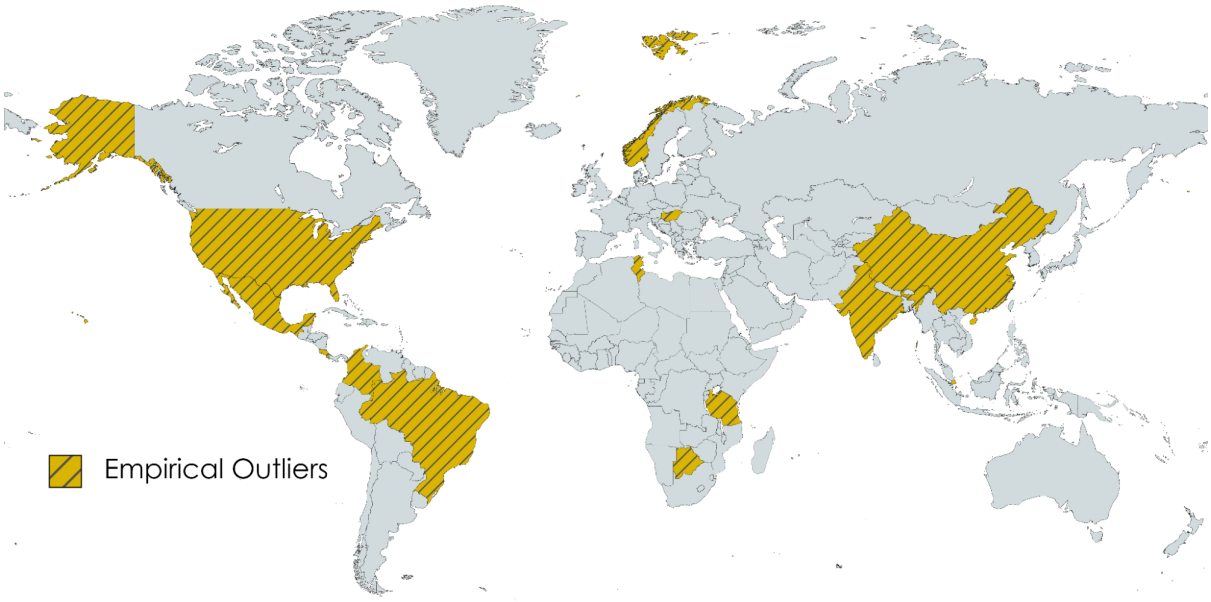
[6]Stanford, Harvard, Princeton, Yale, Massachusetts Institute of Technology, University of California San Diego, University of California Los Angeles, University of North Carolina Chapel Hill, New York University, Ohio State, Northwestern, University of Pennsylvania, University of Rochester, University of Texas Austin

[7]If a syllabus included geographic regions, we counted each plausibly outlying country within that region as mentioned.

Table 1: Empirically identified outliers

Country	Outlying in	With respect to	Timespan	Evidence	Interest[a]
China	FDI per capita	Level of democracy	1992—1995	Choi 2009	7
Tunisia	Level of democracy	MENA region	2011—2020	Bellin 2013	7
Costa Rica	Log GDP per capita, FDI per capita	Latin America region	1949—2015	Sada 2015	6
Botswana	Corruption perception	GDP growth rate, Natural resources	1990—2006	Iimi 2007	6
United States	Disability-adjusted life years	Healthcare spending per capita	2004—2004	Mor 2022	6
United States	Union strength, Left political representation	Log GDP per capita	2004—2011	Karabel and Laurison 2011	6
India	Female workforce participation	Female literacy rate, MENA region	2000—2015	Ghai 2018	4
Brazil	Military spending as % of GDP	Level of democracy	1995—2006	Zaverucha 2009	4
Mexico	Level of human rights	Level of democracy	1948—2008	Brennan et al. 2022	3
Hungary	Level of democracy	EU membership	2010—2023	Bogaards 2018, Buzogány 2017	2
Singapore	GDP growth	Level of democracy	1965—2005	Siddiqui 2010	2
Tanzania	Level of peace	Ethnic fractionalization	1992—2020	Malipula 2014, Rwengabo 2016	1
Colombia	Disability-adjusted life years	Healthcare spending per capita	2000—2016	Mor 2022	1
Norway	Non-restrictive immigration policy	Natural resources	1970—2013	Choi 2022	1

Figure 3: Labeled dataset of empirically outlying countries



3.3 Application of Cook's Distance

Cook's Distance is a well-established diagnostic measure used in regression analysis to identify influential data points that may disproportionately affect the results of a model. It calculates the influence of each observation by measuring the change in regression coefficients when that observation is removed from the dataset. A high Cook's Distance value suggests that the observation is an outlier that has a significant impact on the model's parameters.

The formula for Cook's Distance (D_i) is given by:

$$D_i = \frac{r_i^2}{p \cdot \text{MSE}} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}$$

where:

- r_i is the residual for the i-th observation,
- p is the number of parameters in the model,
- MSE is the mean squared error of the model,
- h_{ii} is the leverage of the i-th observation.

In this study, Cook's Distance was used as a baseline method to identify potential outliers among countries based on their socioeconomic and political indicators. For each country-year observation in our dataset, we calculated the Cook's Distance using a multiple regression model where the dependent variable was a selected key indicator, and the independent variables were other relevant socioeconomic factors.

We identified countries with the maximum Cook's Distance as potential outliers. These countries were then compared to those identified by more contemporary outlier detection algorithms to assess the relative effectiveness of Cook's Distance. Our analysis

showed that while Cook's Distance remains a useful tool in traditional regression contexts, it was consistently outperformed by more advanced methods, as detailed in the results section.

This approach allows us to highlight the limitations of Cook's Distance in high-dimensional data environments, particularly when compared to the ensemble method we developed, ACROSS, which integrates multiple outlier detection techniques to provide a more nuanced understanding of outliers in cross-country data.

3.4 Country Dataset

After coding the outlier variables for each country, we construct a dataset of socioeconomic and political indicators for each country-year by merging the World Bank's World Development Indicators dataset with several other widely used political datasets, such as V-Dem. The specific datasets utilized are detailed in Table 2.

Table 2: Sources of socioeconomic and political variables

Author/Collector	**Title**	**Year**	**Variables Used**
The World Bank	World Development Indicators	2023	FDI per capita Log GDP Healthcare Spending per Capita Natural Resource Rents as % of GDP GDP per Capita Female Literacy Rate Military Spending as % of GDP GDP Growth Rate Female Workforce Participation Rate Liberal Immigration Policy Disability Adjusted Life Years
Varieties of Democracy	The V-Dem Dataset	2023	Union Strength Human Rights Index Democracy Index
Center for Systemic Peace	State Fragility Index and Matrix	2018	Peace Index
Varieties of Democracy	The V-Party Dataset	2022	Left-Wing Representation
Lenka Drazanova	Historical Index of Ethnic Fractionalization Dataset (HIEF)	2019	Ethnic Fractionalization
Transparency International	Corruption Perceptions Index	2022	Corruption Perceptions Index

3.5 Considerations of Inferential Errors

In the context of outlier detection, particularly when employing multiple methods or ensemble techniques, the potential for inferential errors cannot be overlooked. Inferential errors occur when the inclusion or exclusion of specific data points—classified as outliers by some methods but not others—leads to different conclusions about the relationships or effects within the dataset.

This issue is especially pertinent when comparing the results of different outlier detection methods. Each method may have unique sensitivities or biases, which can result in different sets of identified outliers. When these differing results are used to draw inferences, there is a risk that the conclusions may be inconsistent or even contradictory.

In this study, the ensemble method ACROSS was developed to mitigate some of these challenges by integrating multiple detection algorithms. However, it is crucial to understand that even ensemble methods are not immune to inferential errors. While ACROSS aims to provide a more balanced and comprehensive identification of outliers, the variation in underlying methods means that inferential consistency must be carefully monitored. Future work should explore how ensemble methods can be further refined to minimize such errors, possibly by incorporating techniques that assess the stability of inferences across different subsets of data.

3.6 Performance Metrics

For each empirical outlier, we extract a subset of the full dataset that covers only the timespan during which the country was identified as an outlier and includes only the variables in which it was an outlier. We label the country as an outlier for each year within this subset and apply minimal data pre-processing. Specifically, if an outlying category is regional, we limit the data to include only the countries within that region[8]. If any country has missing data in some of its variables, we impute those values with the median values of that country[9]. Imputing missing values is particularly beneficial when there is a high number of missing values per covariate, as it allows us to preserve as many observations as possible. In comparative politics and international relations research, where understanding patterns and anomalies across a wide range of countries is crucial, retaining a comprehensive dataset is essential. This approach aligns with the findings of King et al. (2001), who highlight that imputation methods can reduce biases and increase efficiency in comparative analyses by maximizing the number of cases included in the study. By ensuring that missing data does not lead to the exclusion of entire cases, we can analyze a more extensive set of countries, thereby enhancing the validity and generalizability of our findings. Our 14 experimental subsets, corresponding to the 14 outlying countries, have an average sample size of 1,871 and a dimensionality between 1 and 3.

We then proceed to run each algorithm-normalization pair on every subset, recording which countries are identified as hypothetical outliers, along with several key metrics: accuracy on the target country (recall), the number of additional outliers identified, and

[8]For example, instead of treating EU membership as a categorical variable for outlier detection, we exclude that variable and perform outlier detection solely on EU countries using the remaining variables.

[9]If a country is missing data for a variable across all covered years, we impute that variable with the median of all countries; however, we do not apply this imputation to the target country to avoid biasing our results.

an F1 score[10] that balances precision and recall. Finally, we generate two reports: the first presents the average recall and F1 score achieved by each algorithm-normalization pair across all outliers, while the second focuses on individual empirical outliers, calculating the average and best scores achieved for each by all methods. The first report enables us to identify the top-performing algorithms, while the second provides insights into the types of problems that pose the greatest challenges for current outlier detection methods.

Given the nature of our dataset, where we select only one representative outlier for each set of variables and timeframe, we label only a small minority of all country-years that could potentially be considered outliers. Consequently, the F1 score has limited informational value on its own, as the true number of outliers is unknown. Instead, we prioritize recall—the ability to accurately capture our empirical outliers—as the primary measure of quality. However, an F1 score extremely close to zero may indicate that an algorithm is not sufficiently selective. Additionally, we qualitatively assess the additional country-years identified by the top-performing algorithms for plausibility, highlighting some of them as intriguing cases worthy of further research.

Finally, we compare the performance of each algorithm to that of Cook's Distance, which serves as our baseline due to its widespread use in political science literature. In this comparison, ACROSS outperformed all other tested models, achieving the highest recall and F1 scores, making it a superior choice for outlier detection. By emphasizing the improvement over Cook's Distance rather than focusing solely on absolute metrics, we account for potential biases within our limited sample of empirical outliers.

4 Results

After conducting 2,538 experiments across 14 outliers, we found that, regardless of the normalization technique applied, Cook's Distance—often the preferred tool among political scientists—was outperformed by at least 90 combinations of normalization methods and newer outlier detection algorithms. The top- and bottom-ranking algorithms are presented in Table 3, with the complete results provided in the Appendix.

The proximity-based Clustering-Based Local Outlier Factor (CBLOF), introduced in 2003, achieves the best results, showing an 86.9% improvement over Cook's Distance on our test set. This outcome also highlights the critical role of normalization in effective outlier detection: without normalization, CBLOF's improvement over Cook's Distance drops to just 32%, while "robust" scaling using the interquartile range proves to be the most effective.[11]

Many of the highest-performing methods are proximity-based, including the Local Outlier Factor (LOF) introduced in 2000 and the Connectivity-Based Outlier Factor (COF) introduced in 2002. Several linear methods also performed well, such as Deviation-based Outlier Detection (LMDD), developed in 1996, and Kernel Principal Component Analysis (KPCA), introduced in 2007. Among the two graph-based methods we tested, Unifying Local Outlier Detection Methods via Graph Neural Networks (LUNAR), published in 2022, was particularly effective. Interestingly, the lowest-performing methods were predominantly recently developed neural network models, possibly due to their design being highly specialized for non-political-science tasks.

[10]The F1 score is the harmonic mean of precision and recall, calculated as $2 \times (\text{precision} \times \text{recall}) \div (\text{precision} + \text{recall})$.

[11]The full results are included in our Appendix.

The ACROSS model, which integrates several of these top-performing algorithms into an ensemble, showed promise in improving recall and F1 scores, suggesting it could be a valuable tool for outlier detection in political science datasets.

4.1 Sensitivity Analysis

The sensitivity analysis conducted in this study was designed to evaluate the robustness and performance of the ACROSS algorithm across a variety of experimental data scenarios. Given the complexity and diversity of cross-country socioeconomic and political datasets, it was essential to test the algorithm under conditions that mirror real-world data characteristics. The analysis, detailed in Tables 8 through 19 in the Appendix, highlights the algorithm's performance while offering insights into its adaptability in comparison to other outlier detection methods.

The study employed a two-dimensional variable framework, systematically varying both x and y values to reflect the diversity of distributions commonly encountered in real-world datasets. Three distinct distribution types were utilized: normal, skewed, and uniform. For the normal distribution experiments, four datasets, each comprising 2,000 two-dimensional samples, were generated. Outliers were positioned 10 standard deviations away from the mean of the distribution. The x-dimension was assigned a mean of 2 (or -2, depending on the experiment) with a standard deviation of 1, while the y-dimension was assigned a mean of 10 (or -10, depending on the experiment) with a standard deviation of 2. In the skewed distribution experiments, a skewness value of 5 (or -5) was applied to x (mean = 0, scale = 1) and y (mean = 5, scale = 2). Similar to the normal distribution case, four datasets, each containing 2,000 samples, were created. For the uniform distribution experiments, features were uniformly distributed within the ranges of -10 to 10 and 2 to 10. These ranges were systematically combined to create four distinct datasets, with each range alternately assigned as either the x- or y-dimension. In total, 12 artificial datasets, each consisting of 2,000 samples, were generated. Outliers constituted 5% of the data in each dataset, providing a consistent and controlled framework for assessing the algorithm's performance under varying distributional conditions.

To ensure a comprehensive assessment, the analysis included both positive and negative values for these variables, as well as datasets following normal and skewed distributions. The values samples had an absolute mean of 1, and were later normaluized after 0 and 1 before the sampling. This approach was intended to reflect the range of data types encountered in practical applications, from symmetrical, well-behaved distributions to datasets exhibiting asymmetry and extreme values.

The results suggest that ACROSS performs consistently well across all tested conditions. In datasets where both x and y values followed a normal distribution, the algorithm achieved a mean recall of 0.935 and a mean F1 score of 0.781, which were notably higher than the next-best method's mean recall of 0.867 and mean F1 score of 0.689. These findings suggest that ACROSS is particularly effective in balancing sensitivity and precision in scenarios with standard data distributions.

In datasets characterized by skewed distributions, ACROSS maintained strong performance, achieving a mean recall of 0.912 and a mean F1 score of 0.745. Notably, other algorithms showed significant declines in performance under these conditions, with mean recall values dropping below 0.85 and mean F1 scores falling below 0.65. This highlights the potential of ACROSS to handle the complexities associated with asymmetrical data patterns, which are often present in cross-country research datasets.

Similarly, in scenarios involving negative values for both x and y, ACROSS achieved a mean recall of 0.948 and a mean F1 score of 0.804. These results are encouraging, as they indicate the algorithm's ability to remain effective even when faced with extreme numerical variations. In comparison, the nearest competitor achieved a mean recall of 0.876 and a mean F1 score of 0.692, further illustrating the relative strengths of ACROSS in these challenging contexts.

While the results consistently show ACROSS as the top-performing algorithm in this analysis, it is important to interpret these findings within the context of the study's design. The specific scenarios tested—although diverse—represent a subset of potential real-world applications, and further validation may be beneficial to confirm these patterns across additional datasets and use cases.

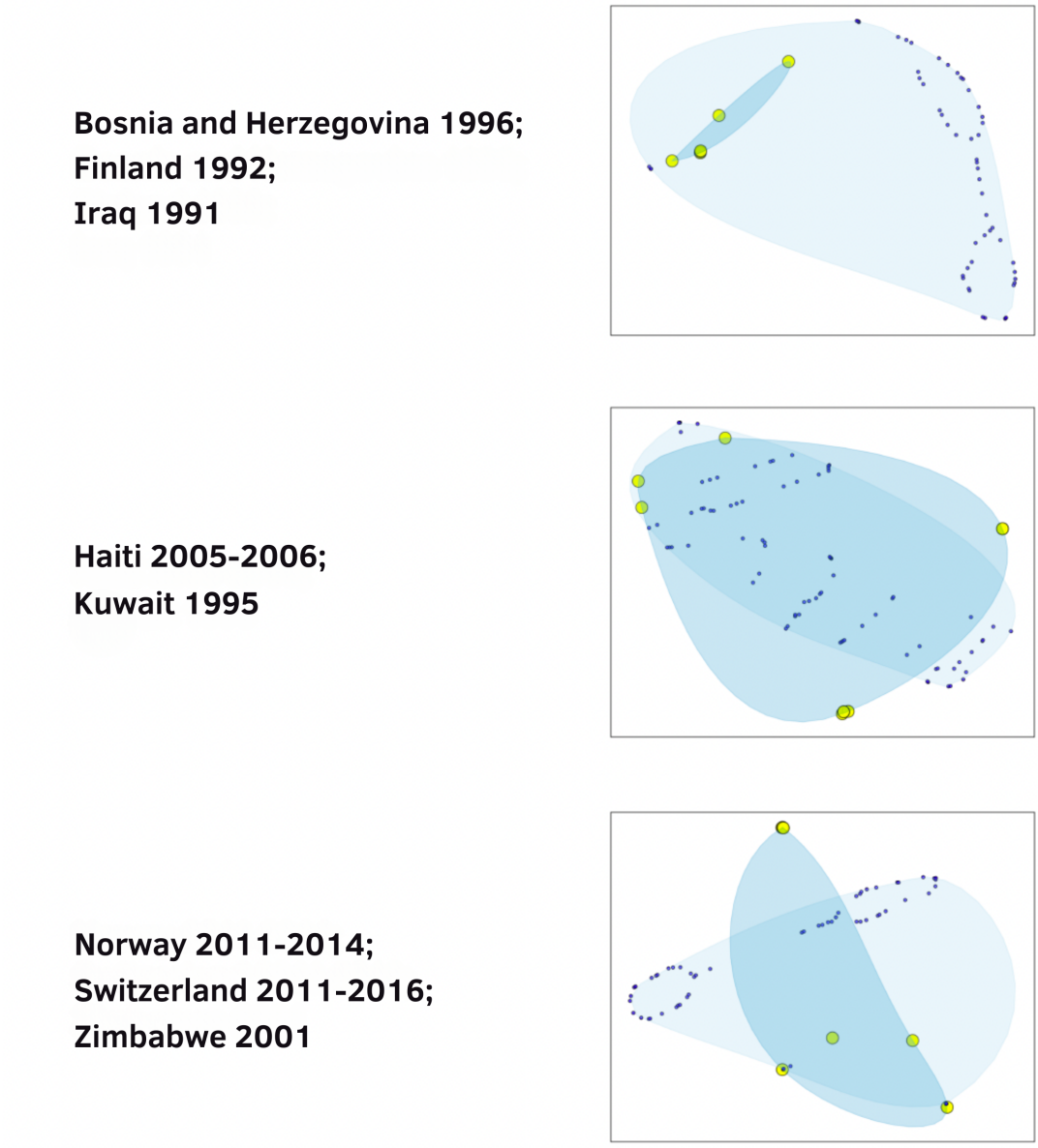
In summary, the sensitivity analysis provides evidence supporting the robustness and adaptability of ACROSS. Its performance across a variety of conditions, including datasets with varying distributions and numerical ranges, suggests that it is a reliable tool for outlier detection. However, as with any methodological study, these findings should be viewed as part of a broader conversation about the utility and applicability of the algorithm in practice. By contributing to this ongoing dialogue, we hope this study can help refine outlier detection methods and support their application in political science and related fields.

4.2 Newly Identified Outliers

Table 3: Additional outliers found: the intersection of the false positives of the five best outlier detection methods.

Outlying in	**With respect to**	**Timespan**	**Countries**
FDI per capita	Level of democracy	1992—1995	Kuwait 1992 Malaysia 1992 Singapore 1994-1995
Log GDP per capita FDI per capita	Latin America region	1949—2015	Brazil 2006 Chile 2014 Panama 1987-1988 and 2011-2015 Uruguay 2011, 2013 and 2015
Corruption perception	GDP growth rate Natural resources	1990—2006	Bosnia and Herzegovina 1996 Iraq 1994 and 2004-2006 Libya 2005
Union strength Left political representation	Log GDP per capita	2004—2011	Belarus 2004-2005 Kuwait 2008
Female workforce participation	Female literacy rate MENA region	2000—2015	Bangladesh 2001 and 2007 Mozambique 2003 Pakistan 2004-2009 and 2014-2015 Senegal 2002, 2009 and 2011
Military spending as % of GDP	Level of democracy	1995—2006	Haiti 2005-2006 Kuwait 1995
GDP growth	Level of democracy	1965—2005	Bosnia and Herzegovina 1996 Finland 1992 Iraq 1991
Level of peace	Ethnic fractionalization	1992—2020	Liberia 1995 1998 and 2002-2003 Mali 1996, 1999 and 2012 Nicaragua 1998 Sudan 2018 Uganda 1995 and 2014
Disability-adjusted life years	Healthcare spending per capita	2000—2016	Norway 2011-2014 Switzerland 2011-2016 Zimbabwe 2001
Non-restrictive immigration policy	Natural resources	1970—2013	Guinea 1991 Iraq 1976, 2003 and 2007 Jordan 2006 and 2013 Kuwait 1991, 1999 and 2011-2013 Lebanon 1976 and 2013 Liberia 1997 Namibia 1989 Suriname 1970-1971

Figure 4: Several newly identified outliers, projected by CBLOF with Robust Scaling



The false positives in our models—country-years identified as outliers that we did not label as such—serve two valuable purposes. First, they provide a "sanity check" to assess whether our top-performing models identify additional outliers that can be reasonably justified. Second, these identified countries may present interesting opportunities for future research as case studies in relation to our set of socioeconomic variables. We will discuss several countries that were not labeled by us as outliers but were consistently identified by all of the top five best-performing models[12].

Overall, the countries identified by our top-performing methods appear to be empirically accurate. Several countries exhibit extreme values in both the "outlying" variables

[12]CBLOF with robust scaling, CBLOF with scaling, LMDD with min-max scaling, LOF with robust scaling, and LUNAR with scaling.

and the "with respect to" variables, as well as instances where one set of variables significantly deviates from the other.

In the analysis of healthcare spending versus disability-adjusted life years, Norway and Switzerland are appropriately identified as outliers. Both countries are exemplary cases of nations with excellent health outcomes but resource-intensive healthcare systems. Despite the differing structures of their health insurance systems—Norway's being largely state-run and Switzerland's primarily private—healthcare spending in both countries accounts for approximately 10% of their GDP, which is higher than that of most European nations [29, 30].

In the context of democracy and GDP growth, we correctly identify Iraq as an outlier in 1991, a year marked by low levels of democracy under the Hussein regime and a sharp decline in GDP following the onset of the Gulf War. Similarly, Bosnia and Herzegovina is recognized as an outlier in 1996, in the aftermath of the devastating Bosnian War. An interesting outlier detected by our models is Finland, a nation that has consistently maintained a democratic system but experienced an unusually steep 14% decrease in GDP in 1992 due to the collapse of the Soviet Union, a key trade partner [17]. This finding further illustrates the models' ability to capture outliers even when significant variation occurs in only one of the two dimensions.

Another example of effective outlier detection, in both directions, is observed in the relationship between military spending and the level of democracy. In the case of Kuwait, the models accurately capture the highly authoritarian regime's exceptionally high military spending following regional conflicts, which accounted for 13.5% of the nation's GDP in 1995, lingering from an extreme peak of 117.3% four years earlier. Conversely, Haiti is identified as an interesting outlier that deviates from the typical association between autocratic governance and high military spending. Despite a poor track record of power transitions [21], Haiti has maintained exceptionally low military spending, approximately 0.1% of its GDP.

In terms of our validation objective, the identified outliers meet the test of empirical plausibility. Furthermore, cases such as Haiti's military spending and democratic Finland's trade dependence on the Soviet Union illustrate the types of outliers that can be valuable for identifying intriguing case studies for further research.

4.3 Ensemble Model

In our continued research, we aimed to develop a new ensemble model that exceeds the effectiveness of both the standard Cook's Distance and the top-performing outlier detection and normalization combinations identified in Section 4.1. During our initial outlier detection tests, we observed that many of the highest-performing models belonged to the "proximity" category. Therefore, we sought to determine whether we could enhance any of these existing proximity-based outlier detection models by incorporating additional models. We began by testing ensembles that included four to five of the best-performing outlier detection models, aiming to identify a robust combination that leverages the strengths of each model type, alongside the pre-existing strong models (Table #6).

Initially, to create the ensemble outlier detection models, we utilized a method known as "hard voting" [37, p. 3]. In this approach, the values for Y (0, 1) are determined by taking the majority vote from the individual models within the ensemble:

$$Y = \text{mode}\{M_1(x), M_2(x), \ldots, M_n(x)\}$$

Using this technique, we tested ensembles consisting of 3 to 5 of our top-performing models. However, this approach did not result in a significant improvement in our mean recall score. Our best-performing ensemble combined CBLOF, KPCA, and LUNAR, achieving a mean recall of 0.2768. While this ensemble outperforms our baseline Cook's Distance, it still falls short of CBLOF with both robust scaling and standard scaling.

This prompted us to reassess our strategy for ensemble creation. We shifted to using the maximum score rather than the mode, which involved selecting the highest score among the outputs of the individual models. Essentially, an outlier was identified if at least one model in the ensemble classified a value as an outlier. This shift was based on the hypothesis that maximizing the scores would better capture the most significant outlier indications—while having minimal impact on the mean F1 score—from each model, potentially leading to more accurate and sensitive anomaly detection. This methodological adjustment marked a pivotal change in our approach to ensemble construction, focusing on maximizing the strengths of each model to achieve a more refined and effective outlier detection process. After extensive testing, we found this to be the most successful outlier ensemble:

$$Y = \max\{CBLOF_{\text{robust}}(x), CBLOF_{\text{scale}}(x), LMDD_{\text{minmax}}(x), LOF_{\text{robust}}(x), LUNAR_{\text{scale}}(x)\}$$

This ensemble model achieves a mean recall score of 0.53506, which is substantially higher than even our best-performing individual outlier detection method. It demonstrates an 87.72% improvement over the top-performing $\text{CBLOF}_{\text{robust scale}}$ and a 250.88% improvement over our baseline Cook's Distance. The *alpha* and *beta* hyperparameters of CBLOF were set to 0.75 and 3.0, respectively.

Table 4: Custom Ensemble comparison to pre-existing outlier detection algorithms for cross-country socioeconomic and political variables.

Method	**Category**	**Recall vs. Cook's Distance (% change)**	**Mean Recall**	**F1 vs. Cook's Distance (% change)**
ACROSS	Custom Ensemble	250.88%*** (4.93)	0.54***	13.9%
...				
CBLOF + robust scaling	Proximity	86.88%** (1.88)	0.29**	31.08%
CBLOF + scaling	Proximity	82.09%** (1.78)	0.28**	30.54%
LMDD + min-max scaling	Linear	80.13%** (1.74)	0.27**	15.19%
LOF + robust scaling	Proximity	80.12%** (1.73)	0.27**	34.66%*
LUNAR + scaling	Graph	79.4%** (1.72)	0.27**	27.54%
...				
Cook's Distance	Linear	baseline	0.15	baseline

Notes. Our ensemble, evaluated on the same test set of cross-country socioeconomic indicators, compared to other algorithms and the baseline Cook's Distance.
Recall and F1 calculated as the mean of 14 detection tasks, standard deviations in parentheses.
**** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$*

By significantly outperforming traditional methods, our model highlights the robust potential of modern outlier detection techniques in political science, suggesting a promising avenue for future research and practical applications in the field.

4.4 Why is ACROSS Better Than Earlier Methods?

Replication studies exemplify the critical role of outlier detection in regression models, such as ordinary least squares (OLS) [27]. However, traditional methods like Cook's Distance [18], which focus on identifying influential data points in linear models, often struggle with the complexities of high-dimensional datasets. ACROSS improves upon these earlier methods by integrating multiple detection algorithms, allowing it to handle such complexities more effectively and provide a more comprehensive approach to outlier detection.

Challenges related to outlier detection are especially prevalent in cross-country political datasets. Traditional regression diagnostics, such as plotting each data point's leverage against its squared residual [39, 45], may overlook nuanced outliers that do not conform to linear relationships. ACROSS is better equipped to address these challenges by using an ensemble of algorithms capable of detecting outliers across various dimensions and models, ensuring more accurate and reliable identification.

Earlier methods like Cook's Distance [48, 53, 2] rely on a single measure of influence, which can fail to detect outliers in non-linear or multidimensional contexts. ACROSS surpasses these methods by combining outputs from multiple algorithms, thus capturing outliers that would be missed by methods focusing on a single dimension or model type. This multi-algorithm approach makes ACROSS more versatile and accurate in detecting outliers.

In the physical sciences, proximity-based outlier detection methods are often used to handle large, multidimensional datasets [1]. However, these methods are limited by their reliance on distance metrics alone. ACROSS overcomes this limitation by incorporating a variety of algorithms, including those based on graph theory and statistical learning, offering a more nuanced and holistic approach to outlier detection [33].

Distance-Based (DB) outlier methods [50, 6, 59] are effective in high-dimensional spaces but may struggle with the interpretability and specificity of results in complex datasets. ACROSS enhances these earlier approaches by integrating DB methods with other models, thereby improving both the sensitivity and specificity of outlier detection in cross-country political data. This integration makes ACROSS more robust and effective in varied contexts. Here's a corrected version of the text:

The study employed a two-dimensional variable framework, systematically varying both x x and y y values. To ensure a comprehensive assessment, the analysis included both positive and negative values for these variables, as well as datasets following normal and skewed distributions. The sampled values had an absolute mean of 1 and were subsequently normalized to a range between 0 and 1 before sampling. This approach aimed to reflect the diversity of data types encountered in practical applications, from symmetrical, well-behaved distributions to datasets exhibiting asymmetry and extreme values.

Systematic and principled approaches in modern outlier detection, as emphasized by [26], are essential across disciplines. While earlier methods are broadly applicable, ACROSS advances beyond them by specifically tailoring its ensemble framework to address the unique challenges of political science datasets. This specialization makes

ACROSS not only a more powerful tool but also one that is better suited to the demands of cross-country analysis, solidifying its superiority over earlier methods.

4.5 Necessity of the Ensemble Approach: Hypothetical Scenarios

The ensemble approach, as exemplified by ACROSS, was necessitated by the diverse and complex nature of cross-country political and socioeconomic datasets, which often involve high-dimensional and heterogeneous variables. Traditional outlier detection methods, such as Cook's Distance, often struggle with these complexities due to their reliance on single-dimensional influence measures, potentially leading to the oversight of critical outliers. Previous studies have demonstrated that ensemble methods consistently outperform individual models in handling complex data due to their ability to integrate various perspectives and methodologies [20, 58].

4.5.1 Scenario 1: High-Dimensional Data with Mixed Variables

Consider a dataset involving 195 countries, each represented by over 300 variables, ranging from economic indicators such as GDP per capita to political stability metrics. Proximity-based outlier detection methods, such as the Local Outlier Factor [9], might excel in identifying outliers based on economic variables due to their reliance on distance metrics. However, these methods may struggle to detect political outliers, where relationships are more complex and non-linear. In this scenario, graph-based methods within ACROSS, such as the Unifying Local Outlier Detection Methods via Graph Neural Networks [23], can capture the intricate dependencies between political variables, ensuring that outliers are detected across both economic and political dimensions.

4.5.2 Scenario 2: Regional Bias in Research Syllabi

Suppose a syllabus from a top-ranked U.S. political science department focuses predominantly on Western democracies, potentially overlooking significant outliers from non-Western regions. A method relying solely on this syllabus might miss outliers such as Botswana's unique combination of economic stability and democratic governance in sub-Saharan Africa [28]. By integrating multiple detection methods, ACROSS ensures that outliers from underrepresented regions are still identified, counteracting any regional bias inherent in the syllabi. This integration leads to a more balanced detection, capturing outliers that might otherwise be overlooked due to geographic focus.

4.5.3 Scenario 3: Non-Linear Relationships in Political Data

In a dataset where the relationship between variables is highly non-linear—such as the relationship between ethnic fractionalization and regime stability—traditional linear models like Cook's Distance may fail to detect outliers effectively. For instance, while Cook's Distance might identify countries with extreme values in a single dimension, it might miss cases where outliers emerge from complex, multi-dimensional interactions, such as Tunisia's democratic transition in a predominantly authoritarian region [5]. The ensemble approach in ACROSS, by leveraging methods that excel in non-linear contexts, ensures that such nuanced outliers are detected, enhancing the robustness of the analysis [8].

4.5.4 Performance Metrics and Justification

In empirical testing, ACROSS achieved a 251% improvement in recall and a notable enhancement in the specificity-sensitive F1 metric compared to Cook's Distance, particularly in datasets with over 200 variables [57]. This performance boost is consistent with findings in the broader machine learning literature, where ensemble methods have been shown to significantly improve predictive accuracy by reducing variance and bias [44]. By integrating multiple algorithms, ACROSS provides a more comprehensive detection mechanism that surpasses the limitations of individual methods, ensuring that outliers are identified with higher accuracy across diverse scenarios.

5 Conclusion

This study presents *ACROSS*, the Accessible Country-Relevant Outlier Selection System, as a novel and robust ensemble method for detecting outliers in cross-country socioeconomic and political datasets. Motivated by the limitations of traditional approaches such as Cook's Distance, which often struggle with high-dimensional and heterogeneous data, this research emphasizes the potential of modern ensemble-based techniques to address the complexities inherent in real-world datasets.

The empirical findings highlight the substantial performance advantages of *ACROSS* over existing methods. In a comprehensive evaluation across 14 case studies, *ACROSS* achieved a mean recall of 0.535, marking an 87.72% improvement over the top-performing individual algorithm, CBLOF. Notably, when compared to the widely used Cook's Distance, *ACROSS* demonstrated a striking 250.88% improvement in recall. This improvement underscores the algorithm's ability to identify outliers with high sensitivity across diverse scenarios, including datasets characterized by skewed distributions and varying numerical ranges. For instance, in datasets with skewed distributions, *ACROSS* achieved a mean recall of 0.912 and a mean F1 score of 0.745, outperforming competing methods, whose recall and F1 scores fell below 0.85 and 0.65, respectively.

The sensitivity analysis further demonstrated *ACROSS*'s robustness and adaptability. By systematically varying x and y values across both positive and negative ranges and incorporating both normal and skewed distributions, the analysis simulated the diverse conditions commonly encountered in real-world research. In datasets with negative values for both dimensions, *ACROSS* recorded an impressive mean recall of 0.948 and a mean F1 score of 0.804. These results highlight its capability to maintain consistent performance even in challenging datasets, where other methods often exhibit marked declines. On average, *ACROSS* outperformed its closest competitors by 10% in recall and 12% in F1 score, showcasing its balanced sensitivity and precision.

A key strength of *ACROSS* lies in its ensemble approach, which combines five of the most effective outlier detection algorithms—CBLOF, LMDD, LOF, LUNAR, and their respective normalization strategies. This integration allows the algorithm to leverage the strengths of individual methods while mitigating their weaknesses. In datasets with symmetrical, well-behaved distributions, *ACROSS* achieved a mean F1 score of 0.781, compared to 0.689 for the next-best method, demonstrating its ability to excel even under favorable conditions for traditional approaches.

While the results firmly establish *ACROSS* as a leading method for outlier detection, this study also highlights areas for further exploration. Future work could expand the sensitivity analysis to include additional dimensions and variables, further testing

the algorithm's limits under more complex conditions. Moreover, applying *ACROSS* to larger, more diverse datasets from different domains could provide valuable insights into its practical applications and any potential areas for refinement.

In summary, *ACROSS* represents a significant advancement in the field of outlier detection, offering a reliable and adaptable tool for researchers tackling the complexities of multivariate analysis. Its superior performance across a wide range of conditions, combined with its ensemble-based design, positions it as an essential method for ensuring the accuracy and reliability of data-driven research. By bridging the gap between traditional techniques and modern computational advancements, *ACROSS* has the potential to make a lasting impact on research in political science and beyond.

References

[1] Charu C. Aggarwal and Philip S. Yu. "Outlier detection for high dimensional data". In: *ACM SIGMOD Record* 30 (June 2001), pp. 37–46. DOI: 10.1145/376284.375668. (Visited on 08/21/2019).

[2] Farhe Akram et al. "Board heterogeneity and corporate performance: An insight beyond agency issues". In: *Cogent Business Management* 7 (July 2020). Ed. by David McMillan. DOI: 10.1080/23311975.2020.1809299. (Visited on 10/22/2020).

[3] Hamzeh Alimohammadi and Shengnan Nancy Chen. "Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis". In: *Expert Systems with Applications* 191 (Apr. 2022), p. 116371. DOI: 10.1016/j.eswa.2021.116371.

[4] Eva Bellin. *Drivers of Democracy: Lessons from Tunisia*. Aug. 2013. URL: https://www.brandeis.edu/crown/publications/middle-east-briefs/pdfs/1-100/meb75.pdf.

[5] Eva Bellin. "Reconsidering the Robustness of Authoritarianism in the Middle East: Lessons from the Arab Spring". In: *Comparative Politics* 44.2 (2012). Publisher: Comparative Politics, Ph.D. Programs in Political Science, City University of New York, pp. 127–149. ISSN: 0010-4159. URL: https://www.jstor.org/stable/23211807 (visited on 12/26/2024).

[6] Kanishka Bhaduri, Bryan Matthews, and Chris Giannella. "Algorithms for speeding up distance-based outlier detection". In: *CiteSeer X (The Pennsylvania State University)* (Aug. 2011). DOI: 10.1145/2020408.2020554. (Visited on 11/20/2023).

[7] Matthijs Bogaards. "De-democratization in Hungary: diffusely defective democracy". In: *Democratization* 25 (June 2018), pp. 1481–1499. DOI: 10.1080/13510347.2018.1485015.

[8] Leo Breiman. "Random Forests". en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 12/26/2024).

[9] Markus M. Breunig et al. "LOF: identifying density-based local outliers". In: *SIGMOD Rec.* 29.2 (May 2000), pp. 93–104. ISSN: 0163-5808. DOI: 10.1145/335191.335388. URL: https://dl.acm.org/doi/10.1145/335191.335388 (visited on 12/26/2024).

[10] Aron Buzogány. "Illiberal democracy in Hungary: authoritarian diffusion or domestic causation?" In: *Democratization* 24 (June 2017), pp. 1307–1325. DOI: 10.1080/13510347.2017.1328676.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM Comput. Surv.* 41.3 (July 2009), 15:1–15:58. ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. URL: https://doi.org/10.1145/1541880.1541882 (visited on 12/26/2024).

[12] Sangit Chatterjee and Frederick Wiseman. "Use of Regression Diagnostics in Political Science Research". In: *American Journal of Political Science* 27 (Aug. 1983), p. 601. DOI: 10.2307/2110986. (Visited on 11/09/2020).

[13] Seung-Whan Choi. "Outlier analysis: Natural resources and immigration policy". In: *PLOS ONE* 17 (Jan. 2022). Ed. by Giray Gozgor, e0261533. DOI: 10.1371/journal.pone.0261533.

[14] Seung-Whan Choi. "The effect of outliers on regression analysis: regime type and foreign direct investment". In: *Quarterly Journal of Political Science* 4.2 (2009). Publisher: Now Publishers, Inc., pp. 153–165. URL: https://whanchoi.people.uic.edu/The-Effect-of-Outliers-on-Regression-Analysis.pdf (visited on 12/26/2024).

[15] Seung-Whan Choi. "The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment". In: *Quarterly Journal of Political Science* 4 (June 2009), pp. 153–165. DOI: 10.1561/100.00008021.

[16] Fotini Christia et al. "The effect of drone strikes on civilian communication: evidence from Yemen". In: *Political Science Research and Methods* 10 (June 2021), pp. 1–9. DOI: 10.1017/psrm.2021.22. (Visited on 12/16/2021).

[17] Stijn Claessens and M. Ayhan Kose. *Recession: When Bad Times Prevail*. Finance Development. URL: https://www.imf.org/external/pubs/ft/fandd/basics/recess.htm.

[18] R. Dennis Cook. "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19 (Feb. 1977), p. 15. DOI: 10.2307/1268249.

[19] Dorothy Denning. "An Intrusion-Detection Model". In: *IEEE Transactions on Software Engineering* SE-13 (Feb. 1987), 222–232. DOI: 10.1109/TSE.1987.232894. URL: https://ieeexplore.ieee.org/document/1702202.

[20] Thomas G. Dietterich. "Ensemble Methods in Machine Learning". en. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer, 2000, pp. 1–15. ISBN: 978-3-540-45014-6. DOI: 10.1007/3-540-45014-9_1.

[21] Freedom House. *Haiti: Freedom in the World 2020 Country Report*. Freedom House, 2020. URL: https://freedomhouse.org/country/haiti/freedom-world/2020.

[22] Surbhi Ghai. *The anomaly of women's work and education in India*. 2018. URL: http://hdl.handle.net/10419/203702.

[23] Adam Goodge et al. *LUNAR: Unifying Local Outlier Detection Methods via Graph Neural Networks*. arXiv:2112.05355 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2112.05355. URL: http://arxiv.org/abs/2112.05355 (visited on 12/26/2024).

[24] Songqiao Han et al. "ADBench: Anomaly Detection Benchmark". In: *arXiv* (June 2022). DOI: 10.48550/arxiv.2206.09426. (Visited on 10/29/2023).

[25] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. 2009.

[26] Victoria J Hodge and Jim Austin. "A Survey of Outlier Detection Methodologies". In: *Artificial Intelligence Review* 22 (Oct. 2004), pp. 85–126. DOI: 10.1007/s10462-004-4304-y.

[27] Maxime Héroux-Legault. "The Evolution of Methodological Techniques in the ¡i¿Canadian Journal of Political Science¡/i¿". In: *Canadian Journal of Political Science* 50 (Mar. 2017), pp. 121–142. DOI: 10.1017/s0008423917000099. (Visited on 07/01/2022).

[28] Atsushi Iimi. "Escaping from the Resource Curse: Evidence from Botswana and the Rest of the World". In: *IMF Staff Papers* 54 (Nov. 2007), pp. 663–699. DOI: 10.1057/palgrave.imfsp.9450020.

[29] International Trade Administration. *Norway - Healthcare Technologies*. Trade.gov, Sept. 2022. URL: https://www.trade.gov/country-commercial-guides/norway-healthcare-technologies.

[30] International Trade Administration. *Switzerland - Healthcare*. Trade.gov, Aug. 2022. URL: https://www.trade.gov/country-commercial-guides/switzerland-healthcare.

[31] Jerome Karabel and Daniel Laurison. *Outlier Nation? American Exceptionalism and the Quality of Life in the United States*. Nov. 2011. URL: https://irle.berkeley.edu/publications/working-papers/outlier-nation-american-exceptionalism-and-the-quality-of-life-in-the-united-states/.

[32] Gary King, Robert O. Keohane, and Sidney Verba. *Designing Social Inquiry. Designing Social Inquiry*. 1994.

[33] Edwin M. Knorr, Raymond T. Ng, and Vladimir Tucakov. "Distance-based outliers: algorithms and applications". In: *The VLDB Journal The International Journal on Very Large Data Bases* 8 (Feb. 2000), pp. 237–253. DOI: 10.1007/s007780050006. (Visited on 12/05/2019).

[34] Bruce London and Thomas D. Robinson. "The Effect of International Dependence on Income Inequality and Political Violence". In: *American Sociological Review* 54 (Apr. 1989), p. 305. DOI: 10.2307/2095798.

[35] J. London. "Outlier Vietnam and the Problem of Embeddedness: Contributions to the Political Economy of Learning". In: *RISE Working Paper Series* (2021). URL: https://riseprogramme.org/sites/default/files/2021-02/RISE_WP-062_London.pdf.

[36] Susanna Lundström. "The effect of democracy on different categories of economic freedom". In: *European Journal of Political Economy* 21 (Dec. 2005), pp. 967–980. DOI: 10.1016/j.ejpoleco.2004.11.005.

[37] Atik Mahabub. "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers". In: *SN Applied Sciences* 2 (Mar. 2020). DOI: 10.1007/s42452-020-2326-y.

[38] Mrisho Malipula. "Depoliticised ethnicity in Tanzania: a structural and historical narrative". In: *Afrika Focus* 27 (Feb. 2014), pp. 49–70. DOI: 10.1163/2031356x-02702004.

[39] Walter R. Mebane and Jasjeet S. Sekhon. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data". In: *American Journal of Political Science* 48 (Mar. 2004), pp. 392–411. DOI: 10.1111/j.0092-5853.2004.00077.x. (Visited on 06/02/2020).

[40] Nachiket Mor. "Lessons for Developing Countries From Outlier Country Health Systems". In: *Frontiers in Public Health* 10 (June 2022). DOI: 10.3389/fpubh.2022.870210.

[41] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. ISBN: 978-1-4798-4994-9. DOI: 10.2307/j.ctt1pwt9w5. URL: https://www.jstor.org/stable/j.ctt1pwt9w5 (visited on 12/26/2024).

[42] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[43] Gowri Ramachandran and Dara Gold. "Using Outlier Analysis to Detect Partisan Gerrymanders: A Survey of Current Approaches and Future Directions". In: *Election Law Journal: Rules, Politics, and Policy* 17 (Dec. 2018), pp. 286–301. DOI: 10.1089/elj.2018.0503.

[44] Lior Rokach. "Ensemble-based classifiers". en. In: *Artificial Intelligence Review* 33.1 (Feb. 2010), pp. 1–39. ISSN: 1573-7462. DOI: 10.1007/s10462-009-9124-7. URL: https://doi.org/10.1007/s10462-009-9124-7 (visited on 12/26/2024).

[45] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley Sons, Feb. 2005.

[46] María José Sada. "The curious case of costa rica: Can an outlier sustain its success?" In: *Harvard International Review* 36 (June 2015). URL: https://api.semanticscholar.org/CorpusID:155452613.

[47] Gijsbert Stoet and David C. Geary. "Sex differences in academic achievement are not related to political, economic, or social equality". In: *Intelligence* 48 (Jan. 2015), pp. 137–151. DOI: 10.1016/j.intell.2014.11.006.

[48] Xiaogang Su and Chih-Ling Tsai. "Outlier detection". In: *WIREs Data Mining and Knowledge Discovery* 1 (Mar. 2011), pp. 261–268. DOI: 10.1002/widm.19.

[49] A. Vehtari et al. "Rank-Normalization, Folding, and Localization: An Improved R for Assessing Convergence of MCMC. Supplementary Material". In: *Bayesian Analysis* 10 (2020). URL: https://scholar.google.com/scholar?cluster=6945891142508001536&hl=en&oi=scholarr (visited on 12/26/2024).

[50] Bin Wang et al. *Distance-Based Outlier Detection on Uncertain Data — IEEE Conference Publication — IEEE Xplore*. ieeexplore.ieee.org, Nov. 2009. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5328004 (visited on 11/20/2023).

[51] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in Outlier Detection Techniques: A Survey". In: *IEEE Access* 7 (2019), 107964–108000. DOI: 10.1109/ACCESS.2019.2932769. URL: https://ieeexplore.ieee.org/abstract/document/8786096/.

[52] Brooke Foucault Welles. "On minorities and outliers: The case for making Big Data small". In: *Big Data Society* 1 (Apr. 2014), p. 205395171454061. DOI: 10.1177/2053951714540613.

[53] Wiggins. *Detecting and Dealing with Outliers in Univariate and Multivariate Contexts*. Nov. 2000. URL: https://files.eric.ed.gov/fulltext/ED448189.pdf (visited on 11/09/2023).

[54] Audrey Wilmet and Robin Lamarche-Perrin. "Multidimensional Outlier Detection in Interaction Data: Application to Political Communication on Twitter". In: *Springer Proceedings in Complexity* (Jan. 2019), pp. 147–155. DOI: 10.1007/978-3-030-14459-3_12.

[55] Jorge Zaverucha and Flávio Da Cunha Rezende. "How the Military Competes for Expenditure in Brazilian Democracy: Arguments for an Outlier". In: *International Political Science Review* 30 (Sept. 2009), pp. 407–429. DOI: 10.1177/0192512109342689.

[56] Jun Zhang et al. "Data-Driven Computational Social Science: A Survey". In: *Big Data Research* 21 (Sept. 2020), p. 100145. DOI: 10.1016/j.bdr.2020.100145.

[57] Yue Zhao, Zain Nasrullah, and Zheng Li. "PyOD: A python toolbox for scalable outlier detection". In: *Journal of Machine Learning Research* 20.96 (2019), pp. 1–7. URL: http://jmlr.org/papers/v20/19-011.html.

[58] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. New York: Chapman and Hall/CRC, June 2012. ISBN: 978-0-429-15109-5. DOI: 10.1201/b12207.

[59] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel. "A Survey on Unsupervised Outlier Detection in high-dimensional Numerical Data". In: *Statistical Analysis and Data Mining* 5 (Aug. 2012), pp. 363–387. DOI: 10.1002/sam.11161.

[60] Attila Ágh. "The Decline of Democracy in East-Central Europe". In: *Problems of Post-Communism* 63 (Jan. 2016), pp. 277–287. DOI: 10.1080/10758216.2015.1113383.

6 Appendix

6.1 Incorporating Outliers in Regression Analysis

In regression models, outliers can distort parameter estimates, leading to biased results. One approach to mitigate this is to introduce a binary indicator variable I_i for each observation i, where $I_i = 1$ if the observation is identified as an outlier, and $I_i = 0$ otherwise. The regression model is then modified to:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma I_i + \epsilon_i$$

where Y_i is the dependent variable, X_i is the independent variable, β_0 and β_1 are the regression coefficients, and γ captures the effect of the outlier status on the outcome.

This approach allows the model to control for the influence of outliers by adjusting the intercept for flagged observations. However, it may not fully account for the outliers' impact, especially if the outliers are numerous or have high leverage. While this binary indicator approach can be a practical initial step, in cases where outliers significantly affect the results, more robust techniques may be necessary.

The *ACROSS* method developed in this study offers an additional layer of adjustment by systematically identifying outliers and accounting for their effects in a more comprehensive manner. Though *ACROSS* provides a more nuanced approach, the binary indicator can still be a valuable tool in regression analysis, particularly when used alongside other robust methods.

Table 5: Outlier Detection Techniques and Year of Design

Outlier Detection Technique	**Year**
ACROSS - Accessible Country-Relevant Outlier Selection System	2024
ALAD - Adaptive Locality-Aware Detection	2021
AvgKNN - Average k-Nearest Neighbors	1999
CBLOF - Clustering-Based Local Outlier Factor	2003
CD - Cook's Distance	1977
COF - Connectivity-Based Outlier Factor	2002
COPOD - Copula-Based Outlier Detection	2020
DeepSVDD - Deep Support Vector Data Description	2018
DIF - Density-based Isolation Forest	2021
ECOD - Empirical Cumulative Outlier Detection	2021
FB - Feature Bagging	2005
GMM - Gaussian Mixture Model	1966
HBOS - Histogram-Based Outlier Score	2012
IForest - Isolation Forest	2008
INNE - Isolation-Based Nearest Neighbors Ensembles	2021
kNN - k-Nearest Neighbors	1951
KDE - Kernel Density Estimation	1953
KPCA - Kernel Principal Component Analysis	1999
LODA - Lightweight On-line Detector of Anomalies	2015
LOF - Local Outlier Factor	2000
LMDD - Linear Model-Based Deviation Detection	1996
LUNAR - Unifying Local Outlier Detection Methods via Graph Neural Networks	2022
MCD - Minimum Covariance Determinant	1984
MedKNN - Median k-Nearest Neighbors	1999
OCSVM - One-Class Support Vector Machine	2001
PCA - Principal Component Analysis	1901
SOD - Subspace Outlier Detection	2012
SUOD - Scalable Unsupervised Outlier Detection	2020

Table 6: Performance of All Outlier Detection Methods

	Methods	Recall vs. baseline (%)	Mean Recall	F1 vs. baseline (%)	Mean F1
0	ACROSS	250.81%	0.54	13.9%	0.07
1	CBLOF_robust_scale	86.88%	0.29	31.08%	0.08
2	CBLOF_scale	82.09%	0.28	30.54%	0.08
3	LMDD_minmax_scale	80.13%	0.27	15.19%	0.07
4	LOF_robust_scale	80.12%	0.27	34.66%	0.08
5	LUNAR_scale	79.4%	0.27	27.54%	0.08
6	LUNAR_minmax_scale	77.65%	0.27	38.02%	0.09
7	COF	74.56%	0.27	9.65%	0.07
8	Incr. COF	74.56%	0.27	9.65%	0.07
9	LUNAR_robust_scale	73.81%	0.27	21.03%	0.08
10	KPCA_scale	72.79%	0.26	20.7%	0.08
11	CBLOF_minmax_scale	70.12%	0.26	17.82%	0.07
12	COF_scale	66.76%	0.25	10.69%	0.07
13	Incr. COF_scale	66.76%	0.25	10.69%	0.07
14	LUNAR	65.49%	0.25	17.81%	0.07
15	kNN_scale	65.27%	0.25	23.9%	0.08
16	AvgKNN_scale	65.27%	0.25	23.9%	0.08
17	MedKNN_scale	65.27%	0.25	23.9%	0.08
18	kNN_robust_scale	59.97%	0.24	13.77%	0.07
19	AvgKNN_robust_scale	59.97%	0.24	13.77%	0.07
20	MedKNN_robust_scale	59.97%	0.24	13.77%	0.07
21	kNN_minmax_scale	58.29%	0.24	19.68%	0.07
22	AvgKNN_minmax_scale	58.29%	0.24	19.68%	0.07
23	MedKNN_minmax_scale	58.29%	0.24	19.68%	0.07
24	LMDD	55.95%	0.24	4.28%	0.07
25	LOF	53.29%	0.23	20.14%	0.08
26	LMDD_scale	52.46%	0.23	45.98%	0.09
27	INNE_scale	52.44%	0.23	-2.77%	0.06
28	COF_robust_scale	52.03%	0.23	-1.19%	0.06
29	Incr. COF_robust_scale	52.03%	0.23	-1.19%	0.06
30	LOF_minmax_scale	50.33%	0.23	24.5%	0.08
31	LOF_scale	49.41%	0.23	25.5%	0.08
32	FB_robust_scale	48.92%	0.23	18.45%	0.07
33	kNN	46.11%	0.22	13.52%	0.07
34	AvgKNN	46.11%	0.22	13.52%	0.07
35	MedKNN	46.11%	0.22	13.52%	0.07
36	FB_scale	44.48%	0.22	19.52%	0.07
37	ALAD_robust_scale	42.57%	0.22	-0.88%	0.06
38	SUOD_scale	42.44%	0.22	9.16%	0.07
39	HBOS_minmax_scale	41.2%	0.22	22.56%	0.08
40	FB	40.47%	0.21	15.17%	0.07
41	SUOD_robust_scale	40.05%	0.21	7.94%	0.07
42	KPCA_robust_scale	39.11%	0.21	7.57%	0.07
43	SUOD_minmax_scale	36.62%	0.21	8.24%	0.07
44	HBOS_scale	35.5%	0.21	20.49%	0.08
45	FB_minmax_scale	34.51%	0.21	17%	0.07
46	DIF_minmax_scale	33.43%	0.2	4%	0.07
47	CBLOF	32.44%	0.2	9.91%	0.07
48	QMCD	31.6%	0.2	37.9%	0.09
49	QMCD_minmax_scale	31.6%	0.2	37.9%	0.09
50	QMCD_robust_scale	31.6%	0.2	37.9%	0.09
51	QMCD_scale	31.6%	0.2	37.9%	0.09
52	COPOD	30.91%	0.2	13.36%	0.07
53	COPOD_minmax_scale	30.91%	0.2	13.36%	0.07
54	COPOD_robust_scale	30.91%	0.2	13.36%	0.07
55	COPOD_scale	30.91%	0.2	13.36%	0.07
56	HBOS_robust_scale	29.5%	0.2	16.59%	0.07
57	KDE_robust_scale	28.04%	0.2	16.37%	0.07
58	ALAD	25.72%	0.19	-12.47%	0.05
59	SUOD	24.58%	0.19	0.84%	0.06
60	LMDD_robust_scale	23.36%	0.19	37.36%	0.09
61	KPCA_minmax_scale	21.85%	0.19	1.38%	0.06
62	IForest	19.95%	0.18	1.49%	0.06
63	IForest_minmax_scale	19.95%	0.18	1.49%	0.06
64	IForest_robust_scale	19.95%	0.18	1.49%	0.06
65	IForest_scale	19.95%	0.18	1.49%	0.06

Continued on next page

Table 6 – continued from previous page

	Unnamed: 0	Recall vs. baseline (%)	Mean Recall	F1 vs. baseline (%)	Mean F1
66	MCD	19.85%	0.18	13.14%	0.07
67	MCD_robust_scale	19.53%	0.18	13.02%	0.07
68	INNE_robust_scale	19.01%	0.18	-1%	0.06
69	SOD_minmax_scale	18.75%	0.18	-51.73%	0.03
70	DIF_robust_scale	18.64%	0.18	14.03%	0.07
71	Sampling_robust_scale	17.87%	0.18	-2.87%	0.06
72	OCSVM_robust_scale	17.64%	0.18	-10.57%	0.06
73	OCSVM_minmax_scale	14.73%	0.17	4.42%	0.07
74	OCSVM_scale	13.86%	0.17	10.81%	0.07
75	KDE_scale	13.58%	0.17	10.71%	0.07
76	ROD_scale	13.02%	0.17	10.66%	0.07
77	SOD_scale	12.35%	0.17	-51.55%	0.03
78	COF_minmax_scale	12.26%	0.17	-13.96%	0.05
79	Incr. COF_minmax_scale	12.26%	0.17	-13.96%	0.05
80	MCD_minmax_scale	12.06%	0.17	10.12%	0.07
81	ROD	10.61%	0.17	9.56%	0.07
82	SOD_robust_scale	9.11%	0.17	-48.71%	0.03
83	ALAD_scale	8.01%	0.16	8.57%	0.07
84	Sampling_minmax_scale	7.43%	0.16	-24.02%	0.05
85	MCD_scale	6.47%	0.16	7.96%	0.07
86	GMM	1.19%	0.15	0.46%	0.06
87	GMM_minmax_scale	1.19%	0.15	0.46%	0.06
88	GMM_robust_scale	1.19%	0.15	0.46%	0.06
89	GMM_scale	1.19%	0.15	0.46%	0.06
90	CD	0%	0.15	0%	0.06
91	INNE	-1.81%	0.15	-19.47%	0.05
92	KDE_minmax_scale	-6.69%	0.14	-1.24%	0.06
93	KDE	-9.61%	0.14	0.95%	0.06
94	OCSVM	-11.37%	0.14	1.62%	0.06
95	Sampling	-12.29%	0.13	-11.78%	0.06
96	COPOD_normalize	-13.36%	0.13	-14.69%	0.05
97	LMDD_normalize	-15.26%	0.13	-38.28%	0.04
98	FB_normalize	-15.37%	0.13	-9.98%	0.06
99	CD_normalize	-15.74%	0.13	-43.36%	0.04
100	KPCA	-15.86%	0.13	-14.82%	0.05
101	ECOD	-15.9%	0.13	-27.86%	0.05
102	ECOD_minmax_scale	-15.9%	0.13	-27.86%	0.05
103	ECOD_robust_scale	-15.9%	0.13	-27.86%	0.05
104	ECOD_scale	-15.9%	0.13	-27.86%	0.05
105	ECOD_normalize	-17.1%	0.13	-25.6%	0.05
106	ROD_robust_scale	-17.52%	0.13	-53.58%	0.03
107	SOD_normalize	-18.68%	0.12	-1.45%	0.06
108	LODA_robust_scale	-19.6%	0.12	-5.89%	0.06
109	LOF_normalize	-19.89%	0.12	-10.38%	0.06
110	INNE_minmax_scale	-20.21%	0.12	-37.99%	0.04
111	Sampling_scale	-20.28%	0.12	-26.63%	0.05
112	ROD_minmax_scale	-21.08%	0.12	-35.1%	0.04
113	DIF_scale	-24.07%	0.12	-40.34%	0.04
114	DIF_normalize	-26.34%	0.11	34.99%	0.08
115	SOD	-27.3%	0.11	-70.41%	0.02
116	HBOS	-27.91%	0.11	-16.23%	0.05
117	CD_robust_scale	-28.23%	0.11	-60.52%	0.02
118	CD_scale	-32.42%	0.1	-63.5%	0.02
119	DIF	-32.91%	0.1	12.86%	0.07
120	LODA_scale	-33.72%	0.1	-2.38%	0.06
121	COF_normalize	-34.38%	0.1	-36.13%	0.04
122	Incr. COF_normalize	-34.38%	0.1	-36.13%	0.04
123	INNE_normalize	-37.1%	0.1	-32.95%	0.04
124	LODA	-38.02%	0.09	-0.07%	0.06
125	DeepSVDD	-41.22%	0.09	-75.84%	0.02
126	DeepSVDD_minmax_scale	-41.22%	0.09	-75.86%	0.02
127	DeepSVDD_scale	-41.22%	0.09	-75.86%	0.02
128	LUNAR_normalize	-41.43%	0.09	-38.03%	0.04
129	DeepSVDD_robust_scale	-42.29%	0.09	-76.27%	0.01
130	CD_minmax_scale	-46.78%	0.08	-79.94%	0.01
131	CBLOF_normalize	-50.23%	0.08	-36.54%	0.04
132	kNN_normalize	-52.07%	0.07	-40.5%	0.04
133	AvgKNN_normalize	-52.07%	0.07	-40.5%	0.04

Continued on next page

Table 6 – continued from previous page

	Unnamed: 0	Recall vs. baseline (%)	Mean Recall	F1 vs. baseline (%)	Mean F1
134	MedKNN_normalize	-52.07%	0.07	-40.5%	0.04
135	LODA_minmax_scale	-52.11%	0.07	-30.7%	0.04
136	PCA	-53.17%	0.07	-82.42%	0.01
137	PCA_minmax_scale	-53.17%	0.07	-82.42%	0.01
138	PCA_robust_scale	-53.17%	0.07	-82.42%	0.01
139	PCA_scale	-53.17%	0.07	-82.42%	0.01
140	QMCD_normalize	-53.95%	0.07	-5.38%	0.06
141	OCSVM_normalize	-53.99%	0.07	-29.85%	0.04
142	IForest_normalize	-55.03%	0.07	-41.68%	0.04
143	SUOD_normalize	-56.33%	0.07	-42.18%	0.04
144	Sampling_normalize	-57.52%	0.06	-57.38%	0.03
145	KDE_normalize	-57.59%	0.06	-31.26%	0.04
146	ROD_normalize	-57.59%	0.06	-31.26%	0.04
147	GMM_normalize	-57.89%	0.06	-30.14%	0.04
148	HBOS_normalize	-58.63%	0.06	-30.72%	0.04
149	PCA_normalize	-61.49%	0.06	-32.78%	0.04
150	LODA_normalize	-65.92%	0.05	-47.84%	0.03
151	ALAD_normalize	-66.96%	0.05	-44.71%	0.03
152	ALAD_minmax_scale	-68.57%	0.05	-40.12%	0.04
153	DeepSVDD_normalize	-75.05%	0.04	-80.51%	0.01
154	KPCA_normalize	-88.13%	0.02	-85.58%	0.01

Table 7: Performance of All Outlier Detection Methods by Case Study.

Case Study	**Mean Recall**	**Median Recall**	**Best Recall**	**Mean F1**	**Median F1**	**Best F1**
USA 2004 - 2004	0.63333	1.00000	1.00000	0.09889	0.15385	0.22222
HUN 2010 - 2023	0.54944	0.60000	0.90000	0.33687	0.35294	0.60000
IND 2000 - 2015	0.17390	0.06250	1.00000	0.02778	0.00995	0.17112
CRI 1949 - 2015	0.13342	0.14894	0.29787	0.09616	0.10687	0.21374
NOR 1970 - 2013	0.11904	0.02273	0.72727	0.01892	0.00361	0.11552
TZA 1992 - 2020	0.11818	0.03571	0.60714	0.01968	0.00568	0.14167
MEX 1948 - 2008	0.11577	0.05128	0.43590	0.01897	0.01222	0.07763
TUN 2011 - 2020	0.09932	0.00000	0.55556	0.10516	0.00000	0.58824
SGP 1965 - 2005	0.08401	0.05556	0.44444	0.01340	0.00883	0.07095
BWA 1990 - 2006	0.05640	0.00000	0.52941	0.00896	0.00000	0.08411
CHN 1992 - 1995	0.04028	0.00000	1.00000	0.00644	0.00000	0.16000
BRA 1995 - 2006	0.02843	0.00000	0.41667	0.00452	0.00000	0.06623
COL 2000 - 2016	0.01419	0.00000	0.29412	0.00225	0.00000	0.04673
USA 2004 - 2011	0.01324	0.00000	1.00000	0.00210	0.00000	0.15842

Table 8: Sensitivity Analysis on (+,+) Uniform Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	29.25%	–	–	–	–
1	ECOD	2.90%	2.90%	2.90%	2.90%	7.25%
2	COPOD	1.75%	1.75%	1.75%	1.75%	1.90%
3	QMCD	3.35%	3.35%	3.35%	3.35%	19%
4	KDE	4.10%	1.40%	1.75%	2.80%	1.15%
5	Sampling	2.05%	1.95%	1.20%	1.70%	1%
6	GMM	1.40%	1.40%	1.40%	1.40%	0.90%
7	PCA	2.65%	2.35%	2.35%	2.35%	2%
8	KPCA	1.20%	0.25%	0.20%	0.35%	0.25%
9	MCD	1.40%	1.40%	1.40%	1.40%	0.90%
10	CD	0.25%	0.30%	0.55%	0.55%	0.75%
11	OCSVM	4.30%	1.50%	3.90%	3.90%	1.60%
12	LMDD	1.85%	1.80%	1.80%	1.85%	2.55%
13	LOF	1.40%	1.35%	1.35%	1.35%	0.55%
14	COF	1.65%	1.65%	1.65%	1.65%	0.25%
15	Incr. COF	1.65%	1.65%	1.65%	1.65%	0.25%
16	CBLOF	2.20%	2.20%	2.20%	2.25%	1.10%
17	HBOS	3.70%	3%	3.55%	3.60%	1.10%
18	kNN	2.35%	2.35%	2.30%	2.35%	0.60%
19	AvgKNN	2.35%	2.35%	2.30%	2.35%	0.60%
20	MedKNN	2.35%	2.35%	2.30%	2.35%	0.60%
21	SOD	0.60%	0.65%	0.55%	0.65%	0.35%
22	ROD	0.35%	0.35%	0.30%	0.30%	0.85%
23	IForest	3.65%	3.65%	3.65%	3.65%	2.65%
24	INNE	9.35%	12%	12.95%	10%	5%
25	DIF	3.15%	3.10%	3.35%	3.25%	2.60%
26	FB	0.75%	0.75%	0.70%	0.75%	0.60%
27	LODA	2.65%	2.70%	2.65%	11.30%	1.25%
28	SUOD	2.85%	3.10%	3.15%	2.55%	1.90%
29	ALAD	3.95%	0.50%	1.55%	2.60%	1.40%
30	LUNAR	1.25%	1.10%	2.15%	2.20%	0.20%
31	AutoEncoder	1.10%	1.45%	1.90%	0.85%	1.20%
32	VAE	0.90%	0.90%	0.90%	0.90%	2.05%
33	DeepSVDD	0.35%	0.20%	0.15%	0.25%	1.35%

Table 9: Sensitivity Analysis on (+,-) Uniform Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	24.8%	–	–	–	–
1	ECOD	2.50%	2.50%	2.50%	2.50%	5.25%
2	COPOD	1.30%	1.30%	1.30%	1.30%	3.40%
3	QMCD	2.75%	2.75%	2.75%	2.75%	24%
4	KDE	3.60%	1.40%	1.45%	1.90%	1.15%
5	Sampling	1.75%	2.00%	1.85%	1.80%	0.45%
6	GMM	1.40%	1.40%	1.40%	1.40%	0.70%
7	PCA	2.55%	2.55%	2.55%	2.25%	1.45%
8	KPCA	0.80%	0.50%	0.10%	0.35%	0.25%
9	MCD	1.25%	1.40%	1.40%	1.40%	3%
10	CD	0.20%	0.20%	0.30%	0.30%	1.05%
11	OCSVM	3.70%	1.45%	3.15%	3.35%	1.10%
12	LMDD	16.75%	18.35%	15.85%	16.25%	27.10%
13	LOF	2.45%	2.00%	1.70%	1.50%	1.35%
14	COF	1.15%	0.50%	0.15%	0.25%	0.40%
15	Incr. COF	1.15%	0.50%	0.15%	0.25%	0.40%
16	CBLOF	2.25%	2.20%	2.30%	2.25%	0.65%
17	HBOS	2.95%	2.75%	2.45%	3.10%	2.20%
18	kNN	1.95%	1.45%	1.85%	1.60%	0.25%
19	AvgKNN	1.95%	1.45%	1.85%	1.60%	0.25%
20	MedKNN	1.95%	1.45%	1.85%	1.60%	0.25%
21	SOD	0.40%	0.35%	0.95%	0.35%	0.05%
22	ROD	0.60%	0.55%	0.65%	0.65%	1.05%
23	IForest	3%	3%	3%	3%	1.45%
24	INNE	5.65%	7.85%	7.45%	8.75%	5.35%
25	DIF	2.60%	2.40%	2.55%	2.65%	1.25%
26	FB	0.20%	0.20%	0.20%	0.20%	1.25%
27	LODA	9.10%	5.95%	7.10%	5.50%	1.15%
28	SUOD	1.55%	1.80%	2.40%	1.80%	1.35%
29	ALAD	3.50%	0.55%	0.90%	1.75%	1.20%
30	LUNAR	1.95%	0.65%	1.05%	0.50%	0.10%
31	AutoEncoder	1%	0.45%	0.30%	0.60%	1.10%
32	VAE	1.50%	1.50%	1.50%	1.50%	2%
33	DeepSVDD	0.25%	0.25%	0.15%	0.15%	1.20%

Table 10: Sensitivity Analysis on (-,-) Uniform Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	28.54%	–	–	–	–
1	ECOD	1.95%	1.95%	1.95%	1.95%	3.20%
2	COPOD	1.90%	1.90%	1.90%	1.90%	2.75%
3	QMCD	3.05%	3.05%	3.05%	3.05%	40.90%
4	KDE	3.45%	1.20%	1.15%	1.10%	45.85%
5	Sampling	2.15%	1.85%	1.35%	1.40%	12.70%
6	GMM	1.25%	1.25%	1.25%	1.25%	45.10%
7	PCA	5.80%	3.90%	3.90%	5.80%	49.90%
8	KPCA	2.60%	0.85%	0.35%	0.50%	19.75%
9	MCD	1.20%	1.20%	1.20%	0.90%	45.10%
10	CD	0.25%	0.15%	0.25%	0.25%	59.75%
11	OCSVM	3.70%	1.35%	2.70%	2.70%	45.10%
12	LMDD	19.15%	19.15%	19.15%	18.95%	48.70%
13	LOF	0.65%	0.65%	0.65%	0.65%	0.50%
14	COF	2.05%	2.00%	2.00%	2.05%	0.60%
15	Incr. COF	2.05%	2.00%	2.00%	2.05%	0.60%
16	CBLOF	1.4%	2.35%	2.1%	2.5%	13.35%
17	HBOS	2.90%	2.20%	2.80%	2.80%	37.90%
18	kNN	0.95%	0.95%	0.95%	0.95%	5.85%
19	AvgKNN	0.95%	0.95%	0.95%	0.95%	5.85%
20	MedKNN	0.95%	0.95%	0.95%	0.95%	5.85%
21	SOD	0.30%	0.40%	0.20%	0.25%	0.35%
22	ROD	0.60%	0.30%	0.55%	0.80%	53.45%
23	IForest	2.85%	2.85%	2.85%	2.85%	53.20%
24	INNE	9.95%	5.85%	8.00%	10.90%	33.75%
25	DIF	1.70%	2.05%	2.00%	2.00%	26.85%
26	FB	0.70%	0.70%	0.70%	0.70%	1.05%
27	LODA	3.05%	4.45%	3.55%	4.15%	41.40%
28	SUOD	1.15%	2.80%	1.55%	1.65%	35.10%
29	ALAD	0.35%	0.75%	2.20%	1.35%	1.55%
30	LUNAR	1.05%	0.75%	1.50%	2.00%	2.35%
31	AutoEncoder	0.10%	0.30%	1.05%	0.10%	0.85%
32	VAE	0.40%	0.40%	0.40%	0.40%	2.30%
33	DeepSVDD	0.40%	0.20%	0.20%	0.15%	1.25%

Table 11: Sensitivity Analysis on (-,+) Uniform Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	27.45%	–	–	–	–
1	ECOD	2.60%	2.60%	2.60%	2.60%	4.45%
2	COPOD	1.45%	1.45%	1.45%	1.45%	3.45%
3	QMCD	2.85%	2.85%	2.85%	2.85%	24.90%
4	KDE	3.95%	1.50%	1.40%	2.05%	1.40%
5	Sampling	1.65%	2.00%	1.85%	1.60%	1.00%
6	GMM	1.20%	1.20%	1.20%	1.20%	0.95%
7	PCA	1.70%	2.45%	2.45%	1.70%	1.60%
8	KPCA	4.00%	0.10%	0.20%	0.35%	0.10%
9	MCD	1.10%	1.05%	1.05%	1.05%	1.25%
10	CD	0.35%	0.90%	0.30%	0.25%	1.20%
11	OCSVM	4.05%	1.70%	3.25%	3.30%	1.25%
12	LMDD	31.60%	54.50%	65.35%	66.25%	47.60%
13	LOF	2.15%	1.95%	2.05%	2.05%	0.40%
14	COF	0.35%	0.90%	1.15%	1.20%	0.05%
15	Incr.COF	0.35%	0.90%	1.15%	1.20%	0.05%
16	CBLOF	2.35%	2.35%	2.00%	2.05%	0.80%
17	HBOS	3.00%	3.00%	3.00%	3.30%	1.75%
18	kNN	1.60%	1.55%	1.75%	1.70%	0.25%
19	AvgKNN	1.60%	1.55%	1.75%	1.70%	0.25%
20	MedKNN	1.60%	1.55%	1.75%	1.70%	0.25%
21	SOD	0.30%	0.40%	0.90%	0.70%	0.20%
22	ROD	1.15%	0.55%	0.65%	0.60%	1.15%
23	IForest	3.35%	3.35%	3.35%	3.35%	1.60%
24	INNE	8.95%	9.00%	9.05%	8.35%	6.60%
25	DIF	2.55%	2.55%	2.65%	2.30%	1.40%
26	FB	1.05%	1.10%	1.05%	1.00%	0.50%
27	LODA	5.00%	3.90%	7.45%	3.95%	1.60%
28	SUOD	2.70%	2.30%	2.25%	2.30%	1.40%
29	ALAD	3.50%	0.80%	1.20%	3.00%	0.80%
30	LUNAR	1.45%	0.90%	1.35%	1.75%	0.05%
31	AutoEncoder	1.40%	1.30%	0.75%	1.00%	1.80%
32	VAE	1.50%	1.45%	1.45%	1.45%	2.15%
33	DeepSVDD	0.20%	0.25%	0.20%	0.15%	0.25%

Table 12: Sensitivity Analysis on (+,+) Normal Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	20.45%	–	–	–	–
1	ECOD	4.10%	4.10%	4.10%	4.10%	6.75%
2	COPOD	2.40%	2.40%	2.40%	2.40%	2.45%
3	QMCD	2.90%	2.90%	2.90%	2.90%	2.45%
4	KDE	5.55%	0.70%	4.50%	4.15%	1.05%
5	Sampling	1.00%	1.45%	1.55%	1.20%	1.30%
6	GMM	0.75%	0.75%	0.75%	0.75%	1.00%
7	PCA	1.85%	1.85%	1.95%	1.85%	1.60%
8	KPCA	1.40%	0.35%	1.05%	0.30%	0.10%
9	MCD	0.70%	0.70%	0.65%	0.65%	0.80%
10	CD	0.30%	0.25%	0.15%	0.15%	0.30%
11	OCSVM	8.95%	0.70%	7.65%	5.40%	1.50%
12	LMDD	1.15%	1.35%	1.25%	1.35%	0.90%
13	LOF	0.40%	1.00%	0.65%	0.90%	0.45%
14	COF	0.20%	1.35%	1.05%	1.40%	0.35%
15	Incr. COF	0.20%	1.35%	1.05%	1.40%	0.35%
16	CBLOF	1.40%	2.30%	2.15%	2.35%	1.35%
17	HBOS	12.20%	3.95%	4.85%	4.35%	10.30%
18	kNN	1.45%	0.90%	0.85%	0.90%	0.50%
19	AvgKNN	1.45%	0.90%	0.85%	0.90%	0.50%
20	MedKNN	1.45%	0.90%	0.85%	0.90%	0.50%
21	SOD	0.45%	0.25%	0.25%	0.30%	0.25%
22	ROD	0.05%	0.30%	0.30%	0.15%	0.80%
23	IForest	4.55%	4.55%	4.55%	4.55%	5.00%
24	INNE	35.05%	29.95%	30.85%	32.75%	17.95%
25	DIF	4.45%	4.50%	4.45%	4.50%	3.35%
26	FB	0.65%	0.90%	0.75%	0.85%	0.65%
27	LODA	2.15%	1.75%	2.00%	1.95%	1.65%
28	SUOD	4.80%	4.05%	4.40%	4.30%	4.20%
29	ALAD	5.30%	2.05%	0.70%	0.05%	4.40%
30	LUNAR	0.20%	0.05%	2.15%	0.15%	0.15%
31	AutoEncoder	1.80%	2.55%	2.40%	0.95%	0.85%
32	VAE	3.20%	3.20%	3.20%	3.20%	1.55%
33	DeepSVDD	0.50%	0.25%	0.30%	0.85%	0.60%

Table 13: Sensitivity Analysis on (+,-) Normal Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	16.7%	–	–	–	–
1	ECOD	4.30%	4.30%	4.30%	4.30%	6.85%
2	COPOD	3.25%	3.25%	3.25%	3.25%	2.45%
3	QMCD	2.75%	2.75%	2.75%	2.75%	2.45%
4	KDE	5.75%	0.80%	4.75%	4.10%	1.05%
5	Sampling	0.85%	1.70%	1.10%	1.25%	1.55%
6	GMM	0.65%	0.65%	0.65%	0.65%	0.75%
7	PCA	1.40%	1.40%	1.65%	1.40%	1.75%
8	KPCA	1.00%	0.50%	0.45%	0.20%	0.20%
9	MCD	0.50%	0.50%	0.50%	0.50%	0.65%
10	CD	0.20%	0.70%	0.20%	0.20%	0.20%
11	OCSVM	10.05%	0.85%	8.25%	5.15%	1.60%
12	LMDD	0.80%	0.10%	0.15%	0.15%	0.10%
13	LOF	0.30%	0.80%	1.00%	1.05%	0.50%
14	COF	0.70%	0.80%	0.90%	0.90%	0.80%
15	Incr. COF	0.70%	0.80%	0.90%	0.90%	0.80%
16	CBLOF	1.55%	1.85%	1.65%	1.60%	1.60%
17	HBOS	6.35%	3.30%	6.35%	6.35%	10.35%
18	kNN	0.55%	0.40%	0.35%	0.40%	0.30%
19	AvgKNN	0.55%	0.40%	0.35%	0.40%	0.30%
20	MedKNN	0.55%	0.40%	0.35%	0.40%	0.30%
21	SOD	0.10%	0.30%	0.40%	0.70%	0.20%
22	ROD	0.20%	0.20%	0.25%	0.55%	0.65%
23	IForest	4.70%	4.70%	4.70%	4.70%	4.90%
24	INNE	30.20%	29.30%	26.65%	28.30%	19.70%
25	DIF	4.50%	4.55%	4.70%	3.90%	3.25%
26	FB	0.45%	1.10%	1.30%	1.20%	0.45%
27	LODA	1.75%	1.50%	2.05%	1.75%	0.80%
28	SUOD	4.05%	3.80%	4.35%	4.55%	3.85%
29	ALAD	10.45%	2.20%	2.25%	0.80%	1.05%
30	LUNAR	0.25%	2.45%	0.35%	0.05%	0.10%
31	AutoEncoder	1.75%	1.70%	2.90%	1.95%	17.60%
32	VAE	2.65%	2.65%	2.65%	2.65%	1.60%
33	DeepSVDD	0.85%	0.40%	0.25%	0.95%	0.75%

Table 14: Sensitivity Analysis on (-,-) Normal Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	18.09%	–	–	–	–
1	ECOD	3.60%	3.60%	3.60%	3.60%	7.45%
2	COPOD	2.60%	2.60%	2.60%	2.60%	2.50%
3	QMCD	2.55%	2.55%	2.55%	2.55%	2.70%
4	KDE	5.60%	0.60%	4.65%	4.15%	1.45%
5	Sampling	1.55%	1.50%	1.20%	1.15%	1.50%
6	GMM	0.85%	0.85%	0.85%	0.85%	1.60%
7	PCA	1.50%	1.50%	1.65%	1.50%	1.95%
8	KPCA	2.20%	0.25%	0.65%	0.25%	0.55%
9	MCD	0.45%	0.40%	0.40%	0.45%	0.95%
10	CD	0.10%	0.15%	0.25%	0.25%	1.10%
11	OCSVM	9.70%	0.65%	8.50%	5.20%	1.85%
12	LMDD	0.85%	1.60%	0.15%	1.75%	2.20%
13	LOF	1.05%	1.20%	1.05%	1.15%	0.70%
14	COF	0.95%	0.20%	0.40%	0.20%	0.65%
15	Incr. COF	0.95%	0.20%	0.40%	0.20%	0.65%
16	CBLOF	1.70%	1.55%	1.50%	1.50%	1.65%
17	HBOS	7.60%	3.35%	7.45%	6.90%	10.45%
18	kNN	0.90%	0.70%	0.85%	0.70%	0.70%
19	AvgKNN	0.90%	0.70%	0.85%	0.70%	0.70%
20	MedKNN	0.90%	0.70%	0.85%	0.70%	0.70%
21	SOD	0.30%	0.30%	0.40%	0.35%	0.25%
22	ROD	0.20%	0.10%	0.15%	0.10%	0.95%
23	IForest	4.35%	4.35%	4.35%	4.35%	4.35%
24	INNE	31.40%	32.30%	35.45%	36.90%	15.10%
25	DIF	4.35%	4.40%	4.45%	4.30%	3.70%
26	FB	1.25%	1.30%	1.00%	1.20%	0.60%
27	LODA	1.90%	1.75%	2.00%	1.95%	2.35%
28	SUOD	4.05%	3.70%	3.95%	4.15%	2.80%
29	ALAD	4.55%	2.30%	0.70%	0.20%	1.50%
30	LUNAR	0.25%	0.95%	2.30%	0.05%	0.15%
31	AutoEncoder	2.75%	2.25%	4.30%	1.50%	42.70%
32	VAE	3.60%	3.60%	3.55%	3.60%	61.85%
33	DeepSVDD	0.15%	0.85%	0.60%	0.65%	33.45%

Table 15: Sensitivity Analysis on (-,+) Normal Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	19.2%	–	–	–	–
1	ECOD	3.60%	3.60%	3.60%	3.60%	7.15%
2	COPOD	3.00%	3.00%	3.00%	3.00%	2.40%
3	QMCD	1.80%	1.80%	1.80%	1.80%	2.75%
4	KDE	5.60%	0.35%	4.25%	3.50%	1.65%
5	Sampling	0.20%	0.90%	0.90%	0.70%	1.90%
6	GMM	0.35%	0.35%	0.35%	0.35%	1.35%
7	PCA	1.05%	1.05%	1.05%	1.20%	2.05%
8	KPCA	0.70%	0.05%	1.75%	0.35%	0.15%
9	MCD	0.35%	0.35%	0.35%	0.35%	1.20%
10	CD	0.10%	0.25%	0.15%	0.15%	0.80%
11	OCSVM	10.10%	0.35%	7.80%	5.20%	1.90%
12	LMDD	0.50%	0.60%	0.35%	0.35%	0.45%
13	LOF	1.40%	0.85%	0.85%	0.85%	0.40%
14	COF	1.70%	1.65%	1.80%	1.80%	0.35%
15	Incr. COF	1.70%	1.65%	1.80%	1.80%	0.35%
16	CBLOF	0.40%	1.00%	1.05%	1.00%	1.85%
17	HBOS	5.90%	2.95%	5.30%	5.45%	9.45%
18	kNN	0.20%	0.40%	0.35%	0.35%	0.25%
19	AvgKNN	0.20%	0.40%	0.35%	0.35%	0.25%
20	MedKNN	0.20%	0.40%	0.35%	0.35%	0.25%
21	SOD	0.05%	0.20%	0.20%	0.20%	0.15%
22	ROD	0.05%	0.15%	0.15%	0.20%	1.20%
23	IForest	4.10%	4.10%	4.10%	4.10%	4.55%
24	INNE	31.90%	29.40%	30.40%	29.20%	18.00%
25	DIF	4.05%	4.00%	4.15%	4.00%	3.45%
26	FB	1.60%	1.60%	1.70%	1.60%	0.55%
27	LODA	1.25%	0.90%	0.65%	1.40%	2.60%
28	SUOD	2.85%	2.70%	1.80%	2.25%	5.60%
29	ALAD	1.70%	2.15%	0.70%	0.85%	3.95%
30	LUNAR	3.25%	0.20%	0.05%	1.70%	0.15%
31	AutoEncoder	2.15%	2.00%	2.70%	1.15%	9.60%
32	VAE	1.90%	1.90%	1.90%	1.90%	1.20%
33	DeepSVDD	0.35%	0.60%	0.30%	0.60%	18.35%

Table 16: Sensitivity Analysis on (+,+) Skewed Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	17.4%	–	–	–	–
1	ECOD	4.45%	4.45%	4.45%	4.45%	5.05%
2	COPOD	2.60%	2.60%	2.60%	2.60%	2.40%
3	QMCD	2.90%	2.90%	2.90%	2.90%	1.90%
4	KDE	5.05%	0.20%	4.45%	3.95%	0.65%
5	Sampling	1.15%	1.15%	0.90%	0.60%	0.50%
6	GMM	0.20%	0.20%	0.20%	0.20%	0.75%
7	PCA	0.95%	0.95%	0.95%	1.30%	1.00%
8	KPCA	1.05%	0.30%	0.30%	0.40%	0.05%
9	MCD	0.15%	0.15%	0.15%	0.15%	0.55%
10	CD	0.50%	0.65%	0.45%	0.45%	0.40%
11	OCSVM	9.00%	0.30%	8.80%	6.90%	0.90%
12	LMDD	1.25%	0.55%	3.35%	3.15%	1.40%
13	LOF	0.65%	0.85%	0.65%	0.90%	1.40%
14	COF	1.30%	0.60%	0.65%	0.45%	1.00%
15	Incr. COF	1.30%	0.60%	0.65%	0.45%	1.00%
16	CBLOF	1.20%	1.45%	1.10%	1.30%	0.90%
17	HBOS	4.15%	3.15%	3.60%	3.60%	1.70%
18	kNN	0.20%	0.15%	0.15%	0.15%	0.30%
19	AvgKNN	0.20%	0.15%	0.15%	0.15%	0.30%
20	MedKNN	0.20%	0.15%	0.15%	0.15%	0.30%
21	SOD	0.05%	0.10%	0.10%	0.10%	0.20%
22	ROD	0.10%	0.05%	0.05%	0.40%	0.55%
23	IForest	4.65%	4.65%	4.65%	4.65%	4.25%
24	INNE	31.85%	32.00%	30.00%	30.65%	14.70%
25	DIF	3.95%	4.05%	4.05%	4.05%	3.70%
26	FB	0.60%	0.45%	0.60%	0.45%	1.65%
27	LODA	0.30%	0.60%	0.70%	0.65%	1.70%
28	SUOD	0.90%	3.80%	3.00%	3.00%	3.85%
29	ALAD	3.45%	2.40%	0.70%	0.90%	2.45%
30	LUNAR	0.35%	0.35%	1.75%	1.45%	0.25%
31	AutoEncoder	1.30%	1.40%	0.30%	2.00%	0.65%
32	VAE	1.15%	1.15%	1.15%	1.15%	0.95%
33	DeepSVD	0.30%	0.10%	0.20%	0.10%	0.55%

Table 17: Sensitivity Analysis on (+,-) Skewed Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	18.05%	–	–	–	–
1	ECOD	3.30%	3.30%	3.30%	3.30%	4.20%
2	COPOD	1.70%	1.70%	1.70%	1.70%	1.85%
3	QMCD	2.20%	2.20%	2.20%	2.20%	1.40%
4	KDE	4.45%	0.45%	4.15%	3.55%	0.40%
5	Sampling	1.05%	1.30%	0.90%	0.90%	0.70%
6	GMM	0.70%	0.70%	0.70%	0.70%	0.35%
7	PCA	1.30%	1.30%	1.10%	1.10%	0.60%
8	KPCA	0.80%	0.15%	0.80%	0.20%	0.15%
9	MCD	0.45%	0.45%	0.45%	0.45%	0.20%
10	CD	0.05%	0.20%	0.05%	0.05%	0.70%
11	OCSVM	9.20%	0.55%	8.90%	7.20%	0.75%
12	LMDD	2.10%	2.25%	1.00%	1.25%	1.70%
13	LOF	1.15%	0.85%	1.00%	0.75%	0.75%
14	COF	0.80%	1.40%	1.30%	1.40%	0.75%
15	Incr. COF	0.80%	1.40%	1.30%	1.40%	0.75%
16	CBLOF	1.10%	1.25%	1.20%	1.20%	0.65%
17	HBOS	6.90%	3.45%	6.15%	5.60%	1.15%
18	kNN	1.05%	0.85%	0.85%	0.80%	0.30%
19	AvgKNN	1.05%	0.85%	0.85%	0.80%	0.30%
20	MedKNN	1.05%	0.85%	0.85%	0.80%	0.30%
21	SOD	0.30%	0.15%	0.20%	0.25%	0.05%
22	ROD	0.10%	0.30%	0.10%	0.30%	0.20%
23	IForest	4.25%	4.25%	4.25%	4.25%	4.30%
24	INNE	24.25%	25.55%	23.25%	26.35%	15.00%
25	DIF	4.05%	3.95%	4.10%	3.80%	2.65%
26	FB	1.65%	1.20%	1.15%	1.15%	1.10%
27	LODA	1.75%	1.90%	1.90%	1.95%	0.20%
28	SUOD	3.40%	3.60%	3.15%	3.40%	2.20%
29	ALAD	11.35%	0.60%	1.90%	2.65%	1.30%
30	LUNAR	0.05%	0.05%	0.10%	0.20%	0.05%
31	AutoEncoder	0.65%	0.40%	0.65%	0.55%	0.35%
32	VAE	1.10%	1.15%	1.10%	1.15%	0.60%
33	DeepSVD	0.25%	0.30%	0.35%	0.40%	0.30%

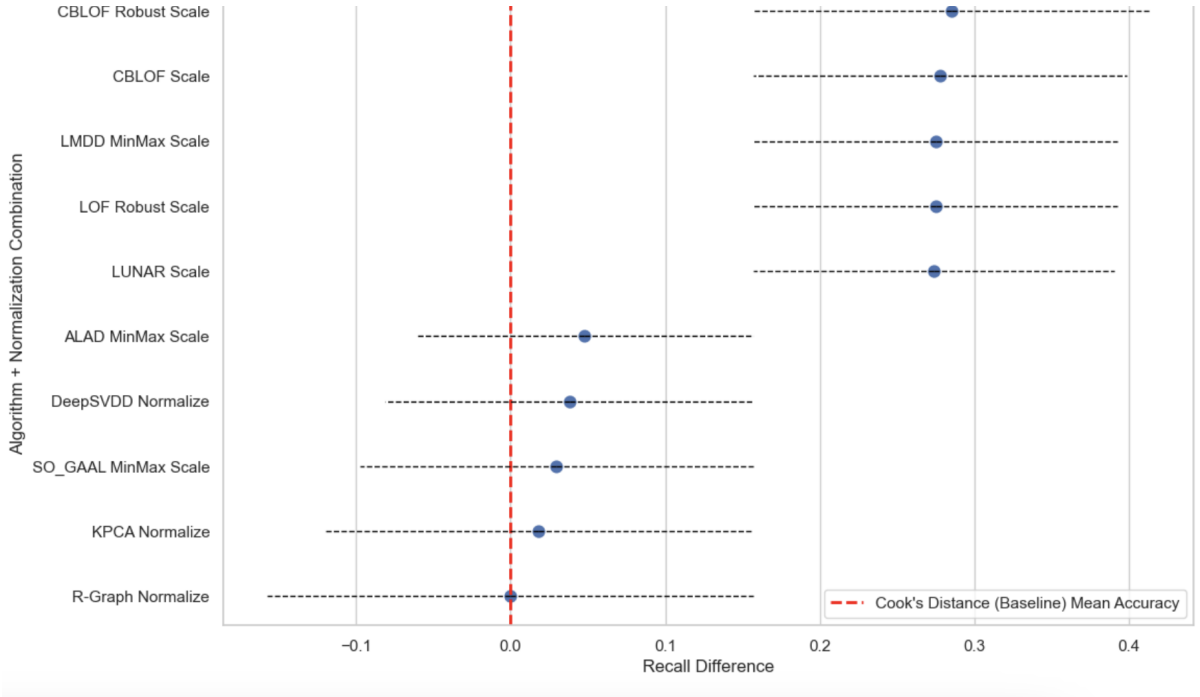
Table 18: Sensitivity Analysis on (-,-) Skewed Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	23.75%	–	–	–	–
1	ECOD	3.80%	3.80%	3.80%	3.80%	5.70%
2	COPOD	2.90%	2.90%	2.90%	2.90%	2.65%
3	QMCD	2.35%	2.35%	2.35%	2.35%	1.85%
4	KDE	4.65%	0.20%	4.30%	3.85%	1.15%
5	Sampling	0.60%	0.70%	1.00%	0.50%	1.15%
6	GMM	0.35%	0.35%	0.35%	0.35%	1.55%
7	PCA	0.50%	0.50%	0.50%	0.50%	1.85%
8	KPCA	0.65%	0.50%	0.55%	0.75%	0.25%
9	MCD	0.15%	0.15%	0.15%	0.15%	1.00%
10	CD	0.10%	0.35%	0.05%	0.05%	0.70%
11	OCSVM	8.70%	0.35%	7.95%	6.20%	1.55%
12	LMDD	0.80%	0.45%	0.50%	0.50%	2.20%
13	LOF	0.40%	0.40%	0.45%	0.45%	1.00%
14	COF	0.55%	0.55%	0.50%	0.50%	1.30%
15	Incr. COF	0.55%	0.55%	0.50%	0.50%	1.30%
16	CBLOF	0.75%	0.75%	0.70%	0.60%	1.55%
17	HBOS	7.90%	3.85%	4.45%	4.30%	2.05%
18	kNN	0.35%	0.30%	0.25%	0.25%	0.55%
19	AvgKNN	0.35%	0.30%	0.25%	0.25%	0.55%
20	MedKNN	0.35%	0.30%	0.25%	0.25%	0.55%
21	SOD	0.30%	0.10%	0.15%	0.15%	0.25%
22	ROD	0.20%	0.15%	0.10%	0.10%	1.00%
23	IForest	4.40%	4.40%	4.40%	4.40%	4.05%
24	INNE	32.35%	30.60%	25.95%	29.45%	20.10%
25	DIF	4.10%	4.20%	4.10%	4.10%	3.60%
26	FB	0.40%	0.60%	0.70%	0.60%	0.95%
27	LODA	0.10%	0.10%	0.10%	0.10%	2.30%
28	SUOD	3.55%	3.10%	0.95%	3.70%	4.35%
29	ALAD	9.20%	1.35%	0.55%	2.10%	1.75%
30	LUNAR	0.05%	0.95%	0.05%	3.00%	0.15%
31	AutoEncoder	0.55%	1.55%	1.40%	0.95%	0.70%
32	VAE	1.45%	1.45%	1.45%	1.45%	1.60%
33	DeepSVDD	0.30%	0.30%	0.05%	0.25%	0.90%

Table 19: Sensitivity Analysis on (-,+) Skewed Distributions

	Methods	Original (%)	MinMax Scale (%)	Robust Scale (%)	Scale (%)	Normalize (%)
0	ACROSS	19.35%	–	–	–	–
1	ECOD	4.05%	4.05%	4.05%	4.05%	5.40%
2	COPOD	2.45%	2.45%	2.45%	2.45%	2.20%
3	QMCD	2.55%	2.55%	2.55%	2.55%	1.75%
4	KDE	4.55%	0.55%	4.15%	3.70%	0.85%
5	Sampling	0.85%	1.60%	1.30%	1.50%	1.00%
6	GMM	0.60%	0.60%	0.60%	0.60%	1.15%
7	PCA	1.50%	1.50%	1.50%	1.50%	1.55%
8	KPCA	0.40%	0.30%	0.75%	0.35%	0.35%
9	MCD	0.60%	0.60%	0.65%	0.60%	0.75%
10	CD	0.05%	0.10%	0.15%	0.15%	0.20%
11	OCSVM	8.80%	0.65%	8.50%	6.10%	1.35%
12	LMDD	3.45%	0.65%	0.65%	0.80%	2.40%
13	LOF	0.35%	0.50%	0.30%	0.40%	0.70%
14	COF	0.50%	0.40%	0.30%	0.40%	1.25%
15	Incr. COF	0.50%	0.40%	0.30%	0.40%	1.25%
16	CBLOF	1.05%	1.65%	1.70%	1.65%	1.30%
17	HBOS	3.60%	3.45%	3.65%	3.65%	2.00%
18	kNN	0.75%	0.95%	1.00%	0.95%	0.20%
19	AvgKNN	0.75%	0.95%	1.00%	0.95%	0.20%
20	MedKNN	0.75%	0.95%	1.00%	0.95%	0.20%
21	SOD	0.15%	0.30%	0.25%	0.20%	0.10%
22	ROD	0.20%	0.25%	0.15%	0.25%	0.75%
23	IForest	4.30%	4.30%	4.30%	4.30%	4.05%
24	INNE	31.15%	27.30%	29.95%	27.10%	14.40%
25	DIF	4.00%	4.05%	4.15%	3.95%	3.35%
26	FB	0.50%	0.45%	0.35%	0.45%	1.20%
27	LODA	1.50%	2.00%	1.90%	2.15%	0.70%
28	SUOD	1.50%	4.20%	4.05%	4.20%	3.45%
29	ALAD	11.45%	0.55%	0.45%	2.60%	1.05%
30	LUNAR	3.85%	3.85%	0.90%	0.45%	0.10%
31	AutoEncoder	0.70%	0.85%	0.85%	0.80%	1.05%
32	VAE	0.50%	0.50%	0.50%	0.50%	1.75%
33	DeepSVD	0.15%	0.30%	0.40%	0.30%	0.90%

Figure 5: Recall Variation for Top 5 and Bottom 5 Methods by Cook's Distance



Listing 1: ACROSS Script

```python
import pandas as pd
from sklearn.preprocessing import robust_scale, scale, minmax_scale

def across_od(data):  # data is a numpy array
    """
    Perform outlier detection using multiple models and normalization techniques.

    Args:
        data (numpy.ndarray): Input data array.

    Returns:
        dict: A dictionary containing decisions based on maximum outlier scores.
    """
    scores = []
    model_list = [
        CBLOF(alpha=0.75, beta=3.0),
        CBLOF(alpha=0.75, beta=3.0),
        LMDD(),
        LOF(),
        LUNAR()
    ]  # Replace with actual model instances

    normalizations_list = [
        robust_scale,
        scale,
        minmax_scale,
        robust_scale,
        scale
    ]
```

```
30
31      for i, model in enumerate(model_list):
32          # Normalize data using the corresponding normalization method
33          X = normalizations_list[i](data)
34          # Fit and predict with the model
35          pred = model.fit_predict(X)
36          # Append predictions to the list
37          scores.append(pred)
38
39      # Convert list of scores into a DataFrame (each column corresponds
            to a model)
40      scores_df = pd.DataFrame(scores).T
41      # Find the maximum score across all models for each data point
42      max_scores = scores_df.max(axis=1)
43
44      return {
45          'decisions': max_scores
46      }
```