Anomaly Analysis in Tax-Exempt Non-Profit Organization Filings:

Applying Unsupervised Learning Method Ensembles on 2013-2021 IRS Form 990s to Perform Outlier Detection as a Measure of Anomalous Activity

Master of Quantitative Economics at University of California, Los Angeles

Tanner M. Woods

Dr. Randall R. Rojas

2 June 2023

Contents

1. Abstract 3

2. Introduction 4

3. Outlier Detection in High-Dimensional Data Spaces 5

3.1 Overviewing the Challenges of Outlier Detection in High-Dimensional Data Spaces 5

3.2 Restructuring the Challenges Presented by Zimek et al. 5

4. Core Challenge #1: Homogenization of the Full-Dimensional Data Space 6

4.1 Shrinking Contrast Between Inliers and Outliers 6

4.2 Bias of L_p-Derived Scores in Higher Dimensions 7

4.3 Enforced Sparsity of the Data Space, or the Equivalence of Hubs and Anti-Hubs 8

4.4 Recursive Nature of Mapping Connected Neighborhoods and Subspaces 9

5. Core Challenge #2: Subspace Selection and Combinatorial Explosion 10

5.1 The Combinatorial Explosion of Total Searchable Subspaces and Absolute Knowledge Bias 10

5.2 Incomparability of Outlier Scores Between Subspaces of Different Dimensionalities 11

6. Confronting the Components of Each Core Issue 11

6.1 Defining Outlierness in the Data Space: Axis-Parallel or Arbitrary? 11

6.2 Possible Pre-Processing for Selection of High-Contrast Subspaces 12

6.3 Proposals for Outlier Detection in High-Dimensional Space 12

7. Defining the Data Space's Path for Determination of Outlierness 13

8. Analyzing the Decision Labels of the Outlier Detection Ensemble 14

8.1 Low-Dimensional Decisions 14

8.2 High-Dimensional Decisions 14

9. Identifying Organizations Exhibiting Anomalous Activity 19

9.1 Non-Composite Identification 19

9.2 Composite Identification 21

9.2.1 Constructing the Composite Outlierness Ranking System 21

9.2.2 Overall and Annualized Composite Ranking 23

10. Future Work 24

11. Conclusion 24

12. References 25

1. Abstract

High-dimensional outlier detection is a dynamic and rapidly advancing field. Drawing from both classic and modern methods, this paper aims to construct an ensemble approach to the identification of outlying high-dimensional data. Following an overview of the challenges inherent within the field, a short summary of the approach to avoiding these problems will be provided. In brief, this paper leans on two core concepts: first, the possibility for characterization of a higher-dimensional space via feature bagging; second, sidestepping interdimensional comparisons via dimensionless characteristics (e.g., Frequency). After defining the structure of the data pipeline, the ensemble will be constructed with components defining outlierness in both the full-dimensional space (iForest, ECOD, COPOD) and within reduced-dimensionality subsets (LOF, CBLOF, and HBOS). The reduced-dimensionality outlierness results will then be implemented as a boosting mechanism to stratify the global outlier rankings, thereby constructing a locality-sensitive measure of global outlierness.

2. Introduction

In response to the economic downturn induced by the COVID-19 pandemic, the federal government implemented the Coronavirus Aid, Relief, and Economic Security (CARES) Act in March 2020. One of most prominent elements of the CARES Act was the Paycheck Protection Program (PPP), which appears to be a major source of fraudulent activities. By 2021, reports indicated that up to $76 billion may have been allocated to applicants under false pretenses and nearly 500 individuals arrested (Stacy, 2021). Stories surrounding such abuse of funds intended for the commonwealth and wellbeing of the body politic runs wild in the popular consciousness. However, rather than positing yet another approach to answer how, why, or to what degree this fraud took place this paper aims to establish a mechanism for unsupervised detection of such activities.

Unfortunately, the individuals and organizations caught committing such activities are understandably reluctant to part with their financial records. Instead, the author proposes the application of an unsupervised outlier detection ensemble to the closest equivalent: the 2013-21 publicly available Form 990s archived by the Internal Revenue Service; form layouts before 2013 were too different to be adaptable. These forms track the reported expenditures, revenue, donations, and even lobbying activities of any organization operating as a tax-exempt non-profit. Using this information as a quantitative foundation, this paper aims to attempt the fulfillment of three goals. First, the identification of outlying observations in the high-dimensional space to measure global outlierness. Second, the identification of outlying observations in feature subsets of the high-dimensional space to measure local outlierness. Third, to provide a method to boost the global outlierness of a filing by incorporating locality-sensitive components of outlierness. Due to the broad scope of the field of outlier detection in high-dimensional spaces, a streamlined overview of the subject will also be provided.

3. Outlier Detection in High-Dimensional Data Spaces

3.1 Overviewing the Challenges of Outlier Detection in High-Dimensional Data Spaces

Despite the breadth of the problem at hand, the survey authored by Zimek et al. on contemporary research efforts and tools towards outlier detection in high-dimensional space provides a wonderfully succinct overview of the core dilemmas:

1. Central limit theorem and its effect on the concentration of scores.
2. Relevant neighborhoods can be disrupted by irrelevant attributes.
3. Recursive nature of needing neighborhoods to choose a subspace and needing to choose a subspace to know the neighborhoods due to breakdown in Euclidean distances.
4. Bias of Lebesgue space (L_p) scores towards higher-dimensionalities due to breakdown in Euclidean distances.
5. Partial and total information loss for comparisons of outlier rankings and outlier scores between subspaces of different dimensionalities, respectively.
6. Combinatorial explosion of total searchable subspaces for increasing dimensionality.
7. Implementing traditional searching methods on combinatorial spaces is both computationally infeasible and induces a pre-cognizance bias due to total knowledge of the data space.
8. The homogeneity of the data space in increasing dimensionalities produces a statistical phenomenon of equal probability for the appearance of "hubs" and "anti-hubs", which can disrupt outlier labeling based on anti-hub detection.

3.2 Restructuring the Challenges Presented by Zimek et al.

In this survey, the authors appear to define the problems of high-dimensional outlier detection as eight distinct issues. However, this paper posits that these eight core issues can be redefined as two core issues with related sub-issues. With some adjustments for succinctness, these two core issues and their components are as follows:

1. Homogenization of the full-dimensional data space.
    a. Shrinking contrast between inliers and outliers.
    b. Bias of L_p-derived scores towards higher dimensionalities.
    c. Enforced sparsity of the data space, or the equivalence of hubs and anti-hubs.

    d. Recursive nature of mapping connected neighborhoods and subspaces.
2. Subspace selection and combinatorial explosion.
    a. The combinatorial explosion of total searchable subspaces.
    b. Incomparability of outlier scores between subspaces of different dimensionalities.

For the justification behind this restructuring, it appears that two of the given issues themselves cannot be solved head-on. No matter the mathematical loopholes, enormous quantities of data in high-dimensional space will induce CLT and selecting many distinct combinations of items from a high-dimensional group of items will induce combinatorial explosion. Instead, the components constituting these two constants must be the focus of our attention — that is, we work towards solving the components to alleviate the unsolvable core constants.

4. Core Challenge #1: Homogenization of the Full-Dimensional Data Space

4.1 Shrinking Contrast Between Inliers and Outliers

In one-dimensional data analysis, one can trivially determine the inlier/outlier status of a data point by its distance from the mean relative to the standard deviation of the distribution; see also Tukey's IQR fences and the modified Thompson-Tau test. Two-dimensional and higher analysis permits the use of more sophisticated contrasting techniques: methods like k-nearest neighbors (k-NN) or local outlier factor (LOF). However, the performance of distance-based methods like k-NN begins to degrade as the dimensionality of a data space increases past approximately ten dimensions. This is a consequence of the increasing difficulty of differentiating between inliers and outliers by means of absolute distance as the dimensionality of a space increases, or in more technical terms[1]:

[1] This is an oversimplification of a complex mathematical reality and should not be taken as a thorough explanation of the characteristics of hypershapes in ever-increasing dimensionalities.

(Simple) Theorem 1: Limits of hyperspheric volume and absolute distance for increasing dimensionality

For a data space of an arbitrary dimensionality, d, whose distance is defined in Euclidean terms, the volume of the accompanying hypersphere is such that…

$$\lim_{d\to\infty} Volume(H.sphere) = 0$$

And therefore, it follows that…

$$\lim_{d\to\infty} \left| Position(x_i) - Position(x_j) \right| = 0$$
$$\forall\, i \neq j$$

Compounding the CLT-induced homogenization, his is further complicated if the correlation between the attributes within a data space is high, thereby lowering the intrinsic dimensionality — that is, the number of dimensions that are informational — of the dataset. As non-informational dimensions increase in proportion to the total space, the latent correlation (i.e., Noise) within the data space likewise increases and produces a further confounding effect upon distance-based decisions.

4.2 Bias of L_p-Derived Scores in Higher Dimensions

As the dimensionality of a space increases, scores based on L_p measures begin to converge towards inseparability. Drawing back on Theorem 1, the convergence upon zero of two given points under infinite dimensionality induces a similar property upon the relationship between Euclidean distances[2]:

(Simple) Theorem 2: Limits of the variance of the mean magnitude-scaled distance distribution

For a data space of an arbitrary dimensionality, d, whose distance is defined in Euclidean terms, then the relationship between the distance distribution and mean magnitude of distance is such that…

$$\lim_{d\to\infty} Var\left(\frac{Distance\ distribution}{E[Magnitude\ of\ distance]}\right) = \frac{Var(Distance\ distribution)}{(E[Magnitude\ of\ distance])^2} = 0$$

[2] In the same spirit as Theorem 1, this should not be taken as a rigorous proof.

4.3 Enforced Sparsity of the Data Space, or the Equivalence of Hubs and Anti-Hubs

Intuitively, the concept of observations acting as hubs for a cluster of similar neighbors is rather straightforward. We will define the hubness score, $N_k(x_i)$, of an observation as follows:

(Simple) Theorem 3: Hubness score of an observation.

Let the hubness score of an observation, $N_k(x_i)$, be…

$$N_k(x_i) = \sum_{i=1}^{n} \left(x_i \in kNN\{x_j\}\right) \; \forall \; i \neq j$$

Then, an observation is designated as a hub, antihub, or mundane point under the following…

$$x_i \coloneqq \begin{cases} Hub \; ; \; N_k(x_i) > \; 2\sigma * \overline{N_k}(X) \\ Antihub \; ; \; N_k(x_i) \; < \overline{N_k}(X) / \; 2\sigma \\ "Mundane" \; ; \; elsewise \end{cases}$$

However, the earlier discussion of decreasing contrast between points for increasing dimensionality in Section 4.1 rears its head to disrupt the discussion of hubs. If hubs are points meeting the criteria in the definition of hubness score above, then this implies that the frequency of occurrence for hubs should increase with dimensionality. Indeed, a Spearman correlation test between dimensionality and the skewness of $Pois(\lambda)$-distributed k-occurrences using 50 empirical data sets produced a statistically significant relationship between the two (Radovanović et al., 2010). Albeit, not as strong as shown by the authors in prior synthetic data, but this is justified as consequence of real-world data possessing a tendency towards lower intrinsic dimensionality.

Having defined the phenomenon and its positive relationship with dimensionality, the question of its relevance to outlier detection stands open. As can be seen in the above definition, an antihub designates an observation of distinctly low hubness — that is, whose position in the data space is very distant from other neighbors. Therefore, antihubs can be considered as Euclidean-based distance outliers. However, we already understand that there must be something askew here given previous discussion of the breakdown of Euclidean-based distances. Indeed, turning back to Theorem 2 provides some insight: the variance of the scaled distance distribution *converges upon* zero, yet it will *never equal* zero. Therefore, due to the homogeneity of the data space there is equal probability of producing either a hub or an anti-hub (Radovanović et al., 2010), which renders anti-hub occurrence

as a measure of outlierness detection futile. This further complicates the issue of determining most representative subspaces in the next section.

4.4 Recursive Nature of Mapping Connected Neighborhoods and Subspaces

While this component bleeds over into the second core challenge, it is also a product of the statistical equivalence of hubs and anti-hubs. As previously stated, any given point has equal probability of being labeled a hub or anti-hub. If the latter are deemed outliers, then it follows that any given point has an equal probability of exhibiting outlierness; this is visualized in Figure 1. In the accompanying Figure 2, one can similarly understand the importance of identifying the most representative subspace in neighborhood-based characterizations of outlierness and relation.

If every point can exhibit outlierness, then any process of determining global and/or local outlierness must simultaneously determine two items of interest for every point: first, the neighborhood of said point; second, the most representative subspace of said neighborhood. Of course, this is quite an impossible task given the two are intrinsically linked. This may be alleviated through a total and thorough exploration of all possible subspaces, but the issues present within this approach will be discussed in the upcoming Section 5.1.

Figure 1: Effect of increasing dimensionality on distribution probabilities under an $\widetilde{iid}\ N(\mu, \sigma^2)$.

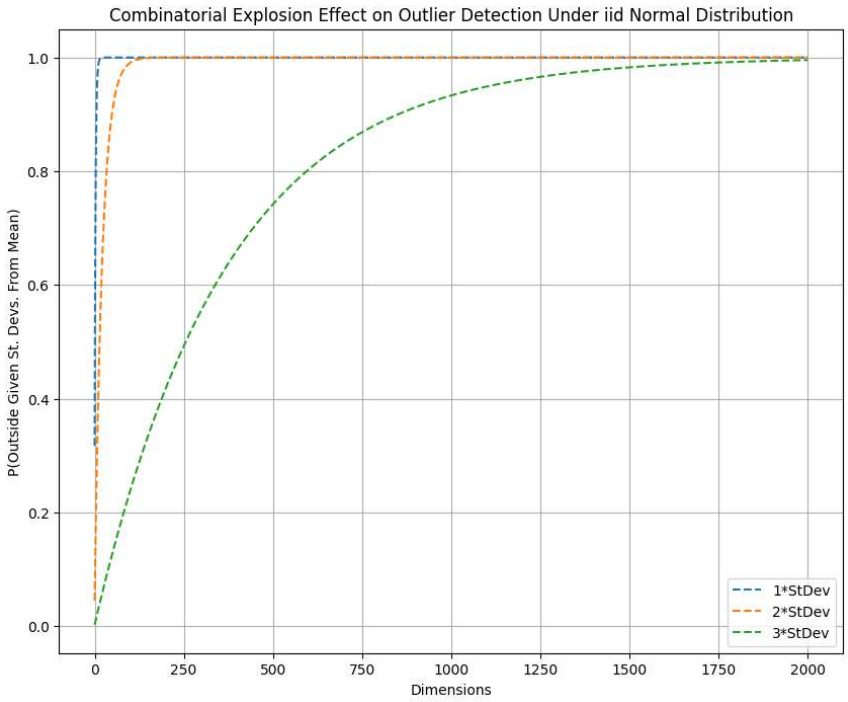
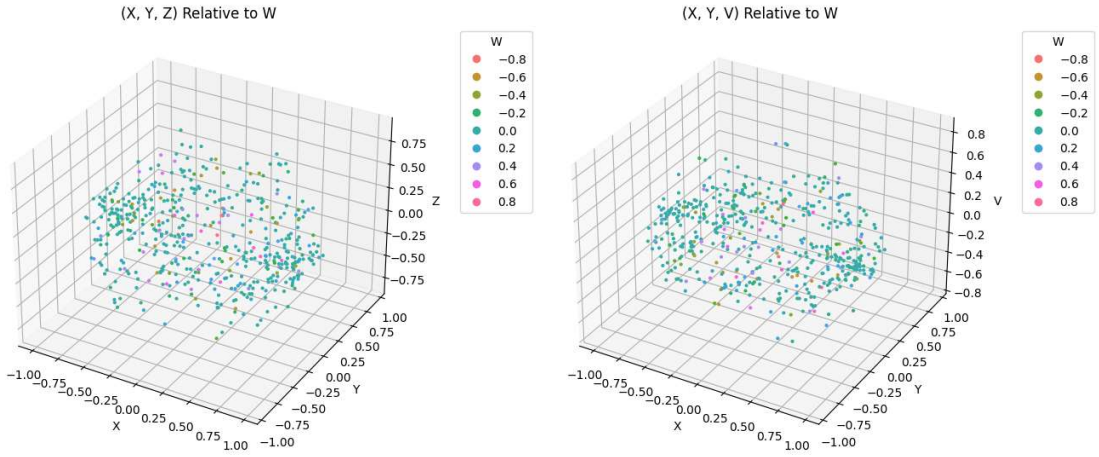


Figure 2: Effect of selecting different subspaces on determining "neighborness".

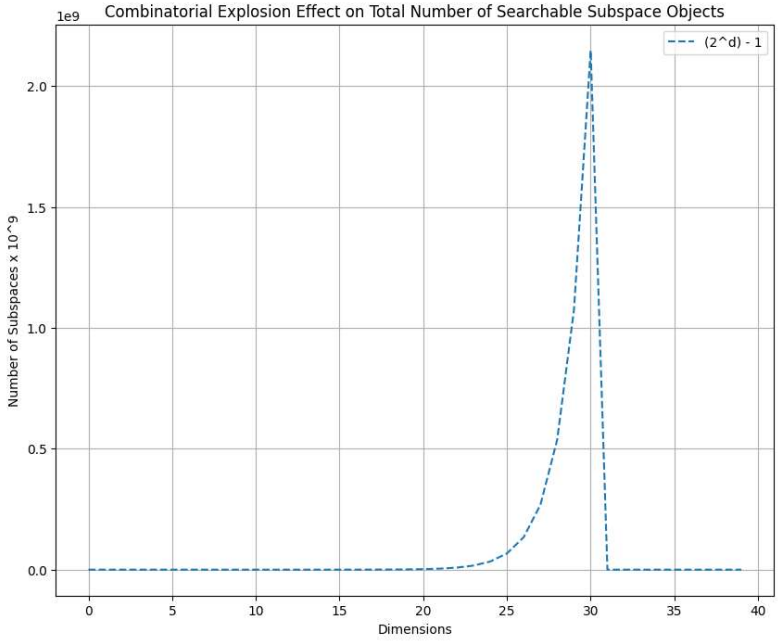


5. Core Challenge #2: Subspace Selection and Combinatorial Explosion

5.1 The Combinatorial Explosion of Total Searchable Subspaces and Absolute Knowledge Bias

As previously discussed, the issue of mapping neighborhoods to subspaces and vice versa, may be solved by a total exploration of all possible combinations of dimensionalities. However, Figure 3 below exhibits the impossibility of this method:

Figure 3: Effect of increasing dimensionality on the number of possible subspaces.



For data spaces of dimensionality d, there exists a total of $2^d - 1$ possible unordered spaces to search. It suffices to say that this obviously presents computational and sanity[3] issues when attempting to select an optimal subspace. For an amusing example of the scale, allow each distinct selection from possible subspace groups of a data space with $d = 105$ to be the equivalent of 1 kg: the sun's mass is covered roughly 20 times over by the number of searchable objects; the fact that only approximately 63.77% of the solar mass is accounted for at $d = 100$ should provide a further idea of the issue we face. On a more technical note, one will begin to brush up against the integer memory overflow limit at even "low" dimensionalities, as can be seen in the combinatorial collapse at $d > 30$. Setting aside its infeasibility, an exhaustive search is also not necessarily desired. Assuming that such a search is performed, then the knowledge gained from a total understanding of the data space is poisonous to the validity of any estimates of that data space.

5.2 Incomparability of Outlier Scores Between Subspaces of Different Dimensionalities

Referring back to Theorem 2, Zimek et al. discuss how the given formula continues to hold in situations where the Weak Law of Large Numbers fails. Such a situation may be described by a scenario where the marginal distance distribution is not static for marginal adjustments in dimensionality. Unfortunately, this describes the solution to the components of the first core challenge. Within one bootstrapped sample, the respective result of one d-dimensional bootstrap as described within Theorem 2 has no connection to the result of another j-dimensional bootstrap. Between the lines, one can surmise that these results are also incomparable for two bootstraps of equal dimensionality, but whose spaces are composed of different elements.

6. Confronting the Components of Each Core Issue

6.1 Defining Outlierness in the Data Space: Axis-Parallel or Arbitrary?

From this point onwards, there are two main paths that one could take in the approach to outlier detection: first, an assumption that outlierness is defined within the axis-parallel space; or second, that outlierness is defined by non-axis-parallel spaces. In other words, that outlierness is contained within lower-dimensional subspaces of the full-dimensional space, or that outlierness is contained within arbitrary lower-dimensional subspaces (). Methods appropriate for the former may not be appropriate for the latter, although the fact of working with arbitrary shapes may lend the

[3] See: "The Jaunt" (1981) by Stephen King from *The Twilight Zone Magazine*

latter some flexibility. For purposes of this paper, the data space of interest will define global and local outlierness as within axis-parallel and arbitrary spaces, respectively.

6.2 Possible Pre-Processing for Selection of High-Contrast Subspaces

Despite former claims that the impact of noise upon outlier detection was dropped for consideration, this was only a half-truth. Before proceeding into any form of detection process, one may wish to de-noise the data space to remove irrelevant data. A possible avenue for this method is to implement the High-Contrast Subspaces (HiCS) algorithm to produce a data space trimmed of non-overlapping and noisy dimensions. While this obviously runs the risk of some information loss, it can improve outlier score computations in data space exhibiting areas of high conditional dependence. To avoid feature creep, this paper — despite the author's best efforts — will allow the algorithm to let lie for the time being.

6.3 Proposals for Outlier Detection in High-Dimensional Space

To confront the issues at hand, this paper will lean heavily upon feature bagging and the stripping of dimensional-dependent characteristics. Through the former, one can hope to avoid the shrinking contrast and enforced sparsity problems by enforcing a fixed low-dimensionality search through the data space. Furthermore, it is possible to sidestep the combinatorial explosion of searchable spaces if the bagging is performed in enough iterations and is sufficiently random, which can permit the estimation of the full-dimensional data space characteristics. However, it is also critical that a method to define global outlierness is implemented that avoids the recursive mapping problem. To this end, an ensemble of methods that avoid a pre-clustering step will be implemented.

In order to weave the local and global outlierness of an observation together, relatively simple transformations will be applied to the local outlierness ensemble results. First, the additive frequency of features labeled as outlying for a given observation will be taken to capture filings that exhibit relatively uniform levels of anomalous activity. Second, the multiplicative frequency will be taken to amplify the abnormality of organizations that perform relatively isolated acts of highly anomalous activity. However, this latter approach is relatively naïve: it is completely dependent on not a single falsely identified outlier, which would result in the multiplicative frequency being zeroed out. A summary of all proposals can be found below:

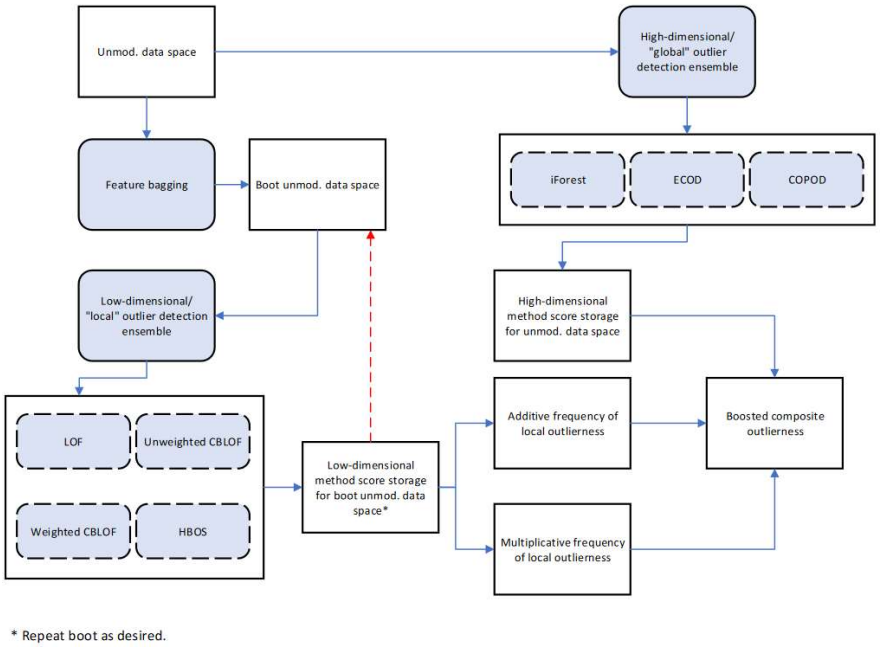
Table 1: Short-Form Proposal to Outlier Detection in High-Dimensional Spaces.

Core Issue	Component	Proposal
Homogenization of the full-dimensional data space due to CLT	Shrinking contrast between inliers and outliers.	Boosting global outlierness with local outlierness.
	Bias of L_p-derived scores towards higher dimensions.	Implement dimensionless characteristics for comparison and consolidation of outlierness.
	Enforced sparsity of the data space.	Feature bagging for local characterization.
	Recursive nature of mapping connected neighborhoods and subspaces.	1. Feature bagging for local characterization. 2. Ensemble of methods free of pre-clustering steps for global characterization.
Subspace selection and combinatorial explosion	The combinatorial explosion of total searchable subspaces and absolute knowledge bias.	Feature bagging to reduce size of searched space.
	Incomparability of outlier scores between subspaces of different dimensionalities.	Implement dimensionless characteristics for comparison and consolidation of outlierness.

7. Defining the Data Space's Path for Determination of Outlierness

At this point, reader fatigue might start to set in: your head is heavy, you're tired, and when was the last time you had some coffee? However, the most exciting elements are finally on the horizon: the components of the data pipeline. These components are outlined as follows:

Figure 4: Comprehensive Data Pipeline.



8. Analyzing the Decision Labels of the Outlier Detection Ensemble

8.1 Low-Dimensional Decisions

While the results of the LOF component in Figure 5A are rather lackluster, the remaining components in Figures 5B-D appear to demonstrate the presence of high-frequency outlying bands running across several of the features. The first band located near Features #110-120 corresponds to a region of below-mean reported values for end-of-year investments that are in "other" securities or are program-related, as well as for intangible assets and deferred revenues. Most clearly visible in the additive frequency in Figure 5E, the second band is found about Features #70-75 and appears to be the result of below-mean reported values for "other" salaries and wages, pension plan contributions, and management fees. Visible in both unweighted CBLOF and additive frequency in Figures 5B and 5E, the third band is located near Features #55-60 and might be consequence of (again) below-mean miscellaneous revenue. Finally, the surface plot of the multiplicative frequency in Figure 5F isolates several clusters of anomalous activity; ironically, one of those clusters is likely due to COVID-19 relief efforts mentioned in the introduction. However, the heatmap of the multiplicative frequency is inherently not particularly informative.

8.2 High-Dimensional Decisions

Unsurprisingly, the consolidated results of the high-dimensional methods — as can be seen in Figures 6 and 7 — tell us little that is not already known by trivial statistical assumption: most filings are not labeled as global outliers. However, that is not to say that the high-dimensional ensemble provides nothing of informative value. Despite identical starting contamination values of 10%, one can see that edge cases exist to produce low and medium levels of global outlierness. Fortunately, this permits one to explore the possibility of first sorting filings by their global outlierness, then afterwards sorting adjusting for local outlierness.

Figure 5A: Frequency of local outlierness under LOF.

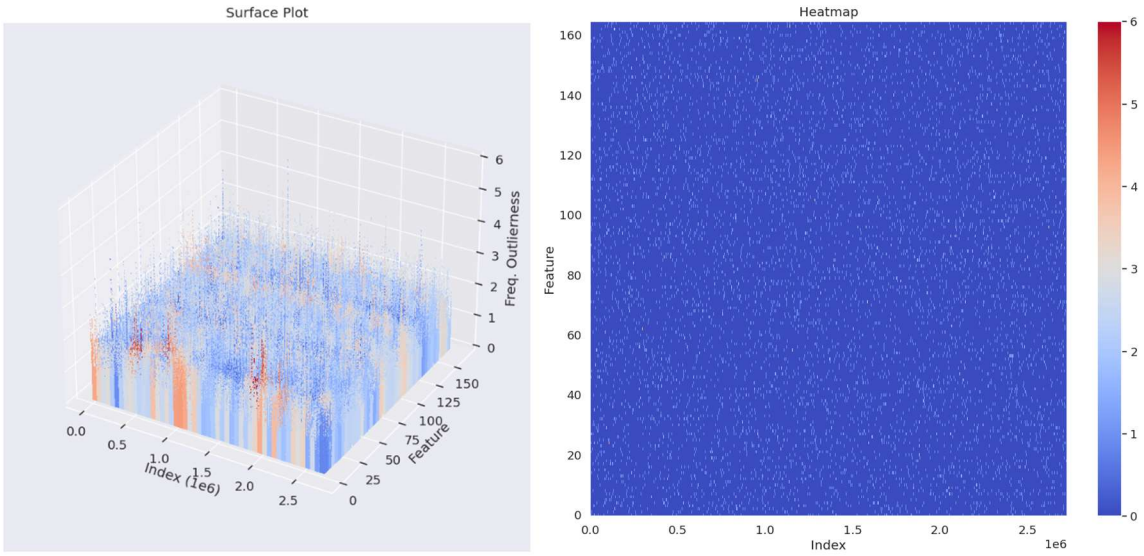


Figure 5B: Frequency of local outlierness under unweighted CBLOF.

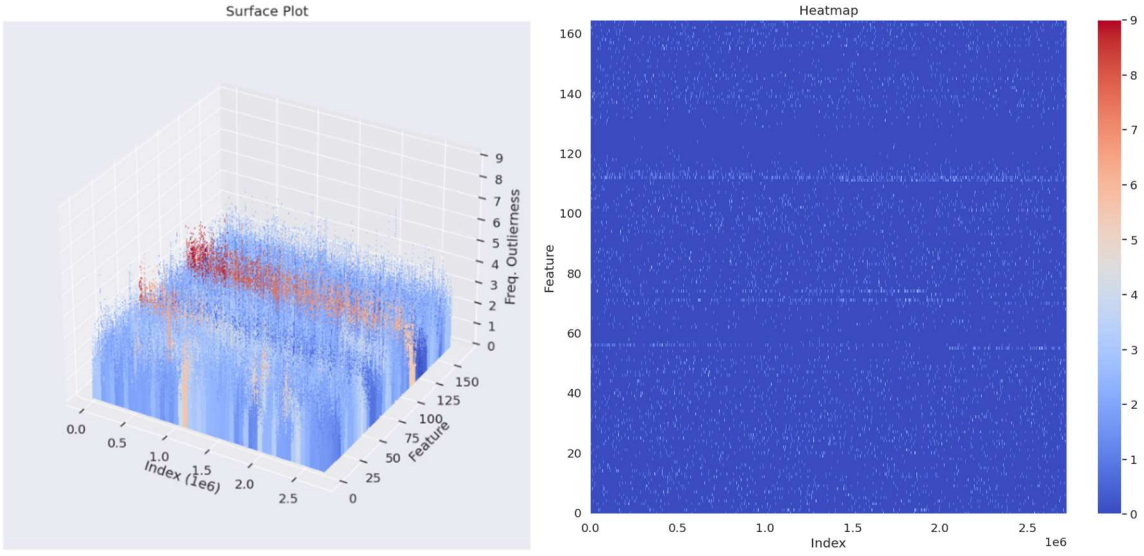


Figure 5C: Frequency of local outlierness under weighted CBLOF.

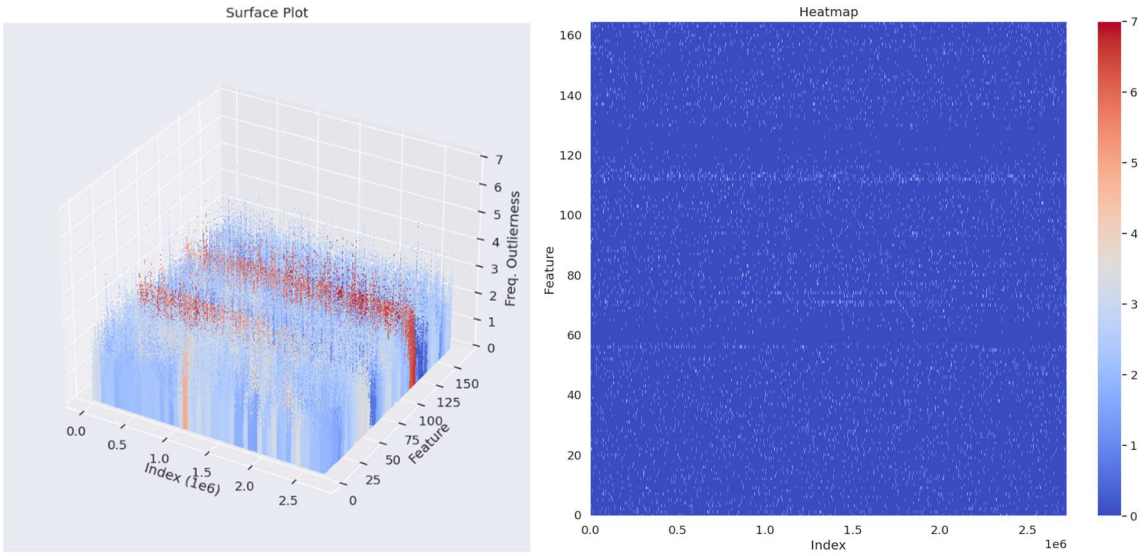


Figure 5D: Frequency of local outlierness under HBOS.

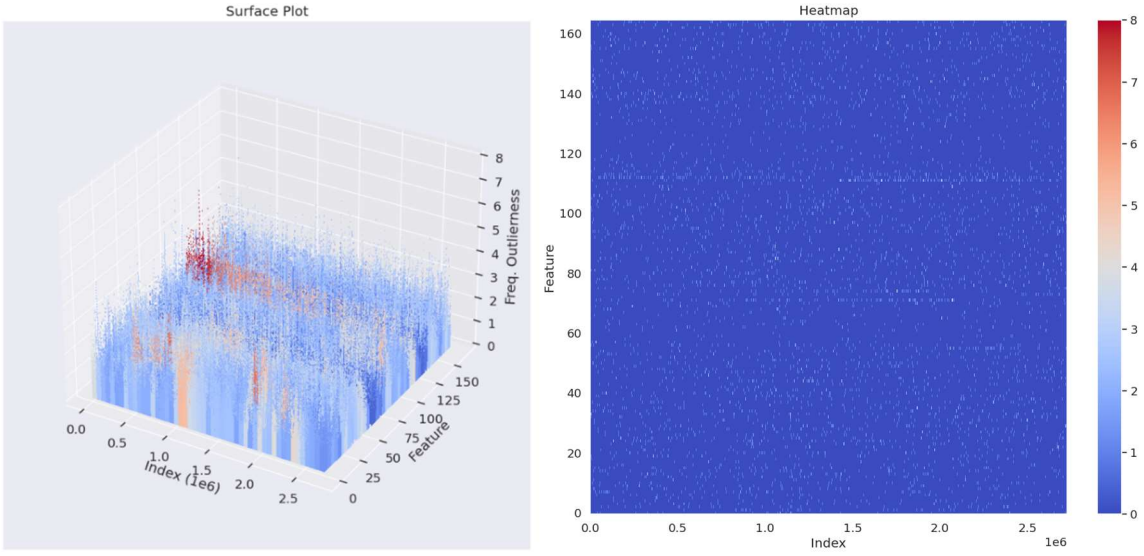


Figure 5E: Additive frequency of local outlierness ($LOF + uwCBLOF + wCBLOF + HBOS$)

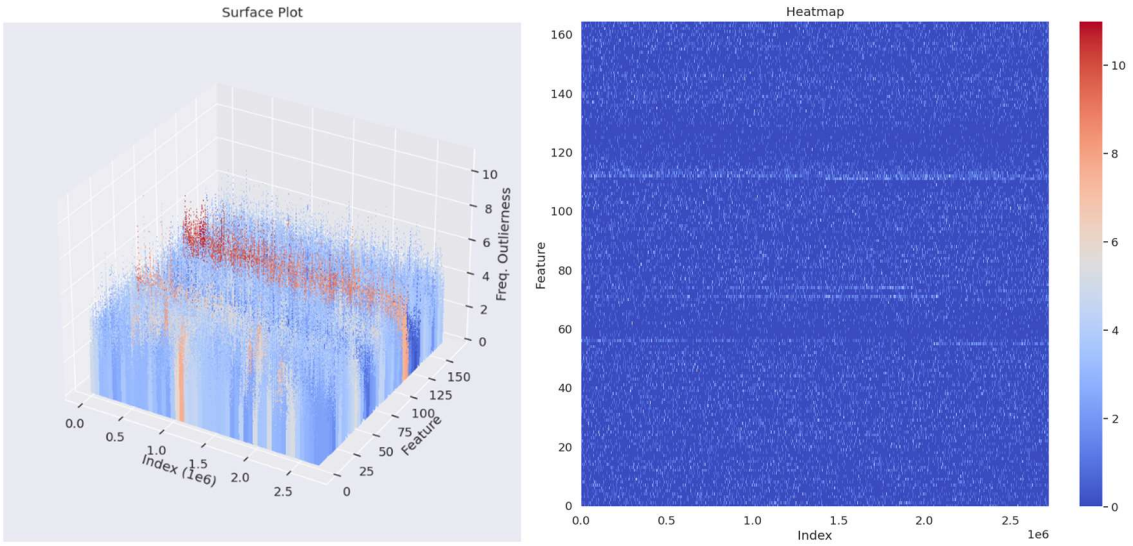


Figure 5F: Multiplicative frequency of local outlierness ($LOF * uwCBLOF * wCBLOF * HBOS$)

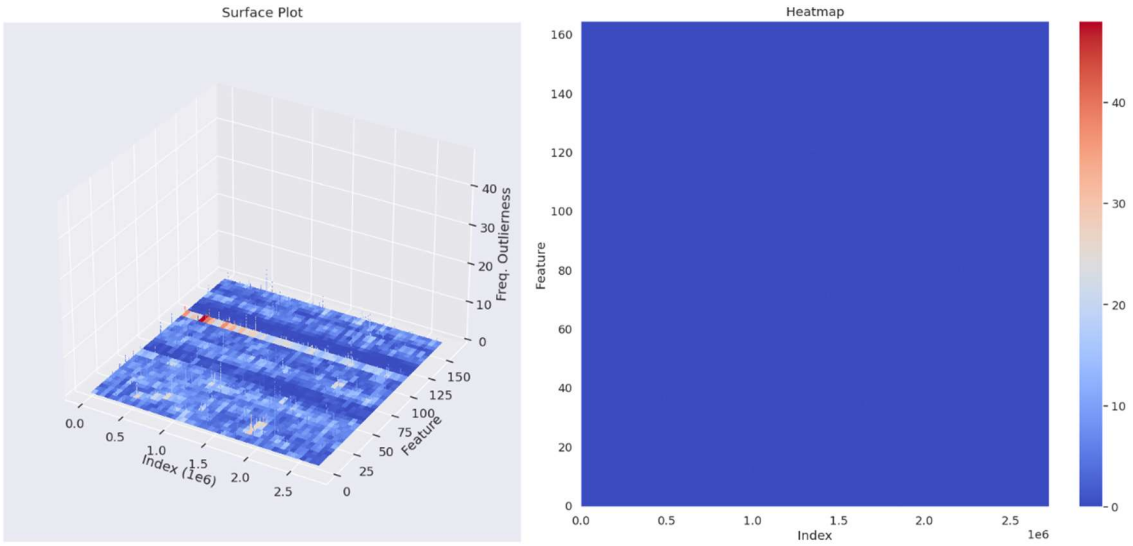


Figure 6: Total frequency of outlierness across high-dimensional methods

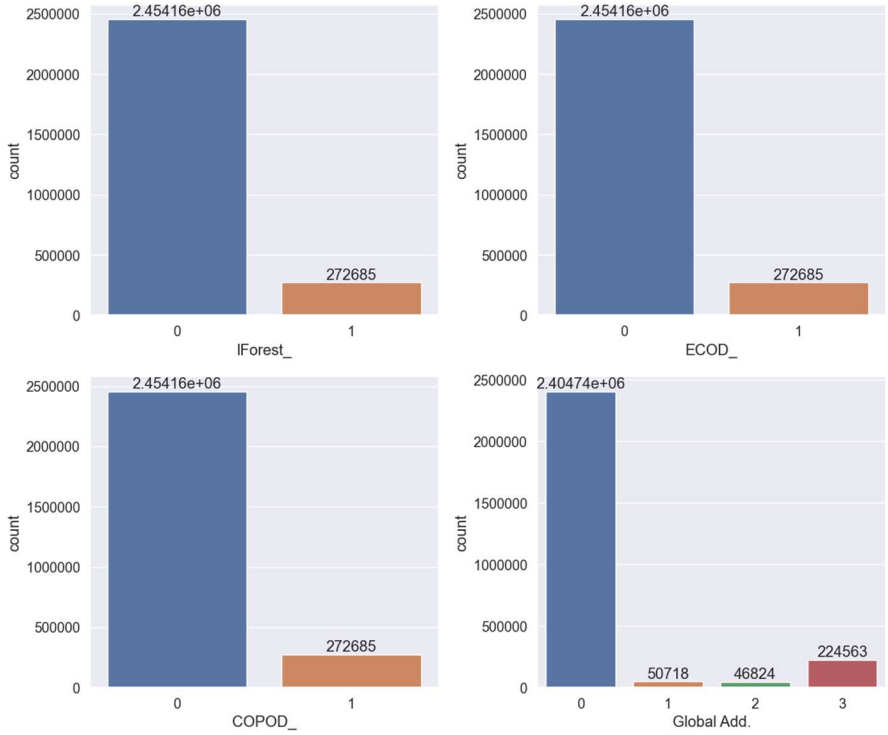
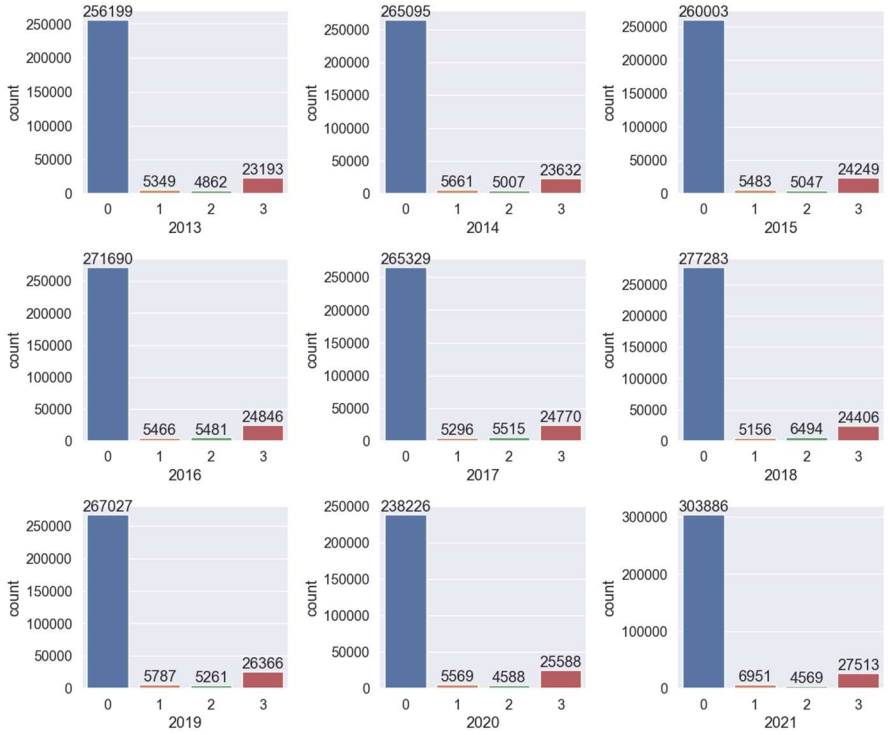


Figure 7: Annual total frequency of outlierness across high-dimensional methods

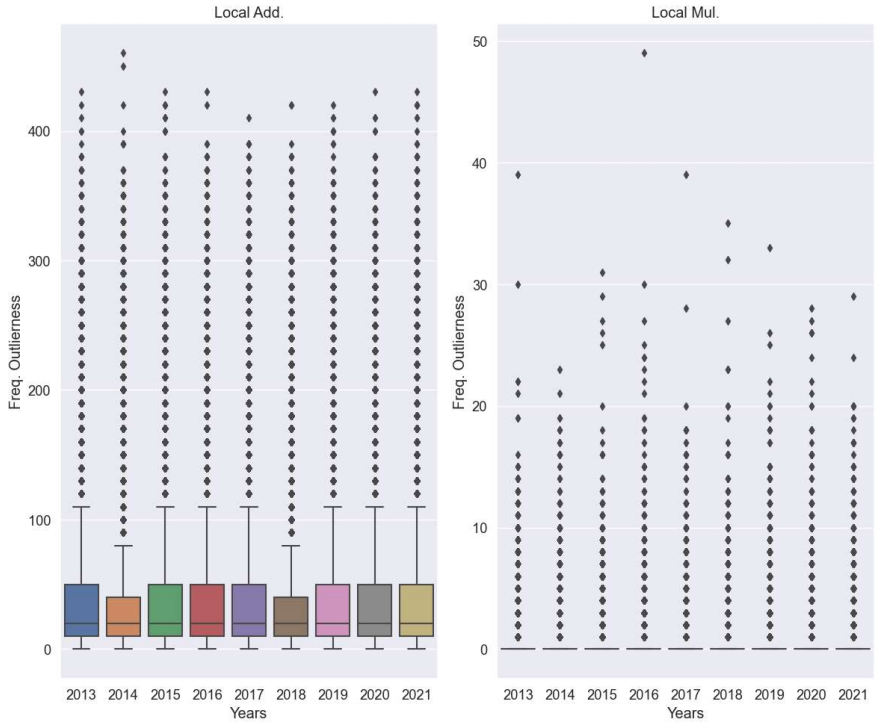


9. Identifying Organizations Exhibiting Anomalous Activity

9.1 Non-Composite Identification

To serve as an approximate measure for anomalous activity within the non-composite relative space, this paper will implement thresholds for anomalous activity relative to the disruption in tailing observations. These thresholds may be seen for local additive and multiplicative frequencies of outlierness at $Frequency\ Outlierness_{Add} \gtrapprox 350$ and $Frequency\ Outlierness_{Mul} \gtrapprox 20$, respectively. While such rule-of-thumb demarcation points might reek of shady statistics, identification of organizations by this metric in Tables 1-3 is intended as a rough baseline to compare composite identification against.

Figure 8: Boxplot of Annual Local Additive and Multiplicative Frequency



Across Tables 2-4, organizations exhibiting high levels of outlierness are concentrated within the 501(c)(3) subsector. In reference to the influential educational, financial, and medical institutions captured, it is unsurprising that these organizations dominate the outlierness ranking. Due to their substantial capital holdings and relative membership, institutions like the Kaiser Foundation and Columbia are outsized on even a national level. However, that is not to say that there are not any observations of interest: namely, the 2013 filing for the Forest Lawn Memorial-Parks and 2020 filing

for the Project Management Institute. For the former, the organization operates six "memorial-parks" and four mortuaries in the Greater Los Angeles Area, but likely exhibits itself as outlying due to a high proportion of celebrity burials on its various premises (Melville, 2022); private plots in the middle of an urban center are not known for their affordability. For the latter, the PMI membership numbers sat just shy of 700,000 in 2021, while revenues and expenses rest at approximately $340,000,000 and $280,000,000, respectively; similar to earlier institutions of influence mentioned, this likely explains their presence.

However, the question still stands as to why 501(c)(3) organizations dominate the rankings. Unfortunately, one can begin to detect the presence of a systematic error in the methodology for measurement of local outlierness. Instead of stratifying by subsector, the ensembles were applied to the entire data space without regard for lumping subsectors together. Given that 501(c)(3)s are the most common form of indicated subsector at approximately 75% of all entries, at a baseline level their presence in outlierness rankings is outsized relative to their peer non-profits. Combined with a tendency for relatively large capital and membership quantities for certain entities within this category, it is unsurprising that they are pushed even higher in representation.

Table 2: EINs of 2020-21 filings where $Frequency\ Outlierness_{Add} \gtrapprox 350$.

Year	EIN	Add. Local	Name	Subsector
2021	112965586	380	Northwell Healthcare Inc	3
2021	133273402	380	The Institute for Urban Family Health, Inc	3
2021	310568628	390	General Electric Credit Union	14
2021	330672915	390	Loma Linda University Health Care	3
2021	366066772	380	Alliant Credit Union	14
2021	382947657	380	Henry Ford Macomb Hospital Corporation	3
2021	520954463	410	Washington Hilton	3
2021	530242652	420	The Nature Conservancy, Inc.	3
2021	582149128	380	Medical Center of Central Georgia Inc	3
2021	742044647	420	National Jewish Health, Inc.	3
2021	752613493	390	Texas Health Physicians Group	3
2021	815086187	420	Blue Meridian Partners, Inc.	3
2021	900779996	380	Childrens Healthcare of Atlanta Inc	3
2021	941156365	410	Board Of Trustees of The Leland Stanford...	3
2021	941340523	430	Kaiser Foundation Health Plan, Inc.	3
2021	951642394	400	University Of Southern California	3
2021	951644600	420	Cedars Sinai Medical Center	3
2021	952096402	380	Los Angeles Opera Company	3
2020	131624158	400	The Rockefeller University	3
2020	135598093	380	Columbia University	3

Table 3: EINs of 2020-21 filings where $Frequency\ Outlierness_{Mul} \gtrapprox 20$.

Year	EIN	Mul. Local	Name	Subsector
2021	390123480	24	Thrivent Variable Annuity Account A	8
2021	541071867	29	Inova Health System Foundation	3
2020	590545223	26	Withlacoochee River Electric Cooperative Inc	12
2020	912145484	26	Verity Health System	3
2020	740650998	28	Gecu Credit Union	14
2020	541141503	22	Community Alternatives, Inc.	3
2020	231887442	24	Project Management Institute, Inc.	6
2020	530259696	27	Optical Society of America	3
2020	111667765	26	Southampton Hospital Association	3
2020	980123241	21	International Olympic Committee	4

Table 4: EINs of 2013-21 filings where both thresholds are exceeded.

Year	EIN	Local		Name	Subsector
		Add.	Mul.		
2019	941340523	400	26	Kaiser Foundation Health Plan, Inc.	3
2018	990073480	420	35	Kamehameha Schools	3
2015	986001141	400	25	Governing Council of The University of Toronto	3
2015	131656633	410	31	The Institute of Electrical and Electronics Engineers...	3
2013	950743320	380	30	Forest Lawn Memorial Parks - And Mortuaries	13

9.2 Composite Identification

9.2.1 Constructing the Composite Outlierness Ranking System

With the information needed to fulfill the primary and secondary objectives, this paper now finally moves towards exploring possible avenues to achieve the tertiary goal. By implementing the Spearman rank-order correlation versus the more traditional Pearson, one can leverage its convenient nonparametric properties to avoid statistical pitfalls from the inherently heavy-tailed distribution of outliers. To this end, a scheme is proposed for the construction of a composite ranking system:

Formula 1: Composite global outlierness ranking with Spearman's rank-order correlation weighting.

Where the Spearman's correlation is defined as $Corr_{Spear}$, then let the Spearman's correlation of additive, Frq_{Add}, and multiplicative, Frq_{Mul}, outlierness frequencies be...

$$Frq_{LocSP} = Corr_{SP}(Frq_{Add}, Frq_{Mul})$$

Then, let the rank-order correlation of global outlierness frequency, $Freq_{Global}$, to both Frq_{Add} and Frq_{Mul} be...

$$Frq_{GloSP,Add} = Corr_{SP}(Frq_{Global}, Frq_{Add})$$
$$Frq_{GloSP,Mul} = Corr_{SP}(Frq_{Global}, Frq_{Mul})$$

For any given rank, R, allow the overall composite ranking system, $R_{Composite}$, to be…

$$R_{Composite} = R_{Global} + \frac{Weights}{3}$$
$$Weights = \left(Frq_{GloSP,Add} * Frq_{Add}\right) + \left(Frq_{GloSP,Mul} * Frq_{Mul}\right) + \left(Frq_{LocSP} * \left(Frq_{Add} + Frq_{Mul}\right)\right)$$

…and for the annualized composite ranking system, $R_{Composite,t}$, over T years to be…

$$R_{Composite,t} = R_{Global} + \frac{Weights_t}{3}$$
$$Weights = \boldsymbol{MinMax}\left(Frq_{GloSP,Add,t} * Frq_{Add}\right) + \boldsymbol{MinMax}\left(Frq_{GloSP,Mul,t} * Frq_{Mul}\right) + \boldsymbol{MinMax}\left(Frq_{LocSP,t} * \left(Frq_{Add} + Frq_{Mul}\right)\right)$$
$$\forall\, t = \{1,2,\dots,T\}$$

By implementing Formula 1, filings can move freely within their global outlierness bins without breaking into neighboring bins. For the first two components of the weighting process, the formula controls for the strength and direction of the dependence between global outlierness and additive/multiplicative frequency. Meanwhile, the final component controls for the possibility of a feedback loop between the two local frequencies in situations where a filing possesses only affirmative outlying decisions. For both overall and annual periods, the values of the Spearman rank-order correlation and their corresponding statistical significances are as follows:

Table 5: Overall and annual Spearman's rank-order correlation coefficients

Type	**Period**	Frq_{LS}		$\boldsymbol{Frq_{GlobalSpear,Add}}$		$\boldsymbol{Frq_{GlobalSpear,Mul}}$	
		Value	*P-Value*	*Value*	*P-Value*	*Value*	*P-Value*
Annual	2013	0.197750	0.00	0.533107	0.00	0.302446	0.00
	2014	0.200999	"	0.532633	"	0.311822	"
	2015	0.200884	"	0.538277	"	0.305561	"
	2016	0.202422	"	0.535755	"	0.307231	"
	2017	0.204571	"	0.540018	"	0.311012	"
	2018	0.187298	"	0.533410	"	0.287467	"
	2019	0.206808	"	0.548253	"	0.308801	"
	2020	0.212651	"	0.562606	"	0.310559	"
	2021	0.201833	"	0.532977	"	0.311741	"
Overall	—	0.201691	"	0.539479	"	0.306478	"

9.2.2 Overall and Annualized Composite Ranking

Finally, one can now apply the previous formula and computed Spearman correlations to a transformation of the original global outlierness rank. For the composite rankings seen in Tables 6-7, allow the following conditions to arbitrarily define the "level" of anomalous activity:

$$Level\ of\ anomalous\ activity = \begin{cases} 0 \leq Composite < 1 \mapsto "Neglible" \\ 1 \leq Composite < 2 \mapsto \quad "Low" \\ 2 \leq Composite < 3 \mapsto "Medium" \\ 3 \leq Composite < 4 \mapsto \quad "High" \end{cases}$$

Table 6: Top 10 overall composite rankings for "High" global outlierness.

Composite Rank	Year	EIN	Name	Subsector
3.902263	2013	360724690	American Dental Association	6
3.872153	2018	990073480	Kamehameha Schools	3
3.861594	2016	611178286	Prospect Crozer, Llc	3
3.829799	2017	310939757	Centerstone Of Kentucky, Inc.	3
3.827551	2015	131656633	The Institute of Electrical and Electronics Engineers...	3
3.782387	2019	232730785	Wellspan Medical Group	3
3.776545	2013	950743320	Forest Lawn Memorial Parks - And Mortuaries	13
3.775421	2019	941340523	Kaiser Foundation Health Plan, Inc.	3
3.767894	2015	986001141	Governing Council of The University Of Toronto	3
3.754525	2021	541071867	Inova Health System Foundation	3

Table 7: Top 10 annualized composite rankings for "High" global outlierness.

Composite Rank	Year	EIN	Name	Subsector
3.884089	2013	360724690	American Dental Association	6
3.857283	2016	611178286	Prospect Crozer, Llc	3
3.833812	2017	310939757	Centerstone Of Kentucky, Inc.	3
3.826416	2018	990073480	Kamehameha Schools	3
3.820648	2015	131656633	The Institute of Electrical and Electronics Engineers...	3
3.792223	2019	232730785	Wellspan Medical Group	3
3.785075	2019	941340523	Kaiser Foundation Health Plan, Inc.	3
3.779360	2020	591479658	Adventist Health System/sunbelt, Inc.	3
3.761459	2015	986001141	Governing Council of The University Of Toronto	3
3.760828	2013	950743320	Forest Lawn Memorial Parks - And Mortuaries	13

From the results above, it appears that the rigid global outlier scores — which previously corresponded only to discrete measures — have been converted to more flexible continuous values. In other words, the tertiary objective of boosting global outlierness to incorporate a locality-sensitive components may have been fulfilled. Unfortunately, the rankings remain dominated by primarily 501(c)(3)s due to their high representation in the data space.

10. Future Work

Throughout the course of this paper, several substantial issues impeded a more thorough understanding of the data space. Foremost, an insufficient capacity for computation resulted in major deviations from the planned state of the data space going into the outlier detection ensembles. These deviations included dropping interactions between dimensions, removing qualitative variables, and an inability to compute determinations for outlierness under COF, LOCI, ROD, and SOD. Investigators in future efforts may wish to procure systems with higher computational capacities than the author's laptop.

Outside of direct problems, those pursuing similar goals using this data space may wish to make some adjustments to the methodologies implemented here. First, the iForest method might serve as a reasonable approximation of a ground truth set given enough iterations. Second, inflation adjustments applied to the data space could —instead of generalizing across organizations by year — more precisely identify the manner and forms of products and services listed under each organization's filings; this may be an unwieldy task given the sheer quantity of such organizations. Finally, it would be prudent for any investigator to implement outlier analysis on subsector-unique subsets of the data space. As discussed previously, failing to do so will result in the domination of more subtle outlying observations due to the tendency of certain financial, health, and educational systems 501(c)(3)s to possess relatively large quantities of capital and members.

11. Conclusion

While the paper appears to have succeeded in accomplishing all three listed objectives, it remains that the failure to stratify the data space by subsector has severely harmed the final composite rankings. It is entirely plausible that echoes of outlying activity have been squashed by the outsized resources available to organizations in entirely different subsectors. Unfortunately, the rankings cannot be broken up by subsector post-facto consequence of their form of construction: the local outlierness frequencies forming the foundation of the boosted composite scores are computed relative to the rest of the data space. In future efforts, investigators would do well to heed the methodological errors incurred over the course of this paper. Regardless, this area of study is an area of rapid advancements and the author's first foray into the area was — if not slightly underwhelming in terms of results — incredibly enlightening.

12. References

Aggarwal, Charu (2017). "High-Dimensional Outlier Detection: The Subspace Method." *Outlier Analysis*, 2nd ed., Springer International Publishing, Cham, 149–184.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *ACM SIGMOD Record*, *29*(2), 93–104. https://doi.org/10.1145/335191.335388

Cowley, S. (2021, August 17). *15% of Paycheck Protection Program Loans Could Be Fraudulent, Study Shows*. The New York Times. https://www.nytimes.com/2021/08/17/business/ppp-fraud-covid.html

He, Z., Xu, X., & Deng, S. (2003). Discovering Cluster-Based Local Outliers. *Pattern Recognition Letters*, *24*(9–10), 1641–1650. https://doi.org/10.1016/s0167-8655(03)00003-5

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: Copula-based outlier detection. *2020 IEEE International Conference on Data Mining (ICDM)*. https://doi.org/10.1109/icdm50108.2020.00135

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. (2022). ECOD: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2022.3159580

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413-422. https://doi.org/10.1109/icdm.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. https://doi.org/10.1145/2133360.2133363

Markus Goldstein and Andreas Dengel. Histogram-Based Outlier Score (HBOS): A Fast Unsupervised Anomaly Detection Algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012.

Melville, G. (2022, September 29). *Inside the Disneyland of Graveyards*. Smithsonian.com. http://www.smithsonianmag.com/history/inside-the-disneyland-of-graveyards-180980510/

Radovanovic, Milos & Nanopoulos, Alexandros & Ivanovic, Mirjana. (2010). Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*. 11. 2487-2531.

Traxl, D., Boers, N., & Kurths, J. (2016). Deep Graphs—A General Framework to Represent and Analyze Heterogeneous Complex Systems Across Scales. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *26*(6), 65303. https://doi.org/10.1063/1.4952963

Zimek, A., Schubert, E. and Kriegel, H.-P. (2012), A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analy Data Mining*, 5: 363-387. https://doi.org/10.1002/sam.11161