Inapplicability of the TVOR method to USHMM Data Outlier Identification

Melkior Ornik*

Abstract

Recent paper "TVOR: Finding Discrete Total Variation Outliers Among Histograms" introduces the Total Variation Outlier Recognizer (TVOR) method for identification of outliers among a given set of histograms. The method relies on comparing the smoothness of each given histogram, given by its discrete total variation, to those of other histograms in the dataset, with the underlying assumption that most histograms in the data set should be of similar smoothness. The paper concludes by applying the TVOR model to histograms of ages of Holocaust victims produced using United States Holocaust Memorial Museum (USHMM) data, and purports to identify the list of victims of the Jasenovac concentration camp as potentially suspicious. In this paper, we show that the TVOR model and its assumptions are grossly inapplicable to the considered dataset. Namely, the dataset does not satisfy the model's critical assumption of the shared smoothness between distributions of the victims' ages across lists, the model is biased in assigning a higher outlier score to histograms of larger sizes, and the dataset has not been reviewed to remove obvious data processing errors, leading to duplication of hundreds of thousands of entries when performing the data analysis.

1 Introduction

Focusing on the problem of identifying compromised data and anomalous behavior, the recently published article "TVOR: Finding Discrete Total Variation Outliers Among Histograms" [1] introduces a novel method named Total Variation Outlier Recognizer (TVOR) for identification of outliers across a set of histograms. The key enabler of the TVOR method is the fact that histograms that describe collections of samples from the same probability distribution should be, after normalization for size, similarly smooth. The method formally describes the smoothness of a histogram as the sum of differences between the number of samples in adjoining bins, i.e., the histogram's discrete total variation (DTV). Guided by asymptotic reasoning, the TVOR method fits the DTV values across histograms to a combination of a linear and a square root function of the number of data points in each histogram. The difference between the histogram's actual DTV and its expected DTV — calculated from the best fit on the entire dataset — is then, after normalization, considered as the histogram's *outlier score*.

As noted by [1], the TVOR method, as many other methods for outlier detection in histogram sets [2, 3, 4], critically relies on the assumption that all histograms in a dataset should come from the same probability distribution, or — in the case of TVOR — should at least have the same smoothness properties. Under such an assumption, [1] displays the strength of TVOR on synthetic data, with most data produced by the same probability distribution, and outliers inserted using another probability distribution. The authors of [1] then use the German census of 1939 [5], generating multiple histograms by randomly choosing subsets of the available data. The TVOR method again performs well — we note that the histograms are all drawn from the same source and thus explicitly follow the same distribution.

Following initial experiments, the work in [1] then proceeds to consider data about Holocaust victims made available by the United States Holocaust Memorial Museum (USHMM) [6]. By comparing histograms obtained from 7601 historical documents such as census records, lists of ghetto inhabitants, lists of casualties,

*Coordinated Science Laboratory and Department of Aerospace Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: mornik@illinois.edu)

and concentration camp population lists — including the lists of victims of the Jasenovac concentration camp [7] differentiated by ethnicity — the authors of [1] claim to have detected "the potentially problematic parts of a sample, which in the case of the Jasenovac list lies in the birth years of Serbian inmates" [1, Appendix D].

In this comment paper we show that the use of TVOR on the USHMM records, in the manner employed in [1], is inappropriate. We identify multiple features that make TVOR and its assumptions inapplicable to the USHMM data, and the data analysis itself faulty:

(i) the critical assumption of the histograms being drawn from distributions with similar smoothness is incorrect; and

(ii) the model, when applied to the particular dataset, is biased towards providing a higher outlier score to larger lists such as the Jasenovac list; and

(iii) the method used in [1] to pull the data from the USHMM database significantly contaminated the dataset by duplicating hundreds of thousands of entries.

The outline of the paper is as follows. In Sec. 2 we recount the formal definition and assumptions of the TVOR model from [1]. Sec. 3 provides an overview of the USHMM dataset used in [1] and the conclusions of the TVOR analysis. We then proceed to describe why TVOR is not applicable to the USHMM dataset. Sec. 4 describes the inapplicability of the assumption of TVOR on a common probability distribution or smoothness levels. Sec. 5 discusses the bias of the TVOR results with respect to the size of each list. Finally, Sec. 6 discusses data processing issues and the flaws of the separate validation of issues in the Jasenovac list claimed in [1].

1.1 Notation.

Symbol $\mathbb{N}$ denotes all positive integers, while $\mathbb{N}_0$ denotes all nonnegative integers. For $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \ldots, N\}$.

2 TVOR Preliminaries

In this section, we provide a brief overview of the Total Variation Outlier Recognizer (TVOR) method introduced in [1]. We refer the reader to [1] for a full theoretical discussion of the method; in this section we focus on the method's fundamentals and underlying assumptions.

TVOR seeks to identify outliers in a finite set of histograms $\{h_1, \ldots, h_k\}$. A *histogram* is defined as a function $h : [n] \to \mathbb{N}_0$, where $h(i)$ signifies the amount of *samples* in a *bin* i. We refer to the sum $\sum_{i=1}^{n} h(i)$ as the *size* of the histogram.

A common purpose of a histogram — and the one adopted by [1] — is to provide an empirical estimate of the probability distribution generating an underlying random variable. Namely, if k samples are taken with a probability distribution $P : [n] \to [0, 1]$ on $\{1, \ldots, n\}$, and $h^k(i)$ indicates the number of samples that equal i, then by the law of large numbers $h^k(i)/k$ almost surely converges to $P(i)$ as $k \to \infty$, for every $i \in [n]$.

Naturally, *if histograms are drawn with samples from the same probability distribution*, we can expect that — after normalizing for their size — these histograms will be similar. Thus, if one histogram in a set is significantly different from the remaining ones, it might not have been generated by the same probability distribution as the rest. TVOR is a method to quantify the difference of a particular histogram from a set of histograms. It does so by measuring the *discrete total variation* (DTV) of the histogram, and comparing it with the normalized DTVs of the other histograms in the set.

Definition 1. *A* discrete total variation *(DTV) of a histogram* $h : [n] \to \mathbb{N}_0$ *is given by* $\|h\|_V = \sum_{i=2}^{n} |h(i) - h(i-1)|$.

Definition 2. *An* expected DTV *for a set of histograms* $\{h_1, \ldots, h_k\}$*, where* $h_i : [n] \to \mathbb{N}_0$*, is a function*

$$m(N) = aN + b\sqrt{N}$$

which best fits the set $\{(N_i, \|h_i\|_V) \mid i \in [k]\}$*, where* N_i *is the size of histogram* h_i*.*

While [1] does not explicitly state the metric of best curve fit for the expected DTV, the presented results indicate that the experiments consider a standard least squares fit or a similar metric.

The TVOR outlier scoring d' for a histogram $h : [n] \to \mathbb{N}_0$ of size N in the set $\{h_1, \ldots, h_k\}$ is then proposed in [1] by normalizing the difference of its DTV with the expected DTV for a histogram of its size:

$$d' = \frac{|\|h\|_V - m(N)|}{\sqrt{N}}. \tag{1}$$

In parallel with TVOR, [1] also proposes its modification

$$d'' = \frac{|\|h\|_V - \hat{\mu}_N|}{\hat{\sigma}_N}, \tag{2}$$

where instead of $m(N)$ and $\sqrt{N}$, the model considers the mean and standard deviation of DTVs of appropriately standard subsets of the German census of 1939 [5], considered in [1] to be a "gold standard" [1, Appendix D].

3 USHMM Dataset

After an initial theoretical discussion of TVOR and experiments on synthetic data and data constructed by subsets of the German census of 1939 [5], the work in [1] focuses on the dataset compiled from the United States Holocaust Memorial Museum (USHMM) lists of Holocaust victims. The vast majority of the data used in the experiments of [1] is provided by the authors online at [8]. All the data that we use in our experiments is contained in that repository.

The full dataset consists of years of birth for more than 3.6 million individual records across 7601 lists. The lists are obtained from historical documents of varying provenance, including ghetto inhabitant lists, death lists, concentration camp lists, and census records [6]. The experiment in [1] pays particular attention to the list of victims of the Jasenovac concentration camp, marked in the USHMM records by ID 45409 [7].

While the full USHMM lists often contain many details about the victims, the dataset considered in [1] is concerned only with their years of birth, with a partial exception of the Jasenovac list, where victim records are further divided by ethnicity into multiple sublists. These sublists do not seem to be available at the authors' dataset repository [8], but they are illustrated in detail in [1].

Each of the 7601 lists in [8] forms a natural histogram, where each bin corresponds to a particular year of birth, between 1850 and 1945, resulting in $N = 96$ total bins. We note that some of the lists contain a small number of obviously impossible dates of birth, as mentioned later; in our experiments, following the figures in [1], we a priori remove all years not in the interval $[1850, 1945]$.

The authors of [1] then apply the TVOR method to identify the outlying histograms. They note that the Jasenovac list results in the highest d'-score, with $d' \approx 43.13$, thus labeling its data as "potentially problematic" [8, Sec. IV-F]. After adding the nationality sublists within the Jasenovac list to the histogram set, the list of Serbian victims also yields a high d'-score of $d' \approx 40.82$ and is also described as potentially problematic [8, Appendix D]. Similar d'' scores are obtained for Jasenovac from (2) by replacing the histogram set statistics with the German census of 1939.

4 Inapplicability of Assumptions

We claim that the application of TVOR does not in reality indicate any potential problems with the Jasenovac list or any of its subsets. In fact, unlike the experiments on synthetic data and synthetic subsets of real census data, TVOR is entirely inapplicable to the USHMM dataset as the dataset does not even come close to the fundamental assumption of these models — that all lists in the set, except possible outliers, should share similar probability distributions or at least similar smoothness, and, in the case of the modified TVOR in (2), also share similar smoothness as the German 1939 census. Such an assumption is explicitly acknowledged in [1, Sec. I] and throughout the paper.

In their only discussion of this challenge, the authors of [1] recognize that the lists may not have similar probability distributions, but argue that "geographical locations of these populations differ, but they still mostly cover the populations whose birth year histograms should have similar discrete total variation properties" [1, Sec. IV-D2]. However, [1] does not offer further qualitative or quantitative support of such a wide-ranging statement, and only gives explanations for why particular lists may or may not be flagged as outliers in terms of DTVs.

The claim that lists across the dataset "should" have similar smoothness properties is easily disproved by a wealth of historical evidence, as well as the data analysis of [1] itself. As mentioned above, the USHMM lists have been obtained from a variety of historical documents [6] of varying quality and purpose. There is no reason to believe either the probability distributions or the smoothness of data across those documents would be similar. The quality and biases of age reporting notoriously depend on the method of data collection [9], and precise population registry books, such as, e.g., those kept by the Judenrat of the Łódź Ghetto [10] will possibly contain smoother data than those collected in the absence of written records, by oral examination [9].

The claim of shared smoothness across the dataset also breaks down when examined through the perspective of data analysis proposed in [1] itself. In [1, Sec. III-A], the authors state the following about the normalized DTV: "[...] if this score is calculated for every sample in a group of samples that are expected to have similar smoothness, then the ones with the highest scores can be considered as outlier candidates". Let us thus consider the normalized DTV score of all 16 lists similar in size to the Jasenovac list (with $39000 \leq N \leq 157000$); we choose lists similar in size to minimize the impact of possible bias in normalization that we discuss in the subsequent section.

Let $m(N)$ fit as the expected DTV of these 16 lists. If the fit was good and all lists, except perhaps Jasenovac, were to have similar smoothness, the signed d' score $d'_{\pm} = (\|h\|_V - m(N))/\sqrt{N}$ would by definition be similar *for nearly all* those lists, where $m(N)$ is fitted as the expected DTV of those lists. Fig. 1 illustrates the values of $d'_{\pm}$ for those 16 lists.

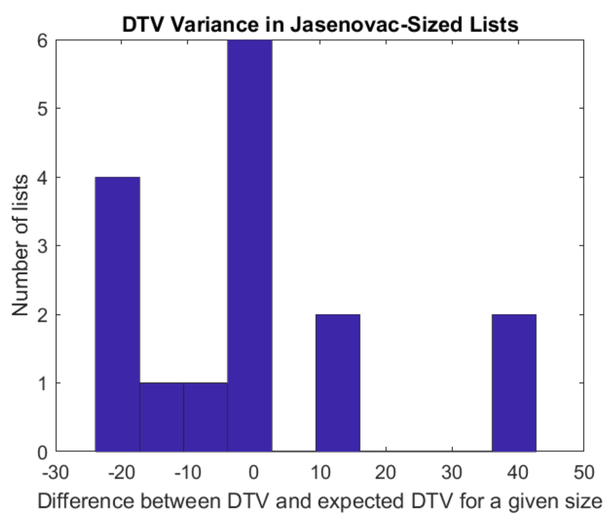

Figure 1: The distribution of the difference between the DTV and expected DTV for lists similar in size to the Jasenovac list.

As shown in Fig. 1, less than half of the considered lists have values in any interval of size 10, and the resulting empirical distribution has seemingly little similarity to a unimodal distribution. Thus, these lists *provably* do not have similar smoothness properties.

5 Histogram Size Bias

Having discussed the assumptions imposed on the dataset by applying the TVOR approach, we now move to show that — even if these assumptions were true for the considered dataset — the TVOR approach is biased towards assigning higher outlier scores to longer lists.

Exactly to avoid size bias, the outlier score is normalized in (1) by dividing the difference between the actual histogram DTV and the expected DTV by $\sqrt{N}$, where N is the histogram size. As the authors comment in [1], such a normalization is a heuristic. In the case of this dataset, it is in fact insufficient.

To show the bias, we fit a standard linear regression model on the set $\{(N_i, d'(h_i) \mid 1 \leq i \leq 7106\}$ of USHMM histograms, where N_i is the size of histogram h_i, and $d'(h_i)$ is its outlier score computed from (1). We easily find that the best fit is given by

$$\overline{d}'(N) = 4.017 \cdot 10^{-5} N + 2.128.$$

Given that the dataset contains 7106 elements, a similar fit is naturally obtained by excluding the Jasenovac list from the fitting. While the coefficient of $4.017 \cdot 10^{-5}$ may not seem significant for small N, we keep in mind that the Jasenovac list contains $N = 78493$ elements.

The identified bias is actually alluded to in [1, Appendix D] itself, in the discussion of the Jasenovac sublists. After noting that even the sublist with the highest outlier d' score has a lower outlier score than the entire list, the authors indeed attribute it to "the fact that the sample with birth years of Serbian inmates has fewer values than the whole Jasenovac list [...]. Because of that, such similar deviations are considered to be less likely on a larger sample and thus the whole Jasenovac list has a slightly larger value of score d'". They, however, still proceed to claim that a high d' score renders the Jasenovac list and the Serbian victims' sublist "potentially problematic" [1, Appendix D].

Keeping in mind the identified bias, let us perform a small numerical experiment. We renormalize d' by considering

$$d'_{ren} = \frac{d'}{\overline{d}'(N_i)}$$

to be the new outlier score; if the original scoring was unbiased, such a renormalization would have no impact. However, in this case it renders Jasenovac list's outlier score $d'_{ren} = 8.15$ *no longer the highest in the dataset*. There are multiple histograms with a higher score, the highest of which ($d'_{ren} = 14.36$) belongs to list ID 40168, a registry book for Block No. A 32 Hohensteiner Str. No. 42 in the Łódź Ghetto [11]. A different regression model motivated by [1], given by $\overline{d}'(N) = aN + b\sqrt{N} + c$, also yields a renormalization in which the Jasenovac list does not feature prominently; after such a renormalization, the Jasenovac list has a lower outlier score than more than 50 other lists.

We present the resulting renormalized outlier scores in Fig. 2. While renormalization still does not entirely remove the bias, it significantly decreases it — an average histogram of size 10^6 would get assigned an outlier score of not more than 7, in contrast to the original TVOR model which would assign an outlier score of ≈ 42.

We emphasize that *our findings do not imply that there is anything potentially problematic* with the Block No. A 32 Hohensteiner Str. No. 42 list, or any of the other lists that score higher than the Jasenovac list after renormalization. For all other reasons mentioned in this paper, the use of TVOR on the USHMM dataset cannot yield any meaningful conclusions even after renormalization.

6 Data Processing and Analysis Issues

Having clarified the modeling issues, we now discuss the errors in the dataset considered in [1]. As noted, the authors of [1] kindly placed all their raw data in an online repository [8]. However, by consulting that

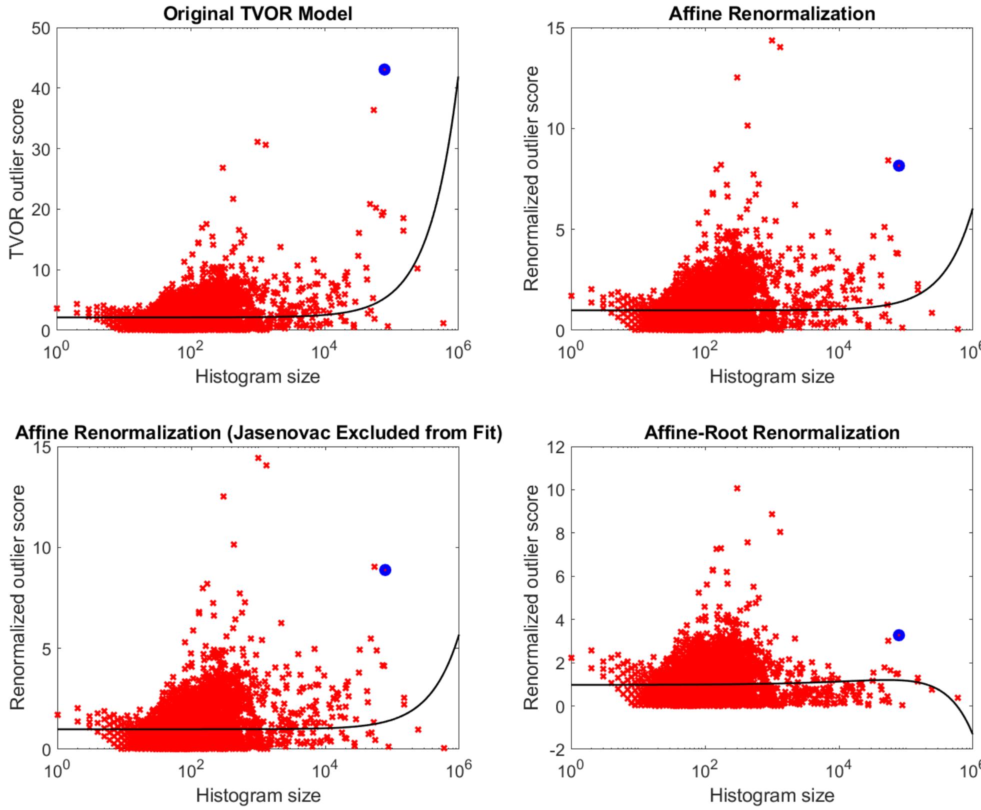

Figure 2: A comparison between the results of the original TVOR model and the results of the three renormalized models: one with the best fitting affine renormalization function $\hat{d}'$, one with the best fitting affine renormalization function after the Jasenovac list is excluded from the dataset, and one with the best fitting function of form $aN + b\sqrt{N} + c$. On all figures, the score for the Jasenovac list is marked in blue. The black line indicates the "expected value" of the outlier score for a histogram of a given size.

repository it becomes clear that it has been downloaded from the USHMM website without any consideration of potential duplication in the data stemming from USHMM storage choices.

Let us describe just several examples. The authors of [1] consider both lists ID 20492 and ID 20493 in their dataset, and even mention ID 20492 as the list with the third highest d'' score [1, Sec. IV-D3]. However, those two lists are essentially the same [12, 13]. They both contain data from the registration cards of Jewish refugees in Tashkent. Their only difference, apart from three entries (out of 148 688 in the dataset) missing from one of the lists, is that one list is in the Latin alphabet, while the other is in Cyrillic. The dataset also considers lists ID 45680 and ID 45681 separately. However, when names not associated with a year of birth are removed, those two lists are entirely the same.

The largest list in the dataset is list ID 25274, *The Elders of the Jews in the Łódź Ghetto*. However, that list is itself compiled from 14 440 sublists which have their own IDs in the USHMM registry (some without accessible years of birth) [14]. Nonetheless, the data analysis in [1] considers these lists separately, in addition to the large list. Given that the total number of lists in the USHMM registry at the time of writing [1] seems to have been around 46 000 (most without accessible years of birth), we imprecisely estimate that around one third of all lists in the dataset of [1] are actually parts of the large Łódź Ghetto list. Whatever the exact number, omissions such as the ones described skew the dataset by adding hundreds of thousands of duplicated entries.

We finally consider the element of the claims of [1] that has little to do with statistics. As corroboration of their identification of potential problems with the Jasenovac list, [1, Sec. IV-F] and [1, Sec. IV-G] mention three of its inconsistent entries. The three entries describe individuals who are apparently likely to not have been in the Jasenovac camp. Without going into the merit of the listings for those three individuals, we first note that the dataset's Jasenovac list has more than 78 000 entries; it is reasonable to expect that any list of 78 000 people will have hundreds of errors.

We then note that the authors of [1] make no effort to identify or prune similar inconsistencies in other lists in the dataset, even when seemingly more egregious and more obvious. In fact, they use those lists directly in their algorithm. For instance, list ID 20652 recording deportations to Theresienstadt is claimed at its beginning to contain exactly 42155 persons [15], but only 42150 are then listed. On the other hand, list ID 1394 is estimated on the USHMM website to contain 14 entries [16], but it actually contains 61. List ID 25274, when converted in [8], seems to contain an individual born in 1771, who would have been around 170 years old at the time of record collection. *We again emphasize that such small inconsistencies are to be expected in any project of the size of the USHMM records and do not indicate any widespread problem with the data*; we point them out only to indicate the lack of importance of the claim in [1] about the three individuals on the Jasenovac list.

Finally, we briefly comment on the discussion of *age heaping* in [1]. Indeed, age heaping — a phenomenon where the year of birth of individuals whose age is uncertain is rounded to a year ending in 0 [9] — is a common phenomenon and can only be expected in the chaotic atmosphere of wartime, concentration camp, and immediate post-war data collection and record-keeping. In Fig. 20 and Fig. 24–25 of [1], the authors, however, draw two more claimed instances of age heaping in the Jasenovac list, a "mid-decade pattern" including the years ending in 4–8, and a heaping in years ending in 2. The paper performs no statistical analysis on these patterns. In fact, the mid-decade pattern is not described anywhere in [1] other than being drawn on the figures. It is unclear how the authors of [1] established the baselines for those years to identify heaping. The same is true for heaping in years ending in 2. While the latter phenomenon is seemingly larger in quantity, the authors do not mention a plausible explanation for it in the context of the Jasenovac camp: regardless of their views on the total number of victims, a wide variety of sources agree that most victims in Jasenovac were likely killed in 1942 (e.g., [17, 18]). Thus, if the original sources recorded the victims' *age* rather than *year of birth*, age heaping will lead to a surplus of individuals seemingly born in a year ending with 2, by rounding the age at death to the nearest decade. We finally remark that the dataset in [8] fails to take into account that some of the years of birth are indeed labeled as disputed on the USHMM website.

7 Conclusion

Paper [1] introduces the Total Variation Outlier Recognizer (TVOR) method for recognizing anomalies in histogram sets. It proceeds to provide a theoretical background for the introduced method, showing that — as any other statistical model — TVOR has its benefits and drawbacks, and may yield reasonable results under particular assumptions. The work of [1] then proceeds to apply the TVOR method to the Holocaust victim lists from the United States Holocaust Memorial Museum, claiming that the list of victims of the Jasenovac concentration camp — and particularly the list of Serbian victims — is potentially problematic.

In this paper, we show that the use of TVOR to the dataset considered in [1] is meaningless. The authors of [1] disregarded the assumptions that they identified in an earlier part of their paper, providing a dataset whose elements could not possibly, and provably did not, come from the same probability distribution or distributions with similar smoothness altogether. Additionally, when analyzing the data, the normalization that they implemented in the TVOR model to avoid a scoring bias to the size of each victim list is incorrect, and the model is still biased toward labeling larger lists as more outlying ones.

We then turn towards the raw data itself and observe that the dataset analyzed in [1] was pulled from the USHMM database without due diligence, resulting in numerous partial or full duplications of the victim lists, further contaminating the model results. Finally, some of the purported patterns presented in [1], notably those on age heaping, were not obtained through the TVOR model at all, their mathematical provenance and statistical importance are unknown, and the authors of [1] fail to identify simple plausible historical and data-collection explanations for them. In conclusion, while the TVOR method may have its uses, attempted analysis of the Holocaust victim lists presented in [1] is not one of them.

References

[1] N. Banić and N. Elezović, "TVOR: Finding discrete total variation outliers among histograms," *IEEE Access*, vol. 9, pp. 1807–1832, 2021.

[2] H. Cramér, *Mathematical Methods of Statistics.* Princeton University Press, 1946.

[3] N. D. Gagunashvili, "χ^2 test for the comparison of weighted and unweighted histograms," *Statistical Problems In Particle Physics, Astrophysics And Cosmology*, pp. 43–44, 2006.

[4] K.-M. Xu, "Using the bootstrap method for a statistical significance test of differences between summary histograms," *Monthly Weather Review*, vol. 134, no. 5, pp. 1442–1453, 2006.

[5] Statistischen Reichsamt, "Statistisches jahrbuch für das Deutsche Reich: Achtundfünfzigster jahrgang 1939/40," 1940.

[6] United States Holocaust Memorial Museum, "Database of Holocaust survivor and victim names." [Online]. Available: https://www.ushmm.org/remember/resources-holocaust-survivors-victims/database-of-holocaust-survivor-and-victim-names

[7] ——, "List of individual victims of Jasenovac concentration camp (ID: 45409)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=45409

[8] N. Banić and N. Elezović, "GitHub - DiscreteTotalVariation/TVOR," 2020. [Online]. Available: https://github.com/DiscreteTotalVariation/TVOR

[9] S. O. Rutstein, G. T. Bicego, A. K. Blanc, N. Rutenberg, F. Arnold, and J. A. Sullivan, "An assessment of DHS-I data quality," Institute for Resource Development, Tech. Rep. MR1, 1990.

[10] United States Holocaust Memorial Museum, "Lodz-names: A record of the 240,000 inhabitants of the Łódź Ghetto (ID: 20619)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=20619

[11] ——, "No. A 32 Hohensteiner Str. No. 42 (ID: 40168)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=40168

[12] ——, "[RG-75.002M, registration cards of Jewish refugees in Tashkent, Uzbekistan during WWII, transliterated data] (ID: 20492)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=20492

[13] ——, "[RG-75.002M, registration cards of Jewish refugees in Tashkent, Uzbekistan during WWII] (ID: 20493)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=20493

[14] ——, "Przełożony Starszeństwa Żydow w Getcie Łódzkim (ID: 25274)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=25274

[15] ——, "Documentation of names and commemoration of the victims (ID: 20652)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=20652

[16] ——, "Deutschen kriminalpolizeiblatt (sonderausgabe) (ID: 1394)." [Online]. Available: https://www.ushmm.org/online/hsv/source_view.php?SourceId=1394

[17] L. M. Adeli, "From Jasenovac to Yugoslavism: Ethnic persecution in Croatia during World War II," Ph.D. dissertation, University of Arizona, 2004.

[18] A. Korb, "Understanding Ustaša violence," *Journal of Genocide Research*, vol. 12, no. 1–2, pp. 1–18, 2010.