

Received December 8, 2020, accepted December 18, 2020, date of publication December 24, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047342

# TVOR: Finding Discrete Total Variation Outliers Among Histograms

NIKOLA BANIĆ<sup>1</sup> AND NEVEN ELEZOVIĆ<sup>2</sup>

<sup>1</sup>Gideon Brothers, 10000 Zagreb, Croatia

<sup>2</sup>Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia

Corresponding author: Nikola Banić (nbanic@gmail.com)

**ABSTRACT** Pearson's chi-squared test can detect outliers in the data distribution of a given set of histograms. However, in fields such as demographics (for e.g. birth years), outliers may be more easily found in terms of the histogram smoothness where techniques such as Whipple's or Myers' indices handle successfully only specific anomalies. This paper proposes smoothness outliers detection among histograms by using the relation between their discrete total variations (DTV) and their respective sample sizes. This relation is mathematically derived to be applicable in all cases and simplified by an accurate linear model. The deviation of the histogram's DTV from the value predicted by the model is used as the outlier score and the proposed method is named Total Variation Outlier Recognizer (TVOR). TVOR requires no prior assumptions about the histograms' samples' distribution, it has no hyperparameters that require tuning, it is not limited to only specific patterns, and it is applicable to histograms with the same bins. Each bin can have an arbitrary interval that can also be unbounded. TVOR finds DTV outliers easier than Pearson's chi-squared test. In case of distribution outliers, the opposite holds. TVOR is tested on real census data and it successfully finds suspicious histograms. The source code is given at <https://github.com/DiscreteTotalVariation/TVOR>.

**INDEX TERMS** Age heaping, anomaly detection, discrete total variation, expected value, fitting, histogram, Myers' index, outlier detection, Pearson's chi-squared test, total variation, Whipple's index.

## I. INTRODUCTION


Outliers can be defined as data patterns that do not conform to an expected normal data behavior [1]. Since identifying outliers or anomalies can often be useful, performing outlier, i.e. anomaly, detection has an important role in many data related areas. For example, with the ever growing application of machine learning in various fields, having clean training sets, free of any unwanted outliers, can often significantly benefit the final production accuracy. On the other hand, in real-time applications such as network traffic or health monitoring, it is usually highly important to detect anomalies that could represent any form of unwanted behavior to prevent their potentially detrimental effects. Alternatively, it may be required to see which samples differ the most from the rest of the data and study them in more detail.

Since there is a relatively high demand for anomaly and outlier detection methods in fields dealing with some form

of data, numerous methods have been proposed for various applications, as can be seen in several review papers [1]–[3].

A particular kind of data are histograms. First introduced by Pearson [4], histograms are by definition estimates of the probability distribution of a continuous variable. If there is a sample of real numbers drawn from the same distribution and all inside a given interval, then histograms can be used as their simple representation, and are also suitable for visual presentation. For histograms to be useful, the bin size should be adjusted accordingly to the data being described [5]–[8]. In certain cases for a group of such histograms it may be interesting to know whether some of them are outliers. This may include histograms describing samples drawn from another distribution different from the one of the majority of the samples, but it may also include histograms just describing some less likely samples from the same distribution. To be clear, in such a case, histograms are not used as tools for outlier detection like in e.g. [9], but they are the data representations to be analyzed for the presence of outliers.

In the simple case when only a single histogram is given, instead of multiple histograms, a straightforward approach

The associate editor coordinating the review of this manuscript and approving it for publication was Giambattista Gruosso .

to check whether it represents a sample that differs from a given distribution would be to use the Pearson's chi-squared test [10]. It tests how likely it is that any observed difference between the bins counts of the given histogram and the expected bin counts occurred by chance. However, for this to work, it is required to know the expected bin counts.

On the other hand, if multiple histograms are given for samples that are assumed to have been drawn from the same distribution, then it is possible to find outliers among them by means of the Pearson's chi-squared test even if the distribution is unknown. Namely, under Glivenko-Cantelli theorem [11] all the given histograms, except the currently tested one, can be used to get a reliable empirical distribution function, which in turn can be used to get the expected bin counts. Over time, numerous other techniques that can be applied in the described cases have been proposed [12]–[15].

While the problem of finding outliers in terms of distribution is common, in some cases it is required to find histogram outliers in terms of some specific histogram property. For example, census data histograms are usually smooth, i.e. the difference between the counts of neighboring bins is relatively low, but in the presence of anomalies such as *age heap-ing* [16], this often stops being the case. One way to measure smoothness is to calculate total variation [17]. This means that by detecting deviations from the expected total variation it could be possible to detect smoothness outliers more easily than by means of some of the previously described techniques. Single-value properties similar to total variation in terms of simplicity, such as skewness, have already been used for outlier detection [18]. As a matter of fact, total variation has also found application in tasks such as classification [19] and outlier detection for graph signals [20].

Therefore, in this paper a new method for outlier detection in terms of discrete total variation (DTV) among histograms that describe samples drawn from the supposedly same, but unknown distribution is proposed. There are several contributions of this paper. First, it is mathematically proven that in terms of the underlying distribution there are only two possible cases of the relation between the sample size and the expected discrete total variation with the first case only being a special case of the second one. Second, a method is proposed that utilizes this relation to detect outliers that deviate from their expected discrete total variation. Third, it is shown that while the proposed method is not supposed to be used as a general outlier detector in terms of distribution, in some special cases it still performs better in this task than Pearson's chi-squared test. Fourth, the proposed method is shown to be able to detect suspicious histograms on real-life census data. The practical applicability and usefulness of the proposed method are shown on synthetic data and real-life census data. The proposed method is simple to implement and it does not require prior knowledge of any distribution.

The paper is structured as follows: in Section II the total variation is formally described, in Section III the theoretical derivation of the proposed method and its underlying model are given, in Section IV the experimental results obtained

on synthetic data and historical real-life census data are presented and discussed, and Section V concludes the paper.

## II. THE TOTAL VARIATION

Total variation of a differentiable function  $f$  is defined as [17]

$$\|f\|_V = \int_{-\infty}^{+\infty} |f'(t)| dt. \quad (1)$$

If  $f$  is non-differentiable, its total variation is given as [17]

$$\|f\|_V = \lim_{h \rightarrow 0} \int_{-\infty}^{+\infty} \frac{|f(t) - f(t-h)|}{|h|} dt. \quad (2)$$

If  $f_n[i] = f * \Phi_n(i/n)$  is a discrete signal obtained with an averaging filter  $\Phi_n(t) = 1_{[0, N-1]}(t)$  and a uniform sampling at intervals  $n^{-1}$ , then its discrete total variation (DTV) is calculated by approximating the signal derivative by a finite difference over the sampling distance  $h = n^{-1}$  and replacing the integral in Eq. (2) by a Riemann, which then gives [17]

$$\|f_n\|_V = \sum_i |f_n[i] - f_n[i-1]|. \quad (3)$$

Despite being relatively simple to calculate, total variation is successfully used in areas such as denoising [21]–[24], image restoration [25]–[28], image super-resolution [29], [30], image enhancement [31], [32], compressive sensing applications [33], [34], computer graphics [35], and others.

## III. THE PROPOSED METHOD

In this section, the proposed method for finding discrete total variation outliers among histograms and the method's underlying model are described. In order to try to avoid any misunderstandings, the structure of this section has purposely been slightly extended. Section III-A gives the general idea of how to use the discrete total variation for outlier detection, Section III-B gives an initial statistical foundation, Sections III-C and III-D use this foundation to derive the relation between the sample size and its expected discrete total variation for two general cases, Section III-E uses this relation to propose the sample models based on the discrete total variation, Section III-F describes the score calculation, Section III-G explains how to combine all these results into a single method, and, finally, Section III-H names this method.

### A. THE GENERAL IDEA

Let there be a sample of  $N$  values,  $\mathbf{x}_n$  its histogram with  $n$  bins, and  $x_i$  the number of values that fell in the  $i$ -th bin with

$$\sum_{i=1}^n x_i = N. \quad (4)$$

Each of the  $n$  bins has an arbitrary interval that can also be unbounded. The bins are not required to be of the same size. Let  $p_i$  be the probability of a value falling in the  $i$ -th bin and

$$\sum_{i=1}^n p_i = 1. \quad (5)$$

Due to randomness the discrete total variation of  $\mathbf{x}_n$ , i.e.  $\|\mathbf{x}_n\|_V$  can differ for each sampling, but it should mostly not differ significantly from its expected value  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  for a given  $N$  and probabilities  $p_i$ . For a given  $\mathbf{x}_n$  the difference between its  $\|\mathbf{x}_n\|_V$  and  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  can serve as a score of how much the sample differs from the expected behavior. Such a score has several drawbacks as well as advantages.

The main disadvantage is that it is required to know  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  for any given  $N$  or at least to know the relation between these two values for proper scaling and comparison.

The main advantage of such a scoring is the simplicity of its calculations due to the very definition of the discrete total variation. Further, because of that it is not necessary to know the desired sample distribution, which significantly widens the application possibilities. Finally, it is not very likely that two samples of the same size have histograms of the same or similar smoothness, i.e. discrete total variation and that their scores differ significantly. That means that 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.

However, in order for this to be practically usable, first an analytical relation between  $N$  and  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  has to be found.

## B. THE STATISTICAL BACKGROUND

The first step in finding a relation between  $N$  and  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  is to examine  $\mathbb{E}[(x_i - x_j)^2]$  in more detail by using the variances of  $x_i$  and  $x_j$ , i.e.  $\text{Var}[x_i]$  and  $\text{Var}[x_j]$ , respectively:

$$\begin{aligned} \mathbb{E}[(x_i - x_j)^2] &= \mathbb{E}[(x_i - \mathbb{E}[x_i] - x_j + \mathbb{E}[x_j] + \mathbb{E}[x_i] - \mathbb{E}[x_j])^2] \\ &= \text{Var}[x_i] - 2\mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\ &\quad + \text{Var}[x_j] + (\mathbb{E}[x_i] - \mathbb{E}[x_j])^2. \end{aligned} \quad (6)$$

The value of  $x_i$  for a given  $i$  has binomial distribution so that

$$\mathbb{E}[x_i] = Np_i, \quad (7)$$

$$\text{Var}[x_i] = Np_i(1 - p_i). \quad (8)$$

For the second term of the last form of Eq. (6) it holds that

$$\mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] = \mathbb{E}[(x_i - \mathbb{E}[x_i])x_j]. \quad (9)$$

The result of Eq. (9) can now be further developed as follows:

$$\begin{aligned} \mathbb{E}[(x_i - \mathbb{E}[x_i])x_j] &= \mathbb{E}[\mathbb{E}[(x_i - \mathbb{E}[x_i])x_j] | x_j] \\ &= \mathbb{E}[(x_i - \mathbb{E}[x_i])\mathbb{E}[x_j | x_i]] \\ &= \mathbb{E}\left[(x_i - \mathbb{E}[x_i]) \frac{p_j}{1 - p_i} (N - x_i)\right] \\ &= -\frac{p_j}{1 - p_i} \mathbb{E}[(x_i - \mathbb{E}[x_i])x_i] \\ &= -\frac{p_j}{1 - p_i} (\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2) \\ &= -\frac{p_j}{1 - p_i} \text{Var}[x_i] = -Np_i p_j. \end{aligned} \quad (10)$$

Combining Eq. (7), Eq. (8), and Eq. (10) develops Eq. (6) to

$$\begin{aligned} \mathbb{E}[(x_i - x_j)^2] &= Np_i(1 - p_i) + 2Np_i p_j + Np_j(1 - p_j) + N^2(p_i - p_j)^2 \\ &= N^2(p_i - p_j)^2 + N(p_i + p_j - (p_i - p_j)^2). \end{aligned} \quad (11)$$

Based on the values of  $p_i$  there are two cases of further actions for establishing a relation between  $N$  and  $\mathbb{E}[\|\mathbf{x}_n\|_V]$ . These two cases are covered in the following subsections.

## C. UNIFORM DISTRIBUTION

### 1) UPPER BOUND

The first case is when the distribution of the sample and consequently the distribution of the histogram are uniform so that the probability of a value falling in the  $i$ -th bin is then

$$p_1 = p_2 = \dots = p_n = \frac{1}{n}. \quad (12)$$

When this is applied to Eq. (11), it eliminates its first term and it simplifies its second term, which then gives the form

$$\mathbb{E}[(x_i - x_j)^2] = \frac{2N}{n}. \quad (13)$$

Taking into account that the square root is a concave function and applying the Jensen's inequality [36] to Eq. (13) gives

$$\mathbb{E}[\sqrt{(x_i - x_j)^2}] = \mathbb{E}[|x_i - x_j|] \leq \sqrt{\mathbb{E}[(x_i - x_j)^2]}. \quad (14)$$

This inequality can then be applied to all neighboring bins:

$$\sum_{i=1}^{n-1} \mathbb{E}[|x_{i+1} - x_i|] \leq \sum_{i=1}^{n-1} \sqrt{\mathbb{E}[(x_{i+1} - x_i)^2]}. \quad (15)$$

Due to the basic properties of the expectation, it holds that

$$\sum_{i=1}^{n-1} \mathbb{E}[|x_{i+1} - x_i|] = \mathbb{E}\left[\sum_{i=1}^{n-1} |x_{i+1} - x_i|\right]. \quad (16)$$

Applying Eq. (3), Eq. (13), and Eq. (16) to Eq. (15) gives

$$\mathbb{E}[\|\mathbf{x}_n\|_V] \leq (n-1) \sqrt{\frac{2N}{n}}. \quad (17)$$

This gives the upper bound for the expected value of the discrete total variation and thus the first relation between  $N$  and  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  if the sample numbers are uniformly distributed.

### 2) EXACT VALUES

Let  $F(n, N)$  denote the expected value of the discrete total variation as a function of two key parameters  $n$  and  $N$ :

$$F(n, N) := \mathbb{E}[\|\mathbf{x}_n\|_V] \quad (18)$$

*Theorem 1:* The exact value of  $F(2, N)$  in closed form is

$$F(2, N) = 2^{-N+1} \lfloor (N+1)/2 \rfloor \binom{N}{\lfloor N/2 \rfloor}. \quad (19)$$

The proof of Theorem 1 is given later in Appendix.

It is relatively easy to show that for each  $r$  it holds that

$$F(2, 2r) = F(2, 2r - 1) \quad (20)$$

and this leads to some unwanted consequences later on in the paper, but there they are mentioned and handled properly.

The case of uniform distribution means that a histogram is a realization of the multinomial distribution and its bins  $x_1, x_2, \dots, x_n$  are random variables. The distribution of each  $x_i$  is  $\mathcal{B}(N, \frac{1}{n})$ , i.e. it is binomially distributed with parameters  $N$  and  $\frac{1}{n}$ . Variables  $x_i$  are not independent, since their sum equals  $N$ . However, because of the symmetry, variables  $x_2 - x_1, \dots, x_n - x_{n-1}$  have the same distribution, which gives

$$\begin{aligned} F(n, N) &= \mathbb{E}[|x_2 - x_1| + \dots + |x_n - x_{n-1}|] \\ &= (n - 1)\mathbb{E}[|x_2 - x_1|]. \end{aligned} \quad (21)$$

Before continuing, for the sake of convenience, first the notation for the multinomial coefficient has to be given as

$$\binom{N}{k_1, \dots, k_n} = \frac{N!}{k_1! \dots k_n!}. \quad (22)$$

**Theorem 2:** The expected value of the total variation of a histogram of uniformly distributed values is calculated as

$$\begin{aligned} F(n, N) &= 2(n - 1) \binom{n - 2}{n}^N \sum_{\substack{k_1 + k_2 \leq N \\ k_1 < k_2}} \\ &\quad \binom{N}{k_1, k_2, N - k_1 - k_2} (n - 2)^{-(k_1 + k_2)} (k_2 - k_1). \end{aligned} \quad (23)$$

The proof of Theorem 2 is given later in Appendix. By using Eq. (23) it is possible to calculate the expected total variation for all reasonable values of  $n$  and  $N$  with some examples being shown in Table 1. However, if using Eq. (23) turns out to be computationally too demanding, the solution is to develop and use some appropriate asymptotic forms.

### 3) ASYMPTOTICS

By taking into account the well-known asymptotic form of the central binomial coefficients that is commonly given as

$$\binom{2r}{r} \approx \frac{4^r}{\sqrt{\pi r}} \quad \text{as } r \rightarrow \infty, \quad (24)$$

it follows that the asymptotic form of  $F(2, N)$  is given as

$$F(2, N) = 2^{-2r+1} r \binom{2r}{r} \approx \sqrt{\frac{2}{\pi}} \sqrt{N}. \quad (25)$$

The experimental calculations suggest that the following hypothesis can be stipulated for the uniform distribution:

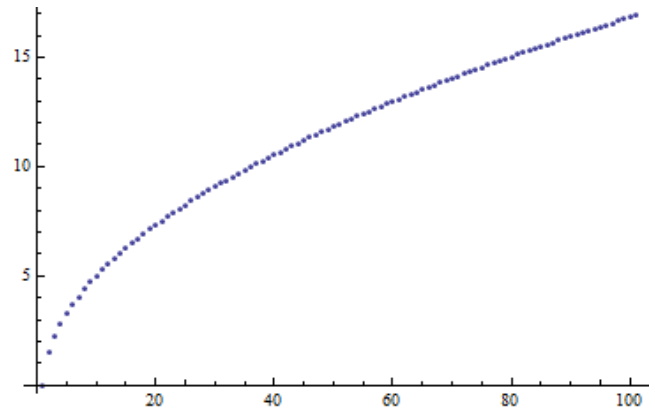
**Hypothesis 1:** For  $N$  sufficiently large, we have

$$F(n, N) \approx (n - 1)F\left(2, \frac{2N}{n}\right). \quad (26)$$

The right side of this equation represents the sum of the discrete total variations of two-binned histograms of the uniform distribution with sample size being equal to the expected

**TABLE 1.** The comparison of the exact values of  $F(n, N)$  with the values obtained by Eq. (27) for some  $n$  and  $N$ .

$n$	$N = 100$		$N = 1000$	
	Eq. (27)	exact value	Eq. (27)	exact value
2	7.97885	7.95892	25.2313	25.2250
3	13.0294	13.0213	41.2026	41.2000
4	16.9257	16.9045	53.5237	53.5170
5	20.1851	20.1472	63.8308	63.8188
6	23.0329	22.9752	72.8366	72.8183
7	25.5892	25.5090	80.9203	80.8950
8	27.9260	27.8207	88.3096	88.2765
9	30.0901	29.9577	95.1533	95.1116
10	32.1142	31.9525	101.554	101.503
20	47.9395	47.3907	151.598	151.427
30	59.7437	58.6681	188.926	188.595
40	69.5808	67.8604	220.034	219.509
50	78.1927	75.7182	247.267	246.522



**FIGURE 1.** The values of  $F(4, N)$  for  $1 \leq N \leq 100$ .

number of values. If this hypothesis is accepted, then the following asymptotic is true for the uniform distribution:

$$F(n, N) \approx \frac{2(n - 1)}{\sqrt{n\pi}} \sqrt{N}. \quad (27)$$

In Table 1 the values obtained by Eq. (27) are compared to the exact values of  $F(n, N)$  for some chosen  $n$  and  $N$ .

Hypothesis 1 and the results of the numerical calculation furthermore suggest that the following hypothesis is true:

**Hypothesis 2:** For each  $n \geq 3$ , the function  $N \mapsto F(n, N)$  is increasing and strictly concave, hence, for each  $0 \leq k \leq N$

$$F(n, k) + F(n, N - k) < 2F(n, \frac{N}{2}). \quad (28)$$

Function  $N \mapsto F(2, N)$  is nondecreasing, but it is not strictly concave, because as demonstrated by Eq. (20) its neighboring values can be equal. The proof of these two hypotheses may be very difficult, but they are not essential for the conclusions that are drawn later in the paper. The diagram in Fig. 1 shows the situation for  $n = 4$  and  $1 \leq N \leq 100$ .

Let  $F_c(n, N)$  denote the expected value of the **circular variation**, which unlike the usual variation has an additional



term  $|x_1 - x_n|$  for the absolute value of the difference between the first and the last bin.  $F_c(n, N)$  is then defined as

$$F_c(n, N) = \mathbb{E}[|x_2 - x_1| + \dots + |x_n - x_{n-1}| + |x_1 - x_n|]. \quad (29)$$

By taking into account Eq. (21), it follows from Eq. (29) that

$$F_c(n, N) = \frac{n}{n-1} F(n, N). \quad (30)$$

Applying Eq. (21) and adjusting the result for later use gives

$$\begin{aligned} F_c(n, N) &= \frac{n}{n-1} (n-1) \mathbb{E}[|x_2 - x_1|] \\ &= n \mathbb{E}[|x_2 - x_1|] \\ &= \frac{n}{2} (\mathbb{E}[|x_2 - x_1|] + \mathbb{E}[|x_n - x_{n-1}|]). \end{aligned} \quad (31)$$

All possible histograms  $\mathbf{x}_n$  can be split into disjoint groups, according to the number of realizations which fall into the first  $n/2$  bins. Let  $q_k$  be the probability that these bins contain exactly  $k$  realizations. Because of the symmetry,  $q_k = q_{N-k}$  for each  $k$ . Since other  $n/2$  bins contain exactly  $N - k$  realizations, the conditional distribution of the realizations in the first  $n/2$  bins is again uniform. Having all this in mind and applying the partition theorem to Eq. (31) gives

$$\begin{aligned} F_c(n, N) &= \frac{n}{2} \sum_{k=0}^N q_k (\mathbb{E}[|x_2 - x_1| | k] + \mathbb{E}[|x_n - x_{n-1}| | N - k]) \\ &= \sum_{k=0}^N q_k \left[ F_c\left(\frac{n}{2}, k\right) + F_c\left(\frac{n}{2}, N - k\right) \right]. \end{aligned} \quad (32)$$

Applying Eq. (28) and the equality  $(\sum_{k=0}^N q_k) = 1$  leads to the following inequality that holds for each even  $n > 4$ :

$$F_c(n, N) < 2F_c\left(\frac{n}{2}, \frac{N}{2}\right). \quad (33)$$

Here  $n$  has to be greater than 4 because having  $n = 4$  effectively leads to use of the function  $F(2, N)$  on the right side of the inequality, and as explained earlier, this is inappropriate for Eq. (28). If  $n = k2^r$  where  $k \geq 3$  and  $r \geq 0$  are integers, then taking the inequality above recursively leads further to

$$F_c(k2^r, N) < 2^r F_c\left(k, \frac{N}{2^r}\right) \quad (34)$$

wherefrom for all suitable  $N$  and  $n$  it then follows that

$$F_c(n, N) < \frac{n}{k} F_c\left(k, \frac{kN}{n}\right). \quad (35)$$

If  $k = 2$  is taken, then the inequality is no longer necessarily valid because of the involvement of  $F(2, N)$ . However, the obtained form yields a better approximation of  $F_c(n, N)$  as

$$F_c(n, N) \approx \frac{n}{2} F_c\left(2, \frac{2N}{n}\right)$$

wherefrom after applying Eq. (30) it then further follows that

$$F(n, N) \approx (n-1) F\left(2, \frac{2N}{n}\right),$$

which in turn is an approximation stipulated in Hypothesis 1.

#### 4) APPROXIMATION ERROR

Fig. 2 shows the difference between the results of Eq. (23) and Eq. (27), which represent the exact and approximated values of  $F(n, N)$ , respectively. It can be seen that in cases where  $N$  is several times greater than  $n$ , the approximation error becomes relatively insignificant for practical purposes. The error only becomes significant when the value of  $N$  is relatively close to the value of  $n$  or below it, but it must be additionally stressed that this rarely occurs in practice since having such values of  $n$  and  $N$  is not too useful. The plots in Fig. 3 further suggests that if required, the approximation error could be modelled accurately. However, for the later use here it is enough to conclude that having a sufficiently large value of  $N$  renders the approximation error insignificant.

#### D. NON-UNIFORM DISTRIBUTION

The second case is when the distribution of the sample and consequently the distribution of the histogram are not uniform. In other words this is the case where Eq. (12) does not hold, i.e. when  $p_i \neq p_j$  for at least one pair of  $i$  and  $j$ . Applying to Eq. (11) all steps that have led to Eq. (17) gives

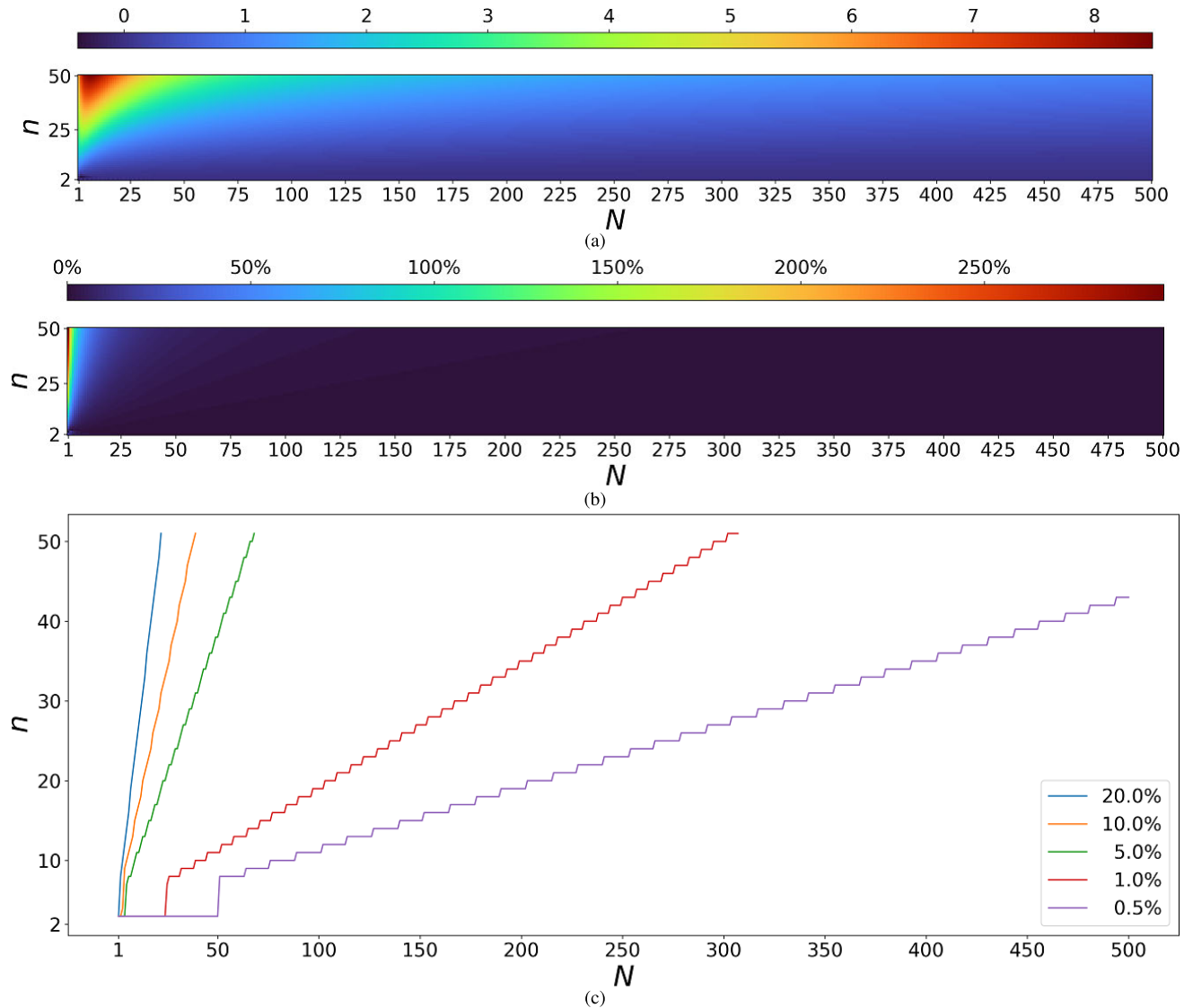
$$\begin{aligned} &\sum_{i=1}^{n-1} \mathbb{E}[|x_i - x_j|] \\ &\leq \sum_{i=1}^{n-1} \sqrt{N^2 (p_{i+1} - p_i)^2 + N (p_{i+1} + p_i - (p_{i+1} - p_i)^2)} \\ &\leq \sum_{i=1}^{n-1} \left( \sqrt{N^2 (p_{i+1} - p_i)^2} \right. \\ &\quad \left. + \sqrt{N (p_{i+1} + p_i - (p_{i+1} - p_i)^2)} \right) \\ &= \sum_{i=1}^{n-1} \left( N |p_{i+1} - p_i| + \sqrt{N (p_{i+1} + p_i - (p_{i+1} - p_i)^2)} \right) \\ &= N \sum_{i=1}^{n-1} |p_{i+1} - p_i| + \sqrt{N} \sum_{i=1}^{n-1} \sqrt{(p_{i+1} + p_i - (p_{i+1} - p_i)^2)}. \end{aligned} \quad (36)$$

If  $\mathcal{D}$  is the sample's theoretical distribution, then the first term of Eq. (36) is the discrete total variation of  $\mathcal{D}$  that is given as

$$\|\mathcal{D}\|_V = \sum_{i=1}^{n-1} |p_{i+1} - p_i|. \quad (37)$$

The second term is a bound for expectation of the deviation of this given sample from its theoretical distribution. A rough estimate for this second term is the value  $2\sqrt{n-1}\sqrt{N}$ . It is obtained by first removing the subtracting part and applying the inequality  $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$  for  $u, v > 0$ , which gives

$$\begin{aligned} &\sum_{i=1}^{n-1} \sqrt{(p_{i+1} + p_i - (p_{i+1} - p_i)^2)} \\ &\leq \sum_{i=1}^{n-1} \sqrt{(p_{i+1} + p_i)} \leq \sum_{i=1}^{n-1} \sqrt{p_i} + \sum_{i=1}^{n-1} \sqrt{p_{i+1}}. \end{aligned} \quad (38)$$



**FIGURE 2.** The difference between the results of Eq. (23) and Eq. (27), which represent the exact and approximated values of  $F(n, N)$ , respectively: a) the absolute error, b) the relative error, and c) the dependance of certain relative errors on  $n$  and  $N$ .

Since  $\sqrt{p_i}$  and  $\sqrt{p_{i+1}}$  are non-negative, the sums in Eq. (38) can effectively be seen as  $L_1$ -norms of  $(n-1)$ -dimensional vectors. Applying the inequality  $\|\mathbf{v}\|_1 \leq \sqrt{d} \|\mathbf{v}\|_2$  where  $d$  is the dimension of the vector  $\mathbf{v}$  [37] to these sums gives

$$\begin{aligned} & \sum_{i=1}^{n-1} \sqrt{p_i} + \sum_{i=1}^{n-1} \sqrt{p_{i+1}} \\ & \leq \sqrt{n-1} \left[ \left( \sum_{i=1}^{n-1} p_i \right)^{1/2} + \left( \sum_{i=i}^{n-1} p_{i+1} \right)^{1/2} \right] \\ & \leq \sqrt{n-1} (1+1) = 2\sqrt{n-1}. \end{aligned} \quad (39)$$

It is useful to know the discrete total variation of some important distributions. Examples of their histograms are shown in Fig. 4. The uniform distribution has a zero total variation. For the triangular distribution  $\mathcal{T}$  with  $n$  bins this

is

$$\|\mathcal{T}\|_V = \frac{4n-8}{n^2} \approx \frac{4}{n+2} \quad (40)$$

for an even  $n$ , while in the case of an odd  $n$  this is given as

$$\|\mathcal{T}\|_V = \frac{4n-6}{n^2} \approx \frac{4}{n+2}. \quad (41)$$

The square distribution  $\mathcal{Q}$  for which  $p_i = Ci^2$  with  $n$  bins has a discrete total variation that can be approximated as

$$\|\mathcal{Q}\|_V \approx \frac{3}{n}. \quad (42)$$

Next, in the case of the square root distribution  $\mathcal{S}$  for which  $p_i = C\sqrt{i}$  and with  $n$  bins the approximation is given as

$$\|\mathcal{S}\|_V \approx \frac{3}{2n}. \quad (43)$$

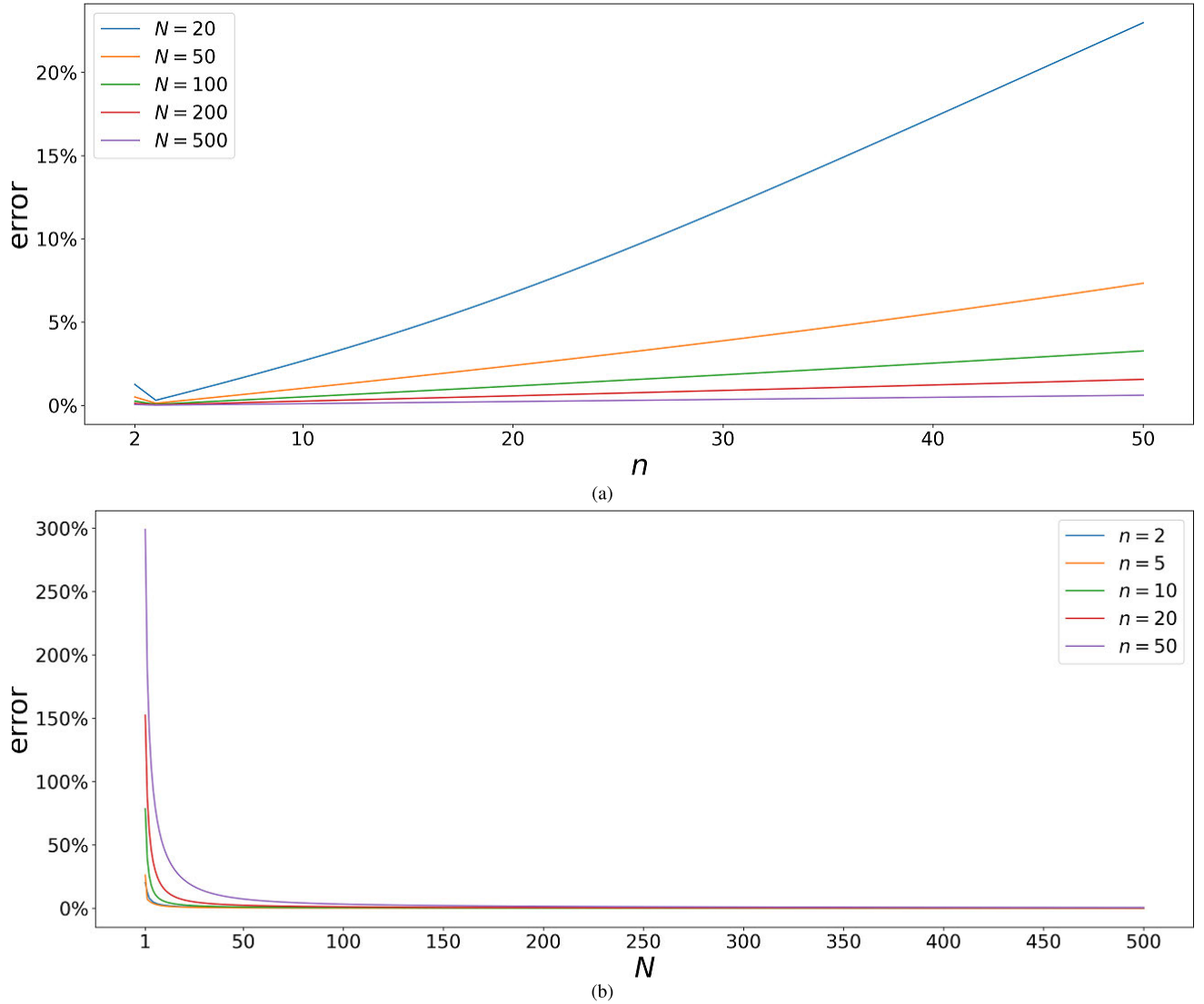


FIGURE 3. The relation between the error when using Eq. (27) and the values of: a) sample size  $N$  and b) number of bins  $n$ .

For the geometric distribution  $\mathcal{G}$  with parameter  $p$  this is

$$\|\mathcal{G}\|_V = p, \quad (44)$$

for the Poisson distribution  $\mathcal{P}$  with parameter  $\lambda > 1$  it is

$$\|\mathcal{P}\|_V \approx \frac{2\lambda^{\lfloor \lambda \rfloor} e^{-\lambda}}{\lfloor \lambda \rfloor!}. \quad (45)$$

The discrete total variation for a unimodal discrete distribution with mode  $M$  is bounded by  $2M$ . The mode for symmetric binomial distribution  $\mathcal{B}(n, \frac{1}{2})$  is  $\frac{1}{2^n} \binom{n}{\lfloor n/2 \rfloor}$  and

$$\|\mathcal{B}\|_V \approx \sqrt{\frac{8}{\pi n}}. \quad (46)$$

The normal distribution  $\mathcal{N}(0, \sigma^2)$  is a continuous one with unbounded support and its theoretical DTV depends on rasterization. The total variation of the probability density function is  $\frac{2}{\sigma\sqrt{2\pi}}$ . If  $[-c, c]$  is essentially the support of the

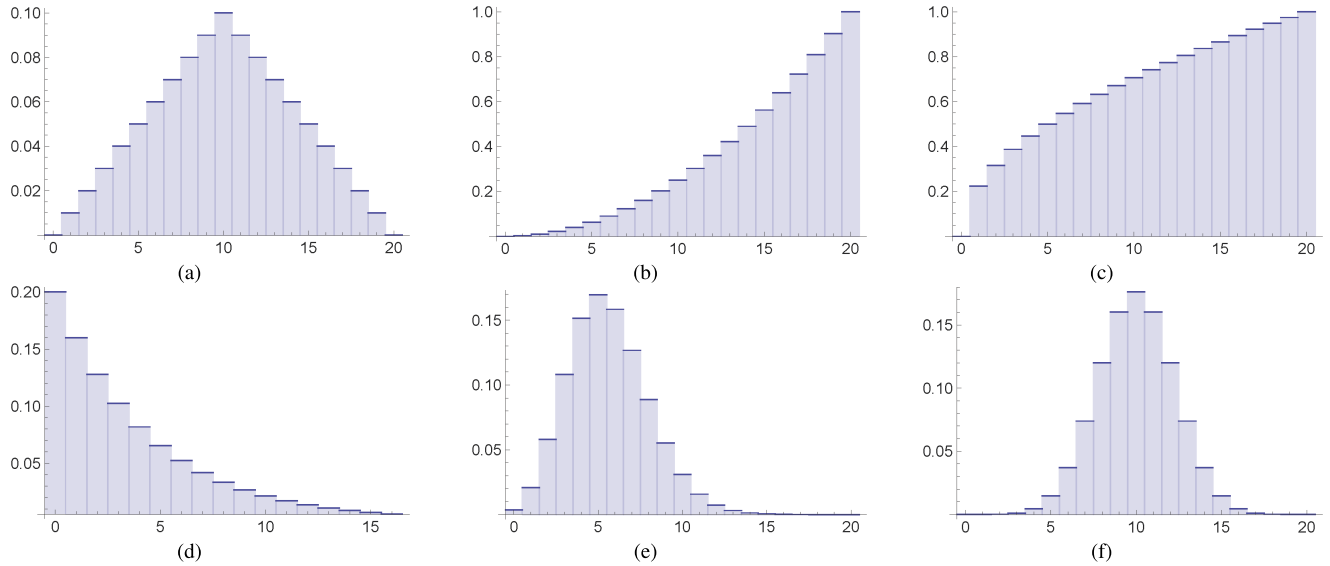
distribution and if  $n \geq \frac{2c}{\sigma}$ , then  $\|\mathcal{N}\|_V$  can be approximated:

$$\|\mathcal{N}\|_V \approx \frac{2c}{n\sigma} \sqrt{\frac{2}{\pi}}. \quad (47)$$

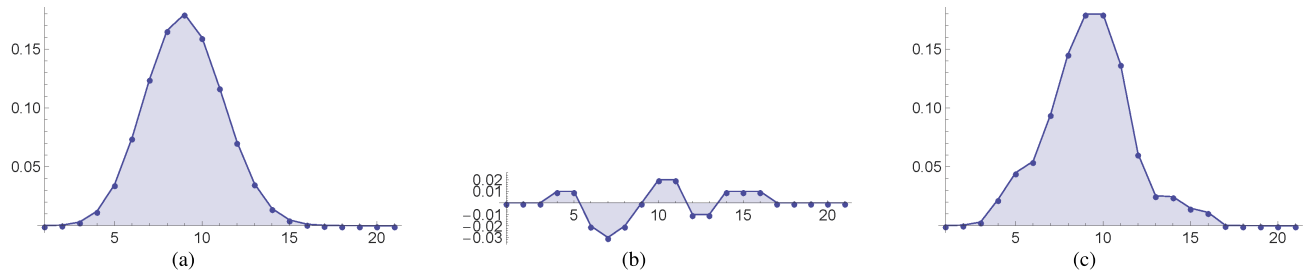
Let  $\mathcal{D}$  be any distribution and  $\mathbf{x}_n$  the histogram with  $n$  bins of a corresponding sample of  $N$  values drawn from the distribution  $\mathcal{D}$ . Then similarly to Eq. (36) it can be written

$$\mathbb{E}[\|\mathbf{x}_n\|_V] \leq \|\mathcal{D}\|_V \cdot N + \mathbb{E}[\|\mathcal{R}\|_V] \sqrt{N} \quad (48)$$

where  $\mathcal{R}$  is a deviation from the theoretical distribution. If there was no randomness and all values were distributed exactly as predicted by the probabilities, then  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  would be  $\|\mathcal{D}\|_V \cdot N$ . Therefore, the second term is due to the randomness. A further thing to notice here is that as  $N$  grows, randomness plays an ever smaller role in Eq. (48) and as  $N$  limits at infinity, the term  $C_1 N$  gets to fully dominate in Eq. (48), which is also expected in accordance with the



**FIGURE 4.** Histograms for the a) triangular, b) quadratic, c) square root, d) geometric, e) Poisson, and f) binomial distribution. The shown histograms are merely for the sake of illustration and the  $x$ -axes do not strictly follow the equations in Section III-D.



**FIGURE 5.** From a) the theoretical distribution by adding b) the deviation due to randomness to c) the final sample distribution.

Glivenko-Cantelli theorem. In Fig. 5 the total variation of the theoretical distribution and the total variation of a sample are equal. This will be the case for all samples which do not alter order between adjacent bins. Therefore, the alteration from the theoretical distribution means that the corresponding sample is essentially different from theoretical one. Deviation from the theoretical distribution can be approximated as total variation of a sample from uniform distribution and therefore the bounds written before can be applied to any distribution.

With regard to the distribution, the use of the discrete total variation that is somewhat similar to the  $L_1$ -norm may be reminiscent of the assumption of the Laplace distribution. However, no minimization, regularization, or any similar process that requires such an assumption is being performed here. Therefore, it should be stressed again that the relations obtained here can be applied to samples of any distribution.

#### E. THE PROPOSED MODEL

After taking into account the previous subsections' results, it is reasonable to consider the model for  $\mathbb{E}[\|\mathbf{x}_n\|_V]$  to be

$$m = aN + b\sqrt{N}. \quad (49)$$

This model can be fitted directly to the sizes and discrete total variations obtained on the given histograms that are to be checked for outliers. If there is not enough given histograms to cover the desired value ranges of  $N$ , then additional ones can be created by randomly subsampling the given ones. In the case where a larger amount of histogram outliers is suspected, then their detrimental influence on fitting of Eq. (49) can be reduced by applying methods such as RANSAC [38].

Alternatively, if the distribution, i.e. the values of  $p_i$  for the histograms' bins are known, then  $a$  and  $b$  can be obtained through Monte Carlo simulation by randomly creating arbitrarily many histograms of various sizes  $N$  and then fitting the model Eq. (49) to their sizes and discrete total variations.

#### F. SCORE CALCULATION

Once the model described by Eq. (49) has been fitted to data, the next step is to assign an outlier score to each of the given histograms. The first step is to calculate a histogram's discrete total variation. Next, the discrete total variation expected for the histograms's size is obtained by using Eq. (49). Finally, the absolute difference between these two values is

$$d = \left| \|\mathbf{x}_n\|_V - m \right|. \quad (50)$$

However,  $d$  cannot yet be used as the score because the standard deviation of the discrete total variation for histograms of random samples varies depending on the samples size  $N$ , which means that the significance of  $d$  depends on  $N$ . This means that first the influence of the sample size on the standard deviation has to be removed. Additionally, the discrete total variation is already a statistic of the sample, which means that its standard deviation is actually the standard error [39]. Many standard errors that do not include division by  $N$  are proportional or close to being proportional to  $\sqrt{N}$ , at least in limit, and in practice this is also the case with the discrete total variation. This can intuitively be seen in the form of the second term of Eq. (48) as discussed earlier. Therefore, for practical purposes the influence of  $N$  on  $d$  can be approximately removed by calculating the distance  $d'$  as

$$d' = \frac{d}{\sqrt{N}} = \frac{\|\mathbf{x}_n\|_V - m}{\sqrt{N}} = \frac{\|\mathbf{x}_n\|_V - aN + b\sqrt{N}}{\sqrt{N}}. \quad (51)$$

The value of  $d'$  can now be used instead of the value  $d$  as the outlier score for the histogram that it was calculated for because it is normalized with respect to the standard error.

It must be mentioned that strictly speaking Eq. (51) is theoretically not correct because the expected value of the discrete total variation is not always proportional to  $N$ . However, during the research conducted for this paper it has been empirically shown that for all tested distributions the standard error was proportional to  $\sqrt{N}$  and that using Eq. (51) is a good practice, even though it may introduce inaccuracies. Since Eq. (51) was specifically designed to comply with the statistical properties related to the discrete total variation as discussed here, using some other score calculation would potentially require a major overhaul of the whole framework.

An alternative to using Eq. (51) that unlike Eq. (51) does not include an explicitly derived formula is to take all data from the given histograms, use it in Monte Carlo simulations to create samples of various desired sizes, for each of these sizes calculate the discrete total variations and their standard deviation, and fit a model to these sizes and their respective standard deviations. If enough data is available, this should result in a relation that is very similar to the one in Eq. (51).

Since  $d'$  is the normalized distance from the expected discrete total variation and since it resembles the  $t$ -statistic, it could be further used to also provide a probabilistic interpretation for a given histogram. However, the goal of this paper is not to propose a new statistical test that can be used in hypothesis testing with predetermined significance levels. The main goal of this paper is just to find the most likely outlier candidates based on the discrete total variation and the distance  $d'$  also suffices for such ranking. Therefore, probabilistic interpretation calculation is omitted in this paper.

## G. APPLICATION

With all the required background given in the previous subsections, it is possible to propose a new method for detecting histogram outliers in terms of the discrete total variation.

## Algorithm 1 The Proposed Method TVOR

**Input:**  $M$  input histograms  $\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)}, \dots, \mathbf{x}_n^{(M)}$

**Output:** scores for input histogram  $d'_1, d'_2, \dots, d'_M$

```

1: for  $i \in \{1, 2, \dots, M\}$  do
2:    $s_i = \sum_{j=1}^n x_j^{(i)}$  ▷ Calculate sample size
3:    $v_i = \|\mathbf{x}_n^{(i)}\|_V$  ▷ Calculate discrete total variation
4: end for
5:  $a, b = \text{FitModel}\left(\bigcup_{i=1}^M (s_i, v_i)\right)$  ▷ Fit Eq. (49) to data
6: for  $i \in \{1, 2, \dots, M\}$  do
7:    $d'_i = \frac{v_i - as_i + b\sqrt{s_i}}{\sqrt{s_i}}$  ▷ Calculate the score
8: end for

```

First, multiple histograms for the samples of various sizes are given as input. The histograms are supposed to have the same bins where each of the bins can have an arbitrary interval. It is also supposed that all these samples are drawn from the same distribution and the goal is to check which of them are most likely to be outliers in terms of the discrete total variation. Next, the discrete total variation is calculated for each of these histograms. Then, model Eq. (49) is fitted to histogram sizes and discrete total variations. Finally, each of the histograms is scored by applying Eq. (51). The histograms for which the highest score values were obtained are the most likely outlier candidates in terms of their discrete total variation. All these steps are summarized in Algorithm 1.

Here it should be additionally stressed that the proposed method has no hyperparameters whatsoever that would have to be tuned or that would otherwise influence the result. It may seem that the number of histogram bins  $n$  is a tunable hyperparameter, but the proposed method is agnostic of the underlying histogram samples - it merely receives already existing histograms as inputs. The histograms are only assumed to have the same bins. It is not even important what the range of the bins is nor is it important whether they are bounded.

## H. THE PROPOSED METHOD'S NAME

Due to the proposed method's model's reliance on the discrete total variation, it was named Total Variation Outlier Recognizer (TVOR) or for the sake of simplicity just Tvor, which is pronounced /tvô:r/ and it means *skunk* in Croatian.

## IV. EXPERIMENTAL RESULTS

In order to validate the proposed method, several experiments have been conducted on both synthetic and real-life data. Additionally, it is shown why the proposed method is more appropriate than some other similar methods. To give a clear and descriptive overview of the method's properties, the structure of this section is purposely slightly more extended. First, Section IV-A describes a baseline method for histogram outlier detection based on the Pearson's chi-squared test [10] to compare its results to the ones of the proposed method. In Section IV-B the behavior of the proposed method in



several scenarios of changing conditions is demonstrated and additionally explained by several experiments for distribution outlier detection among histograms of random samples of different sizes drawn from the normal distribution and the beta distribution with various parameter values. Similar to that, Section IV-C contains experiments for discrete total variation outlier detection among histograms of random samples of various sizes drawn from the beta distribution. The real-life practical use of the proposed method is demonstrated in Section IV-D on the histograms of the birth years taken from census data of several populations from the same time frame. Section IV-E shows the advantage of the proposed method over some other methods that can be used for similar purposes. The obtained results are discussed in Section IV-F. The online repository with the source code and the data required to recreate the results is described in Section IV-G.

### A. THE BASELINE METHOD

The proposed method's goal is to detect outliers specifically in terms of the expected discrete total variation, which can differ significantly from detecting distribution outliers in general. Therefore, the goal of this section is to show the difference in the performance of the proposed method and the Pearson's chi-squared test [10]. This test can be used to check whether a histogram is an outlier by comparing the values of the histogram's bins, which serve here as the categorical variables, to the values that are expected under a supposed distribution. However, since in the problem that is being analyzed in this paper the supposed distribution is unknown, the expected bin values first have to be estimated.

The first step in calculating the  $i$ -th expected bin value is to sum the values of the  $i$ -th bin in all given histograms except the tested one. When this is done for all  $n$  bins, all of the obtained bin sums are divided by the sum of values of all bins in all histograms except the tested one. These normalized sums now represent the estimations of the probabilities that a value will fall in each of the histogram bins. The more histogram are given, the better these estimations are under the Glivenko-Cantelli theorem. Next, all these estimated probabilities are then multiplied by the sum of all bin values in the tested histogram. In that way the sum of the bins in the tested histogram and the sum of the estimated expected bin values are the same. Then, a small positive number is added to all scaled bin values in order to avoid division by zero during the calculation of the Pearson's chi-squared test statistics. Finally, the obtained Pearson's chi-squared test statistic is used as the outlier score for the tested histogram. The described procedure is summarized in Algorithm 2.

### B. SYNTHETIC DATA FOR DISTRIBUTION OUTLIERS

#### 1) THE GOAL

Since there is much freedom in the overall data generation procedure when using synthetic data and less or no limitations when compared to using real-life data, the goal of this subsection is to demonstrate and explain in more detail the behavior

#### Algorithm 2 The Baseline Method

---

**Input:**  $M$  input histograms  $\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)}, \dots, \mathbf{x}_n^{(M)}$   
**Output:** scores for input histogram  $\chi_1^2, \chi_2^2, \dots, \chi_M^2$

- 1: **for**  $i \in \{1, 2, \dots, M\}$  **do**
- 2:    $s_i = \sum_{j=1}^M x_j^{(i)}$  ▷ Calculate sample size
- 3: **end for**
- 4:  $S = \sum_{i=1}^M s_i$  ▷ Calculate the sum of all bins
- 5: **for**  $i \in \{1, 2, \dots, n\}$  **do**
- 6:    $b_i = \sum_{j=1}^M x_i^{(j)}$  ▷ Calculate individual bin sum
- 7: **end for**
- 8:  $\epsilon = 10^{-6}$  ▷ A small positive number
- 9: **for**  $i \in \{1, 2, \dots, M\}$  **do**
- 10:   **for**  $j \in \{1, 2, \dots, M\}$  **do**
- 11:      $O_j^{(i)} = x_j^{(i)}$  ▷ The observed bin value
- 12:      $E_j^{(i)} = \frac{b_j - x_j^{(i)}}{S - s_i} s_i + \epsilon$  ▷ The expected bin value
- 13:   **end for**
- 14:    $\chi_i^2 = \sum_{j=1}^n \frac{(O_j^{(i)} - E_j^{(i)})^2}{E_j^{(i)}}$  ▷ Calculate the score
- 15: **end for**

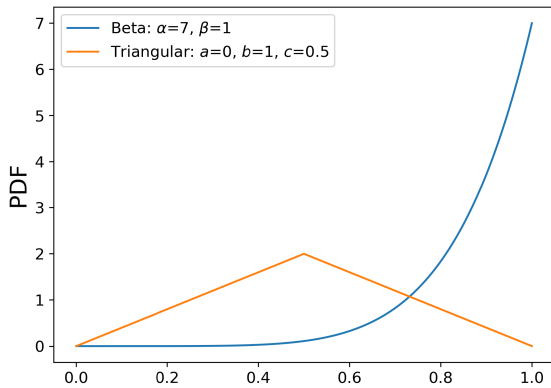
---

of the proposed method depending on gradual changes of various conditions. The performance is here first measured in terms of distribution outlier detection, even though the proposed method was not designed specifically for that task, while the performance in terms of DTV outlier detection is described in the following subsection. The experiments were performed for cases when the inlier and outlier samples for histograms were from the same distribution with changed parameter values and from different distributions.

#### 2) EXPERIMENTAL SETUP

The experiments for distribution outlier detection on synthetic data, i.e. histograms of random samples, were conducted by repeatedly first simulating the mixtures of inlier and outlier samples, then trying to recognize the outlier samples by means of applying the baseline method and the proposed method, and finally examining the results of these simulations. The experiments were conducted for two general cases of inlier and outlier random sample distributions by mixing them in  $10^4$  simulations. In the first case both the inlier and outlier samples were from the normal distribution.

In each simulation of this first case, the inlier data was prepared by generating 100 random inlier samples drawn from the normal distribution with mean 0 and variance 1, i.e.  $\mathcal{N}(0, 1)$ . The size of each individual sample was randomly chosen to be between 500 and 1000. The histogram bins were set to be evenly spaced on the interval  $[-c, c]$  where  $c$  is an arbitrarily chosen value used to check the behavior of various bin arrangements. Each sample value falling outside of the interval  $[-c, c]$  was replaced with the closer one of  $c$  and  $-c$ . Several values of  $c$ , as well as several values of number of bins  $c$ , were used to check the effect of changing conditions.



**FIGURE 6.** The probability density functions of the beta distribution and triangular distribution used in the experiments.

Furthermore, in each simulation, the outlier data was generated by drawing a certain number of random samples from  $\mathcal{N}(0, \sigma^2)$  for various  $\sigma \neq 1$ . The sample size was randomly determined in the same way as for the inlier samples. For both the inlier and outlier data the values of  $c$  and  $n$  were set to the same values to assure having histograms with the same bins. Next, the baseline method and the proposed method were applied to the combined inlier and outlier data to score individual histograms. Finally, the mean value of the rank of all outlier examples obtained by each method was calculated as the performance score of each method. A lower mean rank here means a better performance in terms of outlier detection. For the sake of simplicity, zero-based numbering was used for ranks. This means that in the case of a single added outlier sample, the optimal mean rank of a tested method is 0, while in the case of e.g. 10 added outlier samples, the optimal mean rank is 4.5 since this is the average value of the first 10 zero-based ranks, which should all be assigned to outlier samples' histograms in the case of a method that performs ideally.

In short, every instance of the simulation setup is uniquely determined by the number of histogram bins  $n$ , the number of added outlier samples, the value  $c$  used to determine the interval of the binned values, and the value of  $\sigma$  for outlier distribution. Simulations for each instance were repeated  $10^4$  times to check the performance of the baseline method and the proposed method in various sampling conditions.

In the second general case, the inlier samples were drawn from the beta distribution with parameter values  $\alpha = 7$  and  $\beta = 1$ , while the outlier samples were drawn from the triangular distribution with parameter values  $a = 0$ ,  $b = 1$ , and  $c = 0.5$ . The probability density functions of these distributions are shown in Fig. 6. Similarly to the previous case, several combinations of the number of bins  $n$  and the number of outlier sample histograms added to the 100 inlier sample were checked. For each combination, the results of methods' performance were averaged over  $10^4$  simulations.

### 3) RESULTS

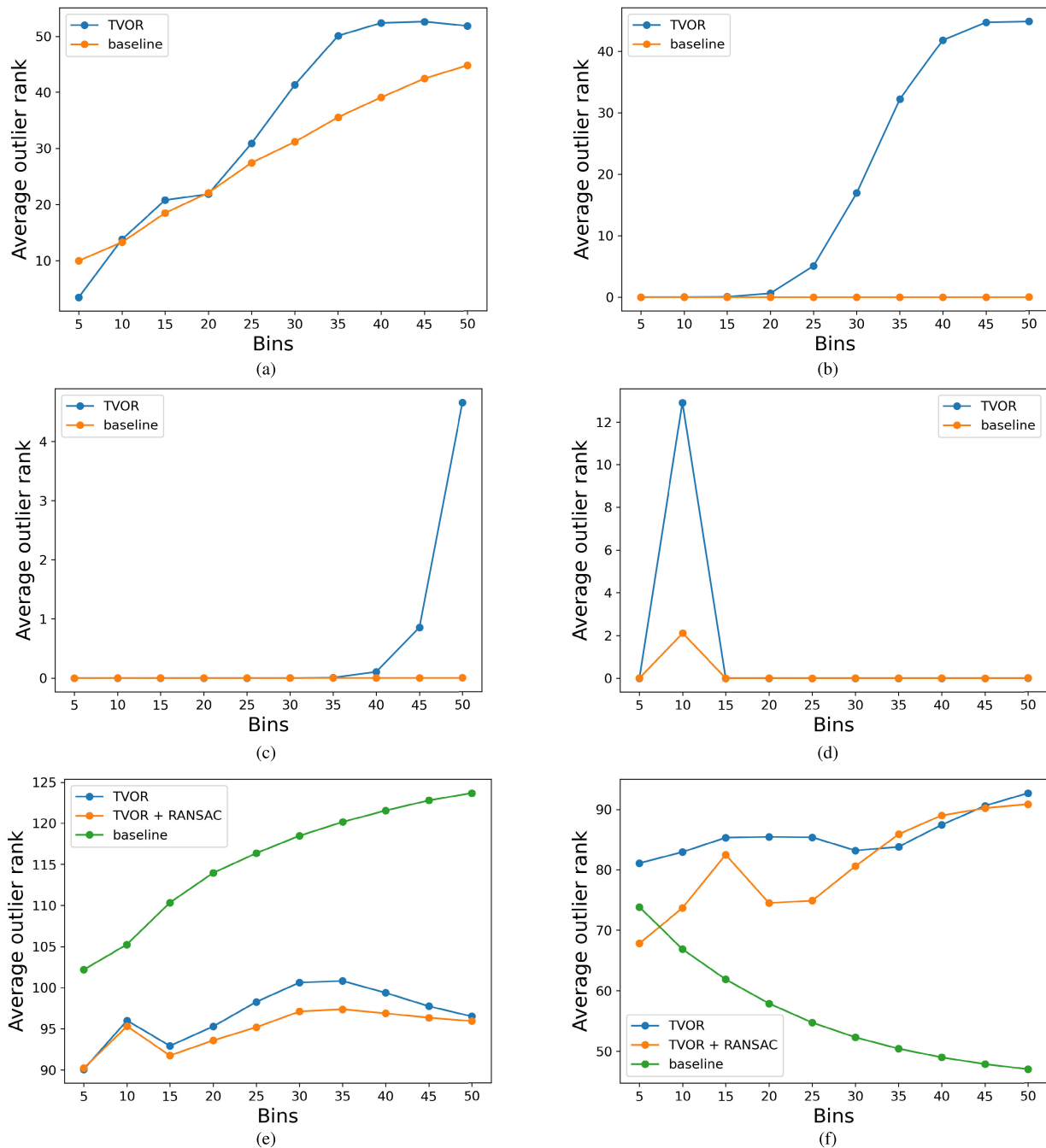
After examining the results of performing simulations for a large number of setups when both the inliers and the outliers

are from the normal distribution, due to the similarity of many of the results, it was decided to show only those that can be used to summarize them all. These results are shown in Fig. 7. The first thing to observe is that in the majority of the cases the baseline method based on the Pearson's chi-squared test performs better in terms of outlier ranking. This is mainly because the proposed method was not designed to find outliers in general, but to find outliers in terms of the discrete total variation. Interestingly, however, the exception to this are the cases when there is a relatively small number of bins, which can be seen in Figs. 7a and 7f, and cases with a high amount of added outlier sample histograms, which can be seen in Fig. 7e where the proposed methods outperforms the baseline method for all given numbers of bins. This means that even if the proposed method was not designed for the same task as the baseline method, in some cases it is still able to outperform it, which may be useful should such cases emerge. A more detailed analysis of the performance results shown in Fig. 7 is given in Appendix, which also explains the sudden drops in the performance such as the one in Fig. 7d.

In short, the proposed method generally performs worse than the baseline method. However, in the cases of smaller values of  $n$ , i.e. in the cases of a smaller number of bins, as well as in the cases with a high amount of outliers, it may perform better. Similar results can be obtained with some other distributions as well and therefore they have been omitted here. If required, any other experiments with a similar setup can be conducted by using the source code publicly available in the repository that is described later.

Next, Fig. 8 shows the results of the experiment where the inlier and the outlier samples were drawn from the beta and the triangular distribution, respectively. As can be expected by viewing Fig. 6, the baseline method outperforms the proposed method in most cases since the difference between the used distributions is significant. Nevertheless, Fig. 8b again shows that the proposed method may be able to outperform the baseline method in the case of a high amount of outliers.

The performance drop of the proposed method for several values of  $n$  shown in Fig. 8a deserves some additional comments. As shown in Fig. 9a, the theoretical DTVs of both distributions are clearly separated for all shown values of  $n$ . This means that if the random samples were sufficiently big, then the performance should significantly improve in accordance with Eq. (48). Namely, in that case the influence of the sample size significantly overpowers the influence of the randomness. As a matter of fact, if the whole experiment is repeated with random samples having their sizes increased by several orders of magnitude, then both the proposed method and the baseline method have the same ideal performance. However, as mentioned earlier, the size of each sample used in the experiment whose results are shown in Fig. 8a was randomly chosen to be between 500 and 1000. For such sizes, the randomness still has a substantial influence on the histograms' DTVs. This is illustrated in Fig. 9b, which shows the mean DTV calculated for  $10^6$  random samples of size 1000 for various values of  $n$  created for both the beta and the

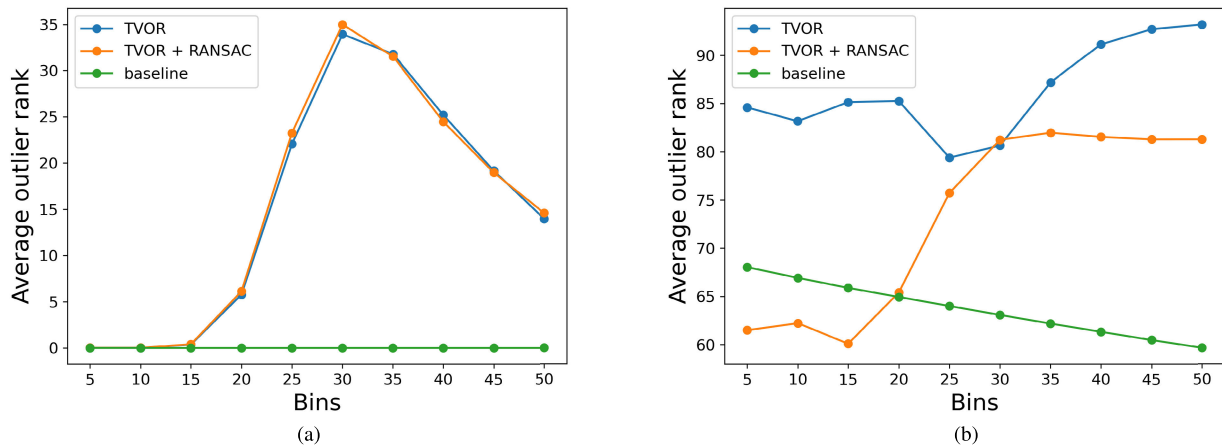


**FIGURE 7.** Comparing the performance of the proposed and baseline methods. First row: performance with 1 added outlier and  $c = 5$  for a)  $\sigma = 0.9$  and b)  $\sigma = 1.5$ . Second row: Performance with 1 added outlier and  $\sigma = 0.5$  for c)  $c = 5$  and d)  $c = 10$ . Third row: Performance with 90 added outliers and  $c = 5$  for e)  $\sigma = 0.9$  and f)  $\sigma = 1.5$ . The results for TVOR + RANSAC was added only in the third row because for the results in the first and the second row the difference was not that significant.

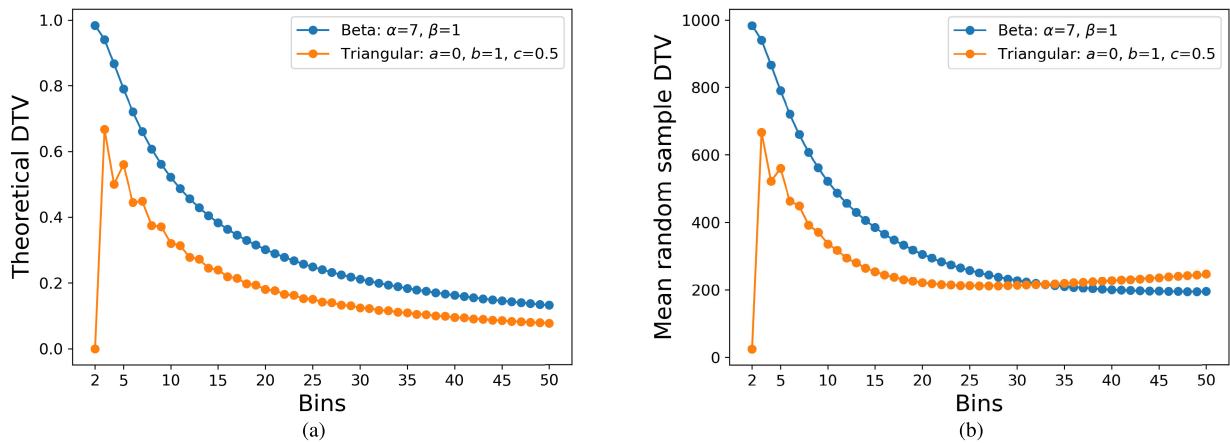
triangular distribution. It can be clearly seen how this differs from the case of the theoretical DTVs and this can be used to explain the particularly low performance of the proposed method when  $n$  is 30 and 35 shown in Fig. 8a. Namely, for these values of  $n$ , the mean values of DTVs become so close that, with the influence of randomness included, it becomes difficult to successfully distinguish between the inlier and the outlier histograms based only on their DTVs. The dependence

of the proposed method's performance on the size of the samples is further analyzed in more detail in Appendix. Based on all the results shown here and in Appendix, it can be concluded that the proposed method's performance improves as the size of the samples increases.

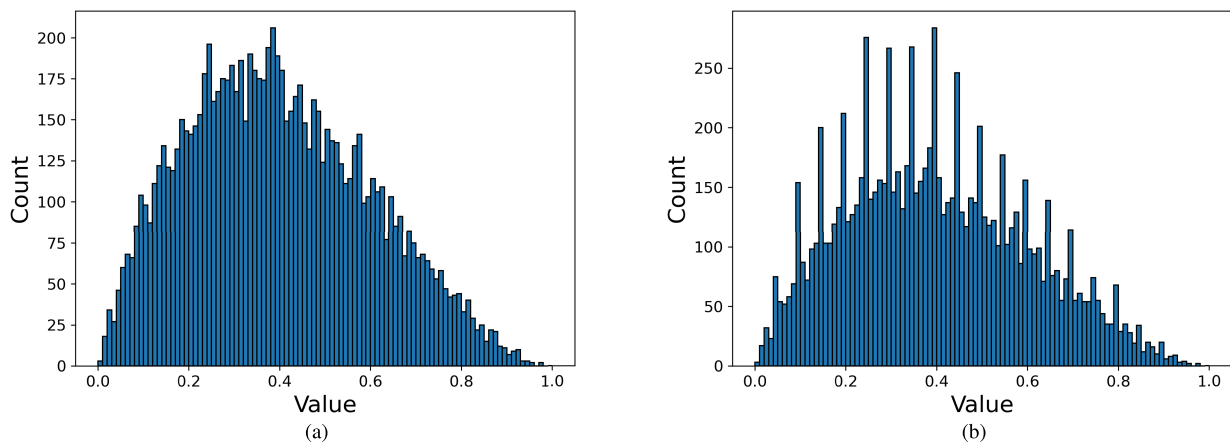
Overall, in terms of distribution outlier detection, the performance of the proposed method is only indirectly dependent on the inlier and the outlier distributions. As shown, it is



**FIGURE 8.** Comparing the proposed and the baseline method in terms of distribution outlier detection performance where 100 inlier random samples are drawn from the beta distribution with  $\alpha = 7$  and  $\beta = 1$ , while the triangular distribution with  $a = 0$ ,  $b = 1$ , and  $c = 0.5$  is used to draw the added a) 1 outlier sample histogram and b) 90 added outlier histograms.



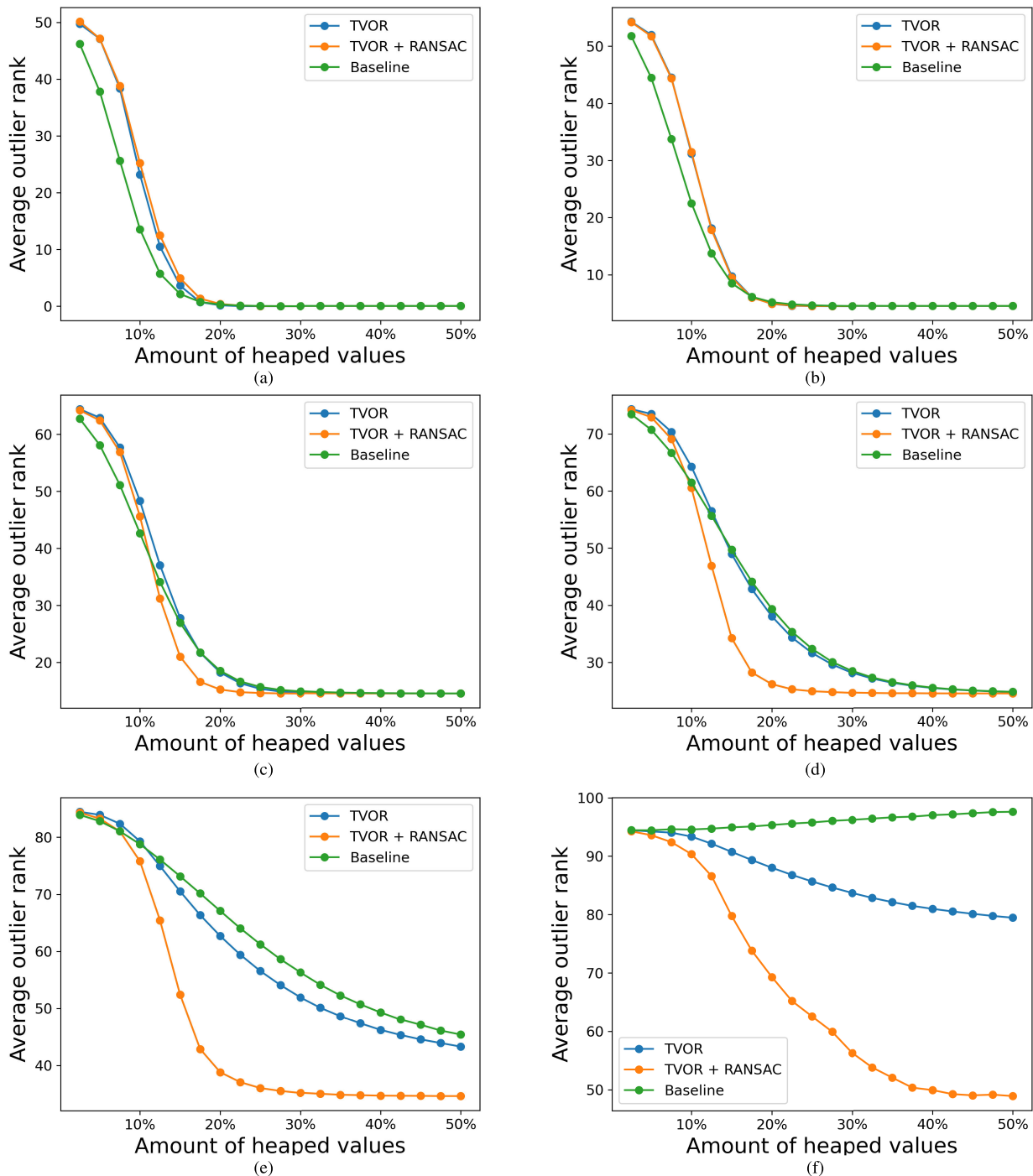
**FIGURE 9.** The comparison of the used beta and triangular distributions in terms of a) the theoretical discrete total variation  $\|\mathcal{D}\|_V$  described in Eq. (37) and b) the mean DTV calculated for  $10^6$  random samples of size 1000 for various values of  $n$ .



**FIGURE 10.** The histogram of a random sample drawn from the beta distribution with  $\alpha = 2$  and  $\beta = 3$  in the case of a) no heaping and b) heaping by moving 10% of randomly chosen items to bins with ordinal numbers divisible by 5 closest to them.

directly dependent on the difference between the theoretical DTVs of these distributions, which is in turn dependent on

the chosen histogram bins. This means that, depending on the histogram bins, the proposed method may perform well



**FIGURE 11.** Comparing the performance of the proposed and baseline methods averaged over  $10^4$  random trials in cases where the number of outlier random samples bin values added to the original 100 inlier random samples was a) 1, b) 10, c) 30, d) 50, e) 70, and f) 90. The inlier and outlier random samples were drawn from the beta distribution with  $\alpha = 2$  and  $\beta = 3$ , but the outlier samples were additionally changed in order to make their histograms have a prespecified amount of heaped bin values.

even when the inlier and the outlier distribution are same, but with slightly different parameters. On the other hand, for significantly different inlier and outlier distributions that have similar theoretical DTVs for the chosen bins, the proposed method may perform poorly. The opposite cases are also possible. Nevertheless, this is not too problematic because the proposed method was not designed for distribution

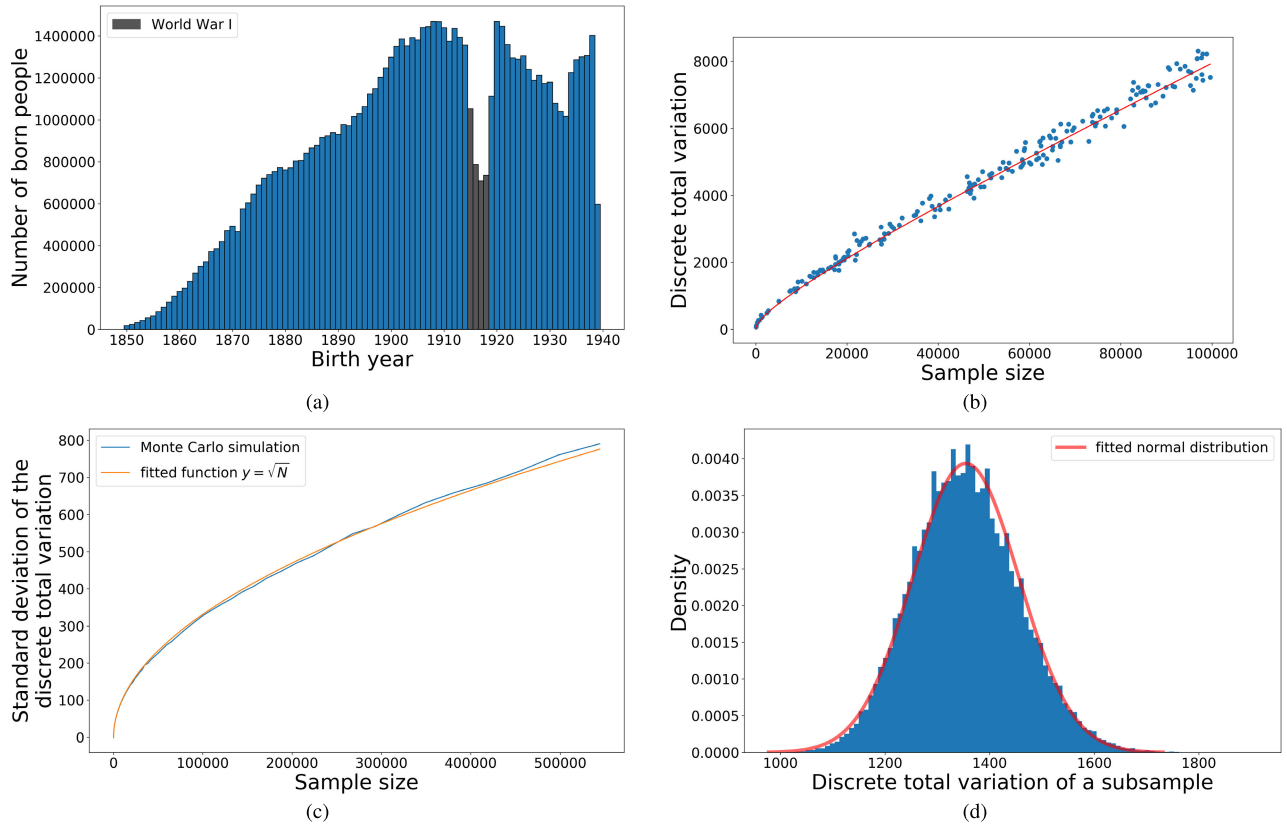
outlier detection, but specifically for the DTV outlier detection.

### C. SYNTHETIC DATA FOR TOTAL VARIATION OUTLIERS

#### 1) THE GOAL

The goal of this subsection is to demonstrate the behavior of the proposed method for the case that it was originally





**FIGURE 12.** The experiments on the German census of 1939: a) histogram of birth years of the German census of 1939 based on the data from [40], [41], starting from year 1850, b) fitting the proposed method's model in Eq. (49) to data for subsamples of the German census of 1939, c) the relation between the sample size and the standard deviation of the discrete total variation obtained through Monte Carlo simulations for the subsamples of the German census of 1939 and a fitted function  $y = a\sqrt{N}$ , and d) the distribution of discrete total variations obtained for 100k subsamples of the German census of 1939 of size 10k.

designed for, i.e. for discrete total variation outlier detection. Additional emphasis is specifically put on cases where the number of outliers gets very close to the number as the inliers.

## 2) EXPERIMENTAL SETUP

Since earlier in the paper it was mentioned that demographics is one of the fields that can benefit from discrete total variation outlier detection, the beta distribution with  $\alpha = 2$  and  $\beta = 3$  was chosen for the inlier samples' distribution. The reason is the resemblance of its histograms to the histograms of some population age distributions. For all experiments the number of bins  $n$  was fixed to 100. The outlier samples were initially also drawn from the same beta distribution and their histograms also had 100 bins. However, the outlier samples' histograms underwent an additional change to simulate the so called age heaping [16]. Namely, for a certain amount of randomly chosen bins with a count greater than 0, their count was decreased by 1 and the count of the closest bin to each of them whose ordinal number was divisible by 5 was increased by 1 as can be seen on the example that is shown in Fig. 10. This was done for various combinations of the amount of outlier samples and the amount of randomly chosen bins that were changed for these outlier samples' histograms. Finally,

the performances of the proposed method and of the baseline method were then compared for all these combinations.

## 3) RESULTS

The obtained results and comparisons are shown in Fig. 11. It can be seen that if there are only a few outliers, then the proposed and the baseline methods are on par with each other and there are only some smaller differences in performance for various amount of heaped values. However, as the number of outliers increases, the proposed method starts to significantly outperform the baseline method, especially in cases where RANSAC is used as suggested in Section III-E. This is especially noticeable in Fig. 11f where the number of outliers is very close to the number of inliers. There the baseline effectively degrades to a random chooser, while the proposed method used in combination with RANSAC excels at outlier detection. This shows the usefulness of the proposed methods for the task of finding the discrete total variation outliers.

## D. CENSUS DATA

### 1) THE GOAL

The goal of this subsection is to test the proposed method on an example of real-life census data with sample sizes

spanning several orders of magnitude and being drawn from slightly different, but similar distributions. Here a closer look is taken at the samples of the top-scoring histograms. This can show the robustness of the proposed method in noisy conditions and its usefulness for real-life data applications.

## 2) EXPERIMENTAL SETUP

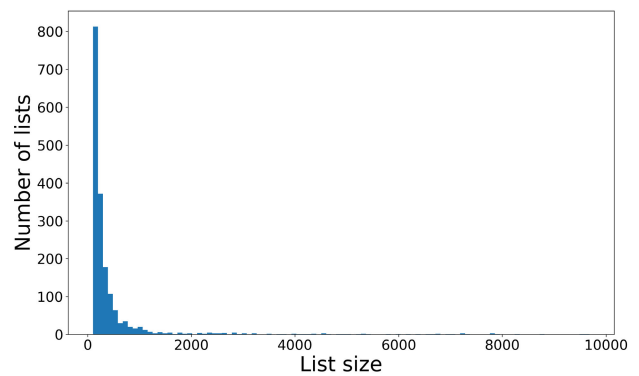
Several census data sources have been used for the experimental setup. The largest of them is the German census of 1939 [40] with the corresponding birth year histogram being shown in Fig. 12a. Since the significant gap for the years of World War I can be traced in age composition of other similar lists and censuses of other countries as well [46], [47], this census data is used here as a gold standard for the discrete total variation of the population histograms for that time.

In addition to that, 7106 variously sized censuses, i.e. lists of people with birth years available at the website of the United States Holocaust Memorial Museum (USHMM) [48] are used since they were composed for the populations from roughly the same time frame. The distribution of the majority of the sizes with the largest ones being excluded for practical purposes is shown in Fig. 13. The geographical locations of these populations differ, but they still mostly cover the populations whose birth year histograms should have similar discrete total variation properties. To make it clear immediately, this does not necessarily mean that the age distributions are similar as well. Namely, one census can have a significantly higher amount of e.g. young people in comparison to other censuses, but as it will be shown later on concrete examples, this should not necessarily affect the discrete total variation of the birth year histograms too significantly. Therefore, these lists available at USHMM constitute an interesting dataset in which to look for outliers in terms of discrete total variation.

## 3) RESULTS

The first experiments that were carried out consisted of simply taking many variously sized subsamples of the birth years from the German census of 1939, calculating the discrete total variations of their birth year histograms, and fitting the proposed method's model in Eq. (49) to the data obtained in this way. Fig. 12b shows the result of this experiment. The proposed model fits well to all data. This also holds for smaller subsamples where the influences of the two terms in Eq. (49) are still on par. It can also be seen how the discrete total variations get more dispersed as the sample size grows. While this may hint at heteroscedasticity, applying weighted regression or variance-stabilizing data transformations did not significantly change the results that are described here.

The relation between the sample size and the standard deviation of the discrete total variation is shown in Fig. 12c. Very similar results are obtained for other distributions as well. It can be seen that the relation is very similar to the square root function, which effectively justifies the use of Eq. (51) for practical purposes. The distribution of the discrete total variations for the subsamples of the same size



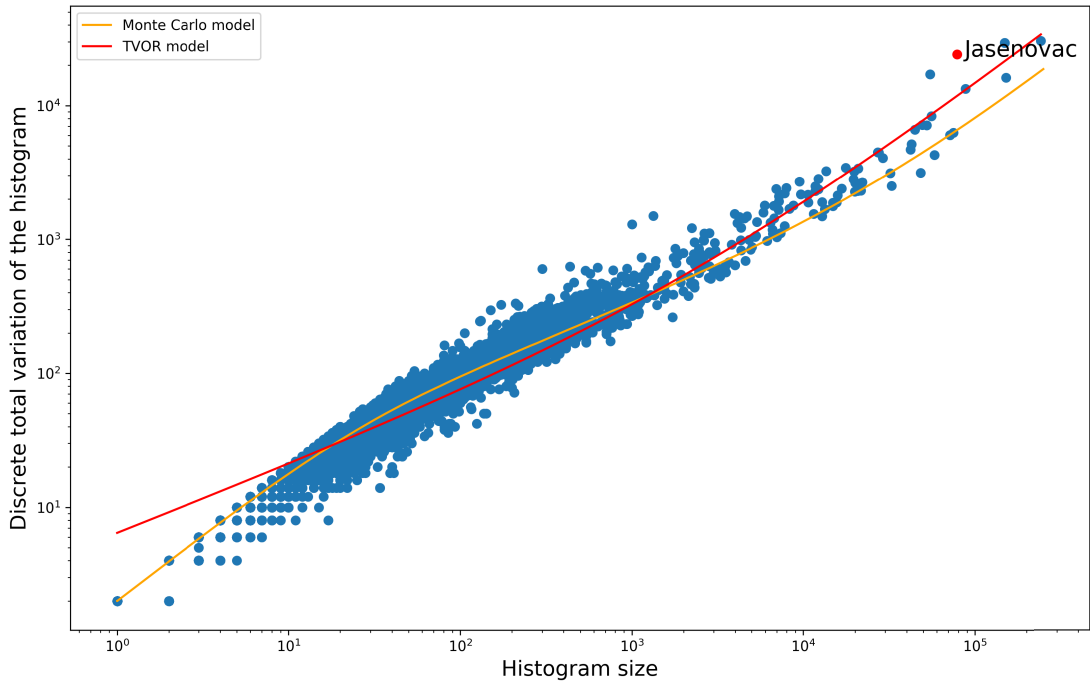
**FIGURE 13.** The distribution of the sizes of the majority of the 7106 USHMM lists that are used for the experiments.

closely resembles the normal one as shown in Fig. 12d with the remark that the discrete total variations there are integers.

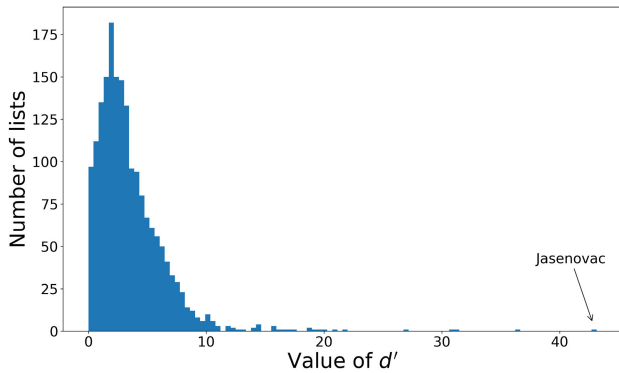
After conducting the relatively simple mentioned experiments in order to get a better insight into the inner workings of the proposed method, the next step was to apply the method to all USHMM lists whose data includes birth years. The distribution of the values of  $d'$  described in Eq. (51) and obtained by the proposed method in this way is shown in Fig. 15, while the relation between the calculated discrete total variations and the predicted ones are shown in Fig. 14.

It can be seen that the majority of the values  $d'$  in Fig. 15 are not spread too widely with the exception of several outliers. Before analyzing these outliers in more detail and commenting on Fig. 14, it must be stressed that in Fig. 14 the plot axes use the logarithmic scale to better accommodate the presentation to the list's size distribution. Therefore, the apparent misfit for the smallest lists can deceive into believing that the proposed model failed to fit properly, while it is actually only a misfit on a small scale. For similar reasons many of the differences between the calculated and the predicted discrete total variations for the larger lists are higher than they may appear to be on the plot. In addition to showing the proposed method's model, the Monte Carlo model based on the average discrete total variations of the variously sized subsamples of the German census of 1939 is shown in Fig. 14 for comparison. It can be seen that on several places its predictions are not quite aligned with the ones of the proposed model, which can be attributed to the distribution shown in Fig. 13, i.e. to the significant influence of samples of certain sizes during the model fitting. This can be alleviated by using techniques such as taking only samples of evenly spaced sizes, but as shown later in this subsection, the top results for the two models do not differ significantly even without applying such techniques. Therefore, the application of such techniques was omitted.

Out of the 7106 lists that were analyzed, the top three outlier lists in terms of  $d'$  were the Jasenovac camp inmates list [42] with  $d' \approx 43.13$ , the list of the Soviet Extraordinary Commission [43] with  $d' \approx 36.5$ , and the list for the Franz Street Number 38 [44] with  $d' \approx 31.29$ . The histograms for



**FIGURE 14.** Applying the proposed method to 7106 lists of the USHMM data. The model based on applying the Monte Carlo simulation to the German census of 1939 is shown for comparison. Note that the plot axes use the logarithmic scale.



**FIGURE 15.** The distribution of values  $d'$  calculated by the proposed method for birth years from the USHMM lists.

these lists are shown in Figs. 16a, 16b, and 16c, respectively. A more detailed analysis of the top-scoring histogram that provides additional insights and explanations of the behavior of the proposed method's scoring is available in Appendix.

In the case of Monte Carlo the score  $d''$  was calculated as

$$d'' = \frac{\|\mathbf{x}_n\|_V - \hat{\mu}_N}{\hat{\sigma}_N} \quad (52)$$

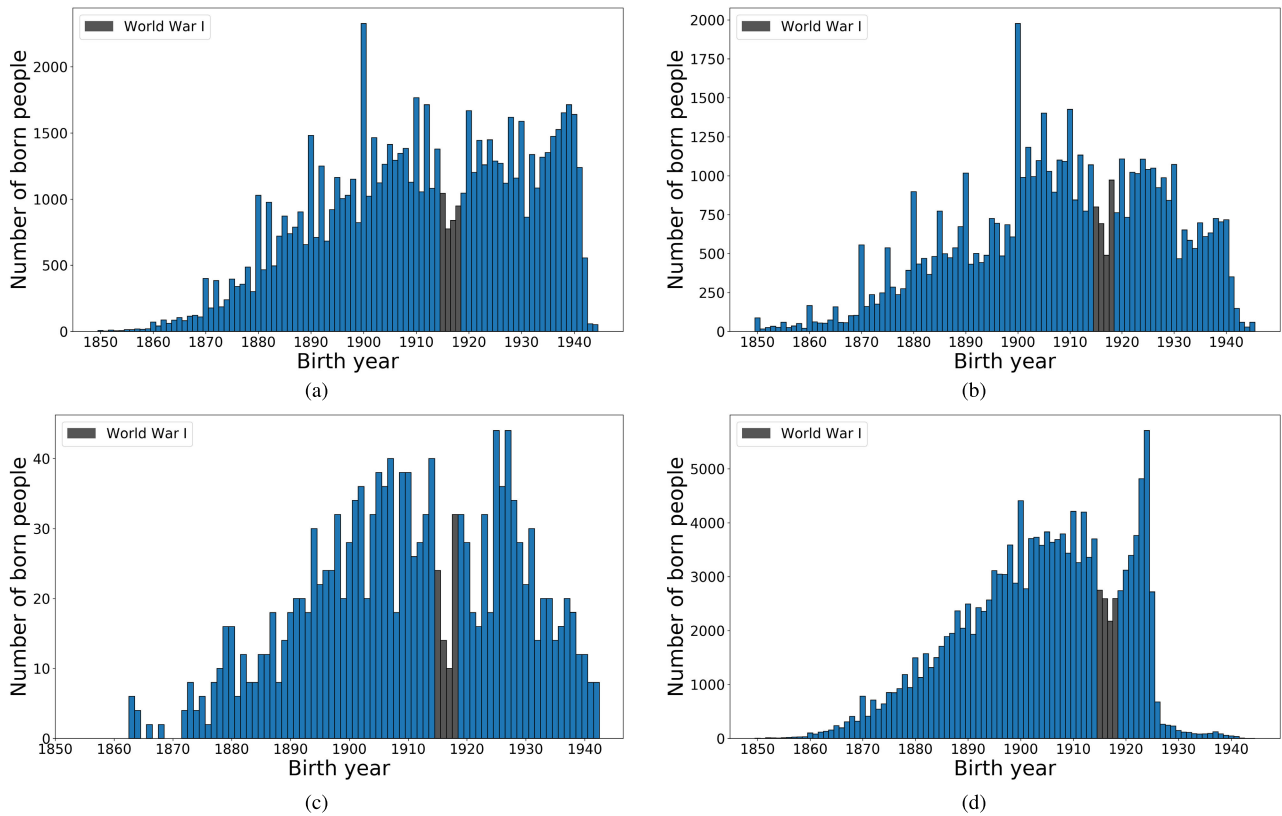
where  $\hat{\mu}_N$  and  $\hat{\sigma}_N$  are the mean and the standard deviation, respectively, of the discrete total variation obtained for a large number of subsamples of size  $N$  of the German census of 1939. The distribution of the values of  $d''$  obtained for all lists from the USHMM is shown in Fig. 17. The lists with the first and second highest value of  $d''$  were the same as for

$d'$  with  $d'' \approx 58.14$  and  $d'' \approx 49.04$ , while the list with the third highest value of  $d''$  was the list of Jewish refugees in Tashkent [45] with  $d'' \approx 44.51$  and with the corresponding histogram shown in Fig. 16d. Already by looking at the mentioned figures for the top-scoring lists it can be seen that their corresponding histograms indeed have high values of discrete total variation with spikes, i.e. individual bins that significantly differ from their neighbors, which contrasts the smoothness of the histogram for the German census of 1939.

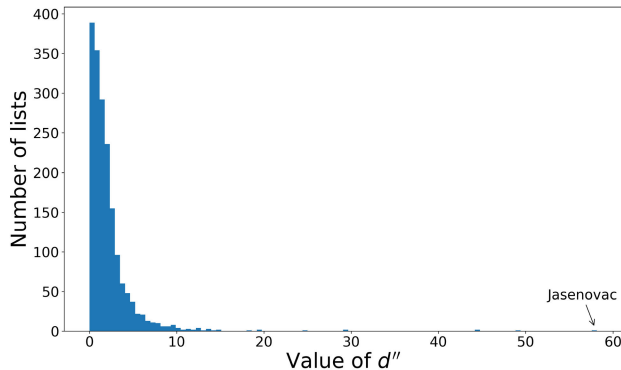
### E. ADVANTAGES OVER EXISTING METRICS

Like for many other groups of population histograms, there is no ground-truth ordering for USHMM lists in terms of their histograms' smoothness or accordance with historical populations. Because of that, the quality of ordering obtained by the proposed method and by existing metrics such as Whipple's and Myers' indices can not be compared directly. However, it is possible to show cases that are problematic for both of these indices, but not for the proposed method.

The first example is the histograms shown by Figs. 18a and 18b, which represent the top-scoring histograms among the USHMM lists' histograms for the Whipple's and Myers' indices, respectively. It can be seen that these histograms are actually relatively smooth, but they also contain only a few non-zero values: the first one 8 and the second one only a single. These histograms can hardly be considered outliers in terms of smoothness when compared to the histograms in Fig. 16, but rather outliers in terms of covered years span, which is different and also detectable by much simpler techniques. Additionally, the lists



**FIGURE 16.** Top-scoring birth year lists out of 7106 checked lists: a) the Jasenovac camp inmates available at USHMM's webpage [42], b) the victims from the Soviet Extraordinary Commission [43], c) the victims from the Franz Street Number 38 [44], and d) persons from the Registration cards of Jewish refugees in Tashkent, Uzbekistan during WWII [45].



**FIGURE 17.** The distribution of values  $d''$  calculated by the proposed method for birth years from the USHMM lists.

that produced these histograms have only a relatively small number of birth years and since the mentioned indices, unlike the proposed method, do not take into account the sample size, they are also more prone to anomalies that arise in smaller samples due to randomness.

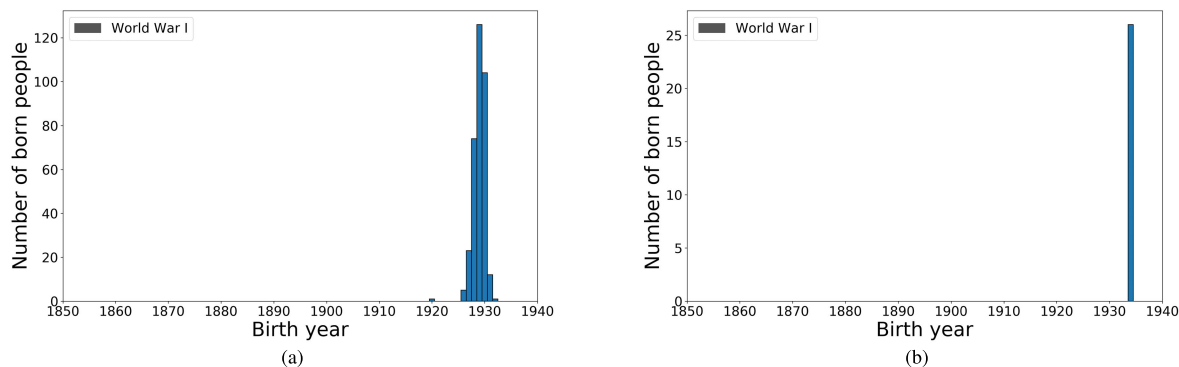
Another problem with metrics such as Whipple's and Myers' indices is that they are mainly concerned with frequencies and do not take into account other properties such as shape or smoothness. Because of that, for different samples that have the same frequencies of last digits of their numbers,

it is still possible to obtain the same values of the mentioned indices even if the samples' histograms differ significantly. An example of this is given in Fig. 19 with a fully smooth histogram that has the same indices values as a histogram that can hardly be considered smooth. While numerous similar examples exist, the ones presented are enough to show the frequency-based weakness of the Whipple's and Myers' indices. On the other hand, the proposed method has no such problems and its values for the histograms in Fig. 19 differ significantly with one being zero and the other one non-zero.

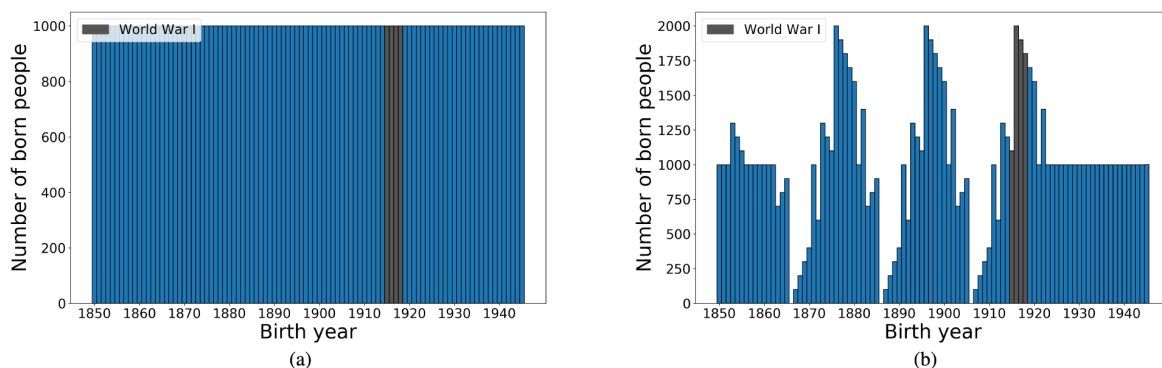
In short, while being widely used and useful in certain cases, metrics such as the Whipple's and the Myers' indices are too simple to properly handle properties such as smoothness. Therefore, the proposed method's ability to specifically target smoothness is its main advantage over other metrics.

## F. DISCUSSION

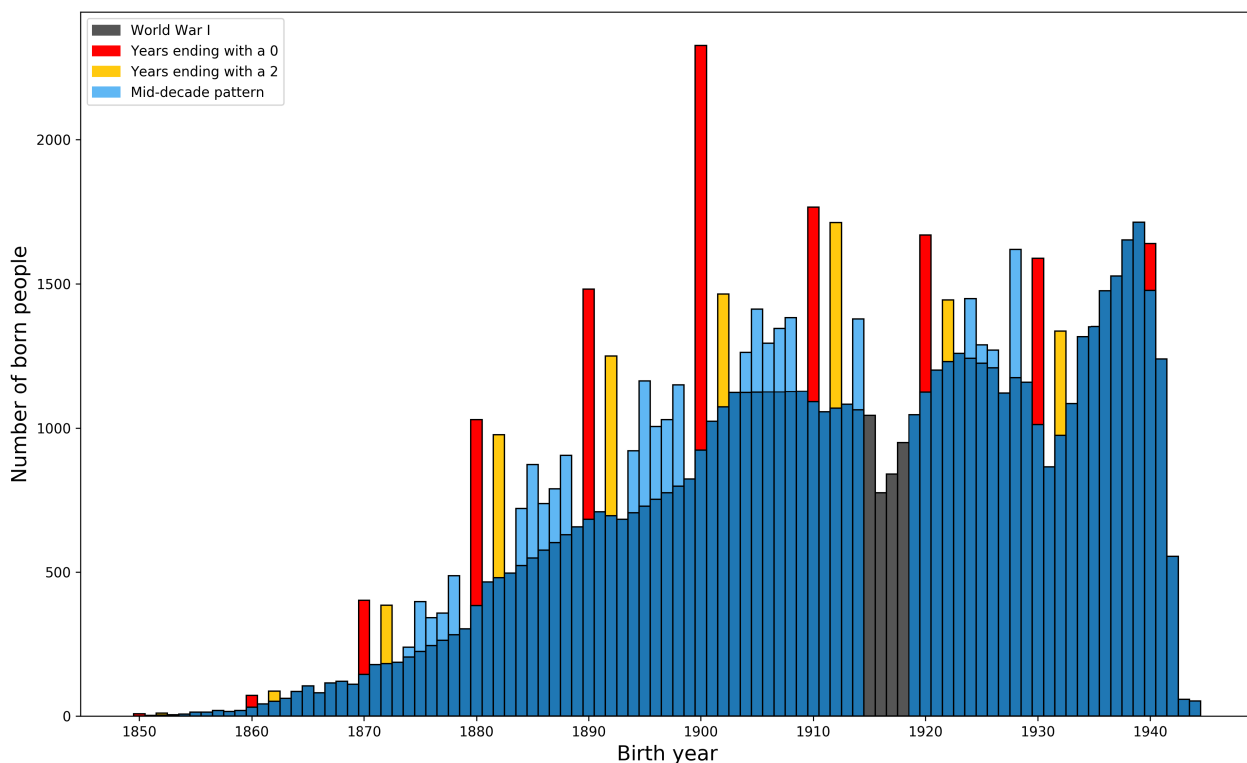
Looking at the distributions shown in Figs. 15 and 17 and observing the significant difference between the majority of the scores and the highest scores, it can be concluded that the histograms of the used USHMM lists that obtained the highest scores are indeed outliers in terms of the discrete total variation. Since the analyzed data consisted of birth years, it may seem that an appropriate tool for identifying outliers such as the ones in Fig. 16 could be the Whipple's index [46],



**FIGURE 18.** Top-scoring birth year lists out of 7106 checked lists for a) the Whipple's index and for b) the Myers' index.



**FIGURE 19.** Examples of histograms for all of which the Whipple's and the Myers' indices have exactly the same values.

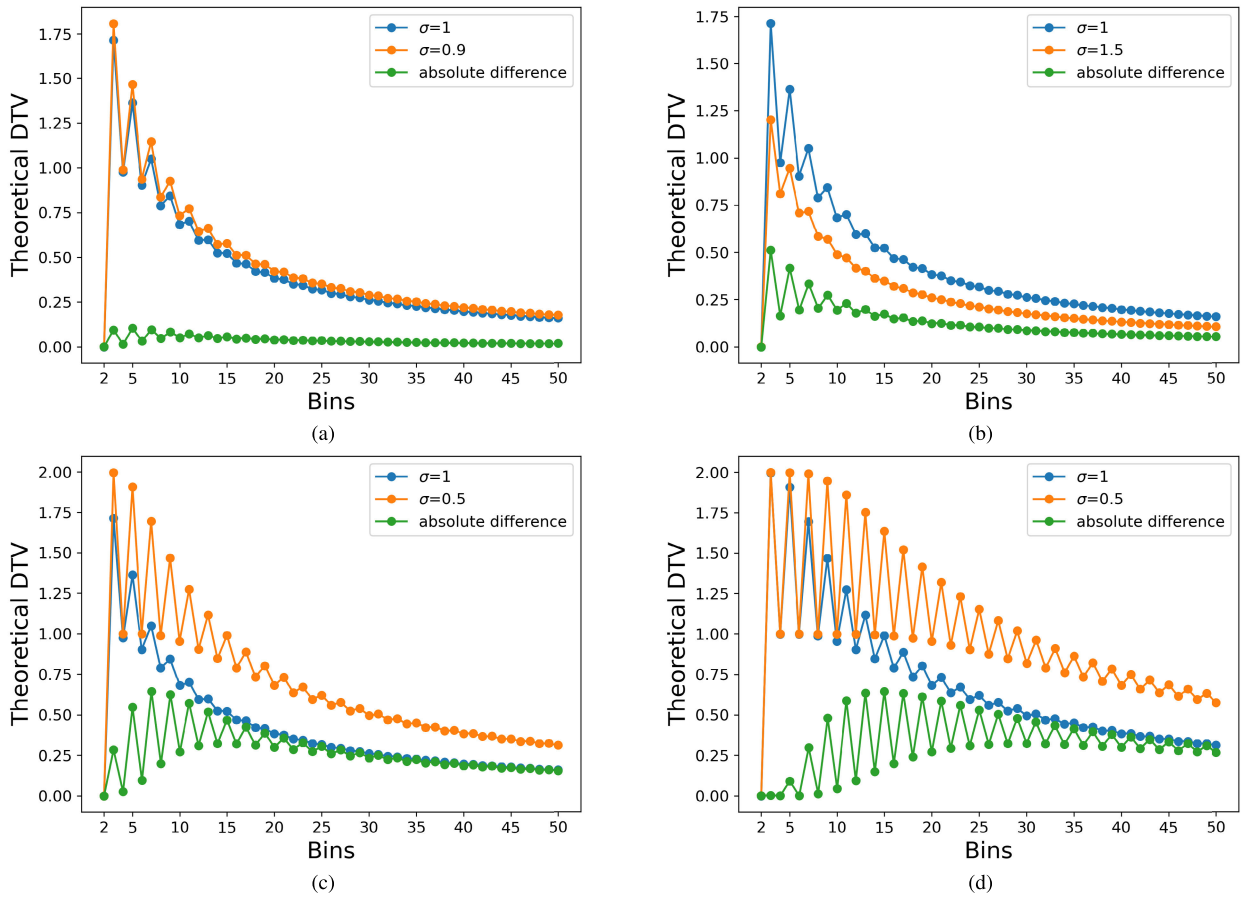


**FIGURE 20.** Same data for Jasenovac inmates as in Fig. 16a, but with additional markings for the age heaping [16] artifacts.

but due to its fixed nature of checking only specific kinds of data, it is often inappropriate [49], [50]. This also holds

in the case of the histogram of the Jasenovac inmates shown in 16a whose artifacts are marked more closely in Fig. 20.





**FIGURE 21.** The comparison of the values of the theoretical discrete total variation  $\|\mathcal{D}\|_V$  of the histograms of normal distribution  $\mathcal{N}(0, \sigma^2)$  for the values in the interval  $[-b, b]$  for various number of histogram bins used to obtain the experimental results that were shown earlier in Fig. 7: a)  $c = 5, \sigma = 0.9$ , b)  $c = 5, \sigma = 1.5$ , c)  $c = 5, \sigma = 0.5$ , and d)  $c = 10, \sigma = 0.5$ .

It can be seen that age heaping occurs in several forms that the Whipple's index not only cannot pick up, but it also gets hampered by them. Namely, in its slightly changed form the Whipple index checks for a surplus of years ending in 0 or 5 when compared to other years, but in the case of Jasenovac there is also a surplus of years ending in 2, which is not checked by the Whipple's index and it actually reduces the overall surplus of years ending in 0 or 5, thus hampering the Whipple's index in detecting the unusual data patterns. Since the proposed method has no such problems, it may be more appropriate in situations similar to the one in this experiment.

Besides all these histogram artifacts, there are other peculiarities with the Jasenovac list. Namely, if it is compared to other USHMM lists used here, it directly contradicts some of them. For example, the list available at [51] states that a certain Stanko Nick survived the war [52], while the Jasenovac list claims that he was killed [53], which is known to be wrong [54]. In another example, the list available at [55] states that a certain Josip Stern arrived at Auschwitz in 1942 [56], while the Jasenovac list claims that he was killed in 1941 [57]. This means that the proposed method can also

be used to detect samples that contain potentially problematic data with properties not always shared with the usual outliers.

### G. SOURCE CODE AND DATA REPOSITORY

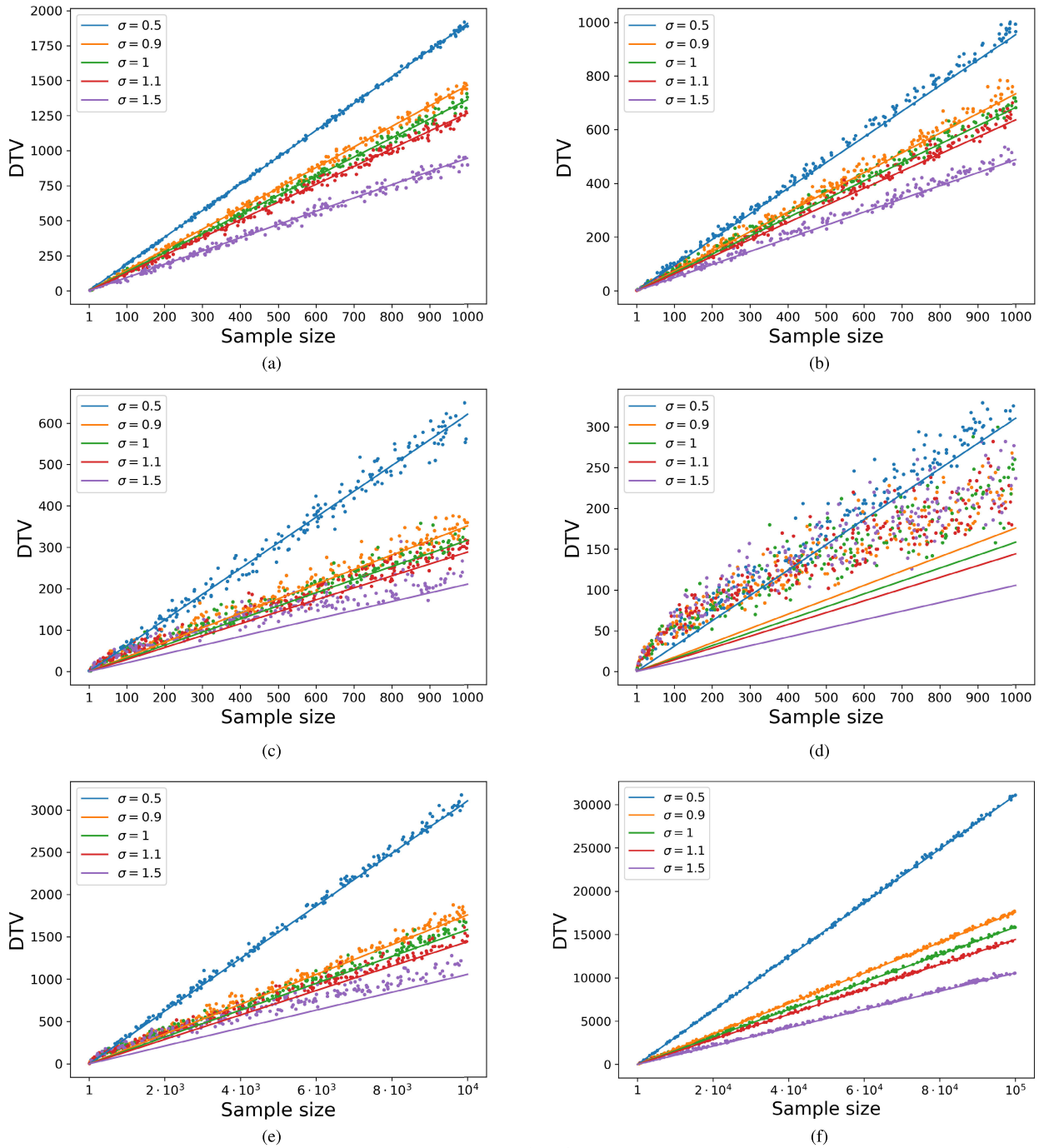
The source code written in the Python programming language and the data required to recreate the results described in this section are publicly available in a dedicated GitHub repository.<sup>1</sup> At the time of writing this paper, the census data used in this section was publicly available at the USHMM website, but for the sake of simplicity of recreating the results, it is also available in the repository. While the census data also contains other information alongside the birth years, only the birth years were copied to the repository in order to avoid data privacy violation for potentially still living persons. For example, according to the Jasenovac camp inmates list [42], which was already shown to be problematic, a certain Stojan Ražokrak [58] was allegedly killed in 1942, but a publicly available video of him<sup>2,3</sup> from 2012 and its transcript<sup>4</sup> clearly

<sup>1</sup><https://github.com/DiscreteTotalVariation/TVOR>

<sup>2</sup><https://www.youtube.com/watch?v=S5IRwT63as0>

<sup>3</sup><https://archive.is/48sKw>

<sup>4</sup><https://archive.is/RtnsJ>



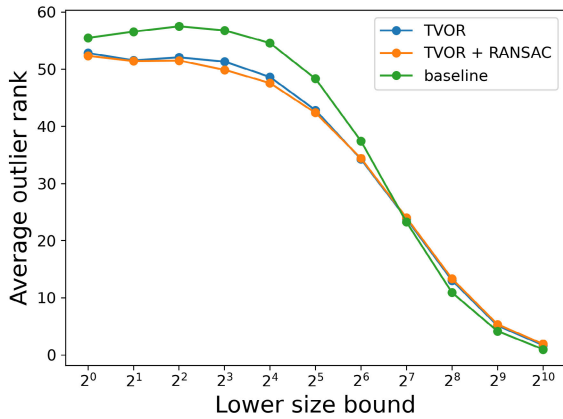
**FIGURE 22.** The DTVs of histograms of random samples drawn from  $\mathcal{N}(0, \sigma^2)$  and of sizes randomly chosen to be between 1 and  $U$ . The number of bins  $n$  and the upper size bound  $U$  are set to a)  $n = 5$  and  $U = 1000$ , b)  $n = 10$  and  $U = 1000$ , c)  $n = 25$  and  $U = 1000$ , d)  $n = 50$  and  $U = 1000$ , e)  $n = 50$  and  $U = 10^4$ , and f)  $n = 50$  and  $U = 10^5$ . The lines represent the value of  $\|\mathcal{D}\|_V$  described in Eq. (37) and multiplied by the sample size, while the dots represent the random samples.

show the opposite. Because of that, it seemed reasonable to copy only the birth years, while any interested reader can check the rest of the data at the USHMM website by using the appropriate list identifier given in the repository.

## V. CONCLUSION AND FUTURE WORK

In this paper, a method for finding discrete total variation outliers among histograms has been proposed. It scores

histograms based on the deviation of their discrete total variation from its expected value. To carry out this scoring, a statistical framework has been proposed. One of the method's main advantages is that in order to work it requires no information about the distribution of the samples that are being described by histograms. In some special cases the proposed method even outperforms the Pearson's chi-squared test when looking for the outlier histograms in terms of the



**FIGURE 23.** The dependence of the performance of the baseline and the proposed method on the random samples' size range when 100 inlier samples are drawn from  $\mathcal{N}(0, 1)$ , a single outlier sample is drawn from  $\mathcal{N}(0, 0.9^2)$ , the number of bins  $n$  is 15,  $c = 5$ , and the size of the inlier and the outlier samples is randomly chosen to be between  $L$  and  $10 \cdot L$  where  $L$  is the lower size bound that is shown on the x-axis.

sample distribution despite the fact that it was not designed for this task. On the other hand, the proposed method clearly outperforms the Pearson's chi-squared test when looking for discrete total variation outliers, especially in cases of a huge amount of outliers. Overall, the proposed method represents a successful proof-of-concept of how discrete total variation that is used in the method's modelling can be applied to histogram outlier detection in terms of discrete total variation, which has been experimentally confirmed on synthetic and real-life data. Future work may include looking for some other histogram properties that can also be used for histogram outlier detection in terms of their smoothness in the cases where the distribution of the histogram samples is unknown. As for improving the proposed method, future work will include at least two things. The first of them is the analysis of variance for the discrete total variation to potentially improve the scoring criteria. The second of them comprises other aspects of the histogram's discrete total variation that could decrease the scores obtained for the inlier samples, but simultaneously keep the scores obtained for the outliers high.

## APPENDIX

### A. PROOFS OF THE THEOREMS

#### 1) PROOF OF THEOREM 1

By the definition of  $F(2, N)$ , it can be developed as follows:

$$\begin{aligned} F(2, N) &= \mathbb{E}[\|\mathbf{x}_2\|_V] \\ &= 2^{-N+1} \sum_{k=0}^{\lfloor \frac{N-1}{2} \rfloor} \binom{N}{k} (N-2k). \end{aligned} \quad (53)$$

For an even  $N = 2r$ , the equality  $\sum_{k=0}^N \binom{N}{k} = 2^N$  leads to

$$\begin{aligned} &\sum_{k=0}^{r-1} \binom{2r}{k} (2r-2k) \\ &= 2r \sum_{k=0}^{r-1} \binom{2r}{k} - 4r \sum_{k=1}^{r-1} \binom{2r-1}{k-1} \end{aligned}$$

$$\begin{aligned} &= r \left( 2^{2r} - \binom{2r}{r} \right) - 2r \left( 2^{2r-1} - 2 \binom{2r-1}{r-1} \right) \\ &= -r \binom{2r}{r} + 2r \binom{2r}{r} = r \binom{2r}{r}. \end{aligned} \quad (54)$$

Since here  $\lfloor (N+1)/2 \rfloor = \lfloor (2r+1)/2 \rfloor = r$  and  $\lfloor N/2 \rfloor = \lfloor (2r)/2 \rfloor = r$ , it follows that Eq. (54) matches Eq. (19). For an odd  $N = 2r + 1$ , a similar calculation as before gives

$$\sum_{k=0}^r \binom{2r+1}{k} (2r+1-2k) = (r+1) \binom{2r+1}{r}. \quad (55)$$

To avoid any possible confusion, it has to be mentioned that the lower index of the binomial coefficient in Eq. (55) can also be set to  $r+1$  because  $N$  is supposed to be odd there. Furthermore, like in the previous case, it can be seen that Eq. (55) also matches Eq. (19), which proves Theorem 1.

#### 2) PROOF OF THEOREM 2

The expectation  $\mathbb{E}[|x_2 - x_1|]$  can be written as follows:

$$\mathbb{E}[|x_2 - x_1|] = n^{-N} \sum_{k_1 + \dots + k_n = N} \binom{N}{k_1, \dots, k_n} |k_2 - k_1|. \quad (56)$$

The right-hand side of Eq. (56) can further be written as

$$\begin{aligned} &n^{-N} \sum_{k_1 + k_2 \leq N} \binom{N}{k_1, k_2, N-k_1-k_2} \\ &\quad \times |k_2 - k_1| \sum_{k_3 + \dots + k_n = N-k_1-k_2} \binom{N-k_1-k_2}{k_3, \dots, k_n} \\ &= n^{-N} \sum_{k_1 + k_2 \leq N} \binom{N}{k_1, k_2, N-k_1-k_2} \\ &\quad \times |k_2 - k_1| (n-2)^{N-k_1-k_2} \\ &= \left( \frac{n-2}{n} \right)^N \sum_{k_1 + k_2 \leq N} \binom{N}{k_1, k_2, N-k_1-k_2} \\ &\quad \times (n-2)^{-(k_1+k_2)} |k_2 - k_1|. \end{aligned} \quad (57)$$

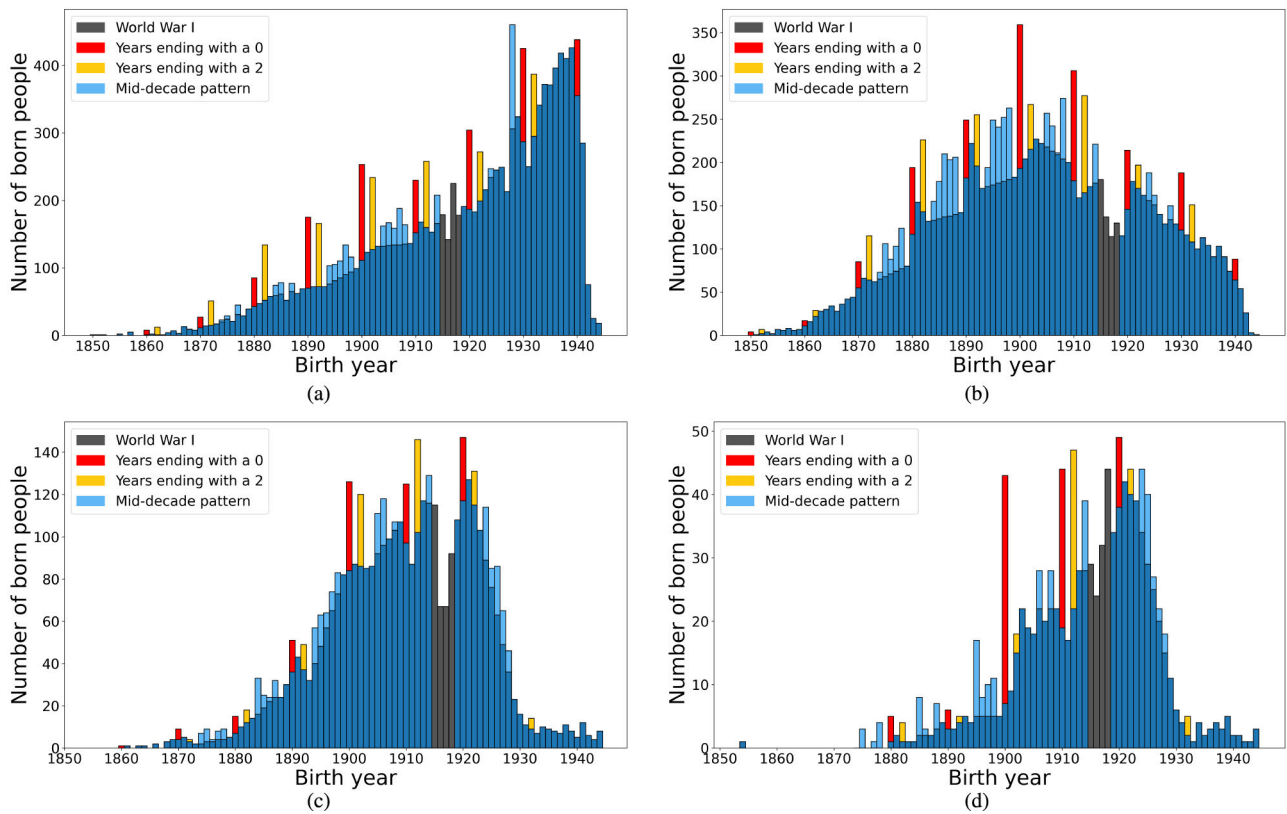
To obtain Eq. (23) from here, it is sufficient to note that

$$\begin{aligned} \mathbb{E}[|x_2 - x_1|] &= \mathbb{E}[x_2 - x_1 \mid x_2 > x_1] \\ &\quad + \mathbb{E}[x_1 - x_2 \mid x_2 < x_1] \\ &= \mathbb{E}[x_2 - x_1 \mid x_2 > x_1] \\ &\quad + \mathbb{E}[x_2 - x_1 \mid x_1 < x_2] \\ &= 2\mathbb{E}[x_2 - x_1 \mid x_2 > x_1]. \end{aligned} \quad (58)$$

Eq. (58) can be applied to Eq. (57), which can then be applied to Eq. (21). This results in Eq. (23), which proves Theorem 2.

### B. THEORETICAL DISCRETE TOTAL VARIATIONS

To facilitate a better understanding of the experimental results that were discussed in Section IV-B and shown in Fig. 7, the comparison of the values of the theoretical discrete total variation  $\|\mathcal{D}\|_V$  of the histograms of normal distribution with parameters used to obtain these results are given in Fig. 21. By comparing Figs. 7 and 21, it is relatively easy to explain phenomena such as the sudden drops in the proposed method's



**FIGURE 24.** Birth year histograms of Jasenovac camp inmates [42] by nationality with markings for age heaping: a) Roma inmates, b) Jewish inmates, c) Croatian inmates, and d) Muslim inmates. Only the histograms for nationalities for which there are more than 1000 listed inmates are shown here, while the histogram for the Serbian inmates is given separately in Fig. 25.

performance that can be seen in Fig. 7d when 10 bins are used. Namely, Fig. 21d clearly shows that for 10 bins the difference between the theoretical DTVs of the distributions used there is very small, which renders the proposed method inadequate for recognizing outlier samples for that specific case. Similar reasoning can also be applied to successful cases where this difference is sufficiently large.

### C. DEPENDENCE OF VARIATION ON THE SAMPLE SIZE

Fig. 9 clearly shows how randomness can have a significant impact on the performance of the proposed method. Nevertheless, as described by Eq. (48), when the samples' sizes grow, this impact becomes ever smaller. However, in order to decrease this impact in cases of e.g. larger values of  $n$ , the samples' sizes have to grow significantly more than in the cases of smaller values of  $n$ . This is illustrated on several examples shown in Fig. 22. There it can be seen that for  $n = 5$  the samples with random sizes up to 1000 are clearly separated, while for the same sizes and  $n = 50$  the samples can hardly be separated. However, as shown in Fig. 22e and Fig. 22f, if the upper bound for the sizes of random samples gets increased even further, the separation again becomes clear. As shown in Fig. 23, this has a direct influence on the performance of both the baseline and the proposed methods.

In short, a successful application of the proposed method assumes a reasonably high ratio between the number of bins  $n$  and the sizes of samples. How high this ratio should be, however, depends on the specific distributions of the samples.

### D. PARTITIONING THE TOP-SCORING HISTOGRAM

In order to describe the behavior of the proposed method in more detail, it may be useful to additionally analyze the top-scoring histogram shown in Figs. 16a and 20. By partitioning the initial birth year sample into more smaller samples, it is possible to examine the behavior of the proposed method when the sample size is changing. One way of partitioning the sample is by nationality of the inmates. The nationalities for which there are more than 1000 listed inmates are, as specified in the Jasenovac inmates list, the following ones: Serbian, Roma, Jewish, Croatian, and Muslim. While the histograms of the Roma, Jewish, Croatian, and Muslim nationalities shown in Fig. 24 all exhibit signs of age heaping similar to the ones in Fig. 20, by far more prominent signs are exhibited by the Serbian nationality as shown in Fig. 25.

If the histograms for separate nationalities are also added to the set of USHMM lists and the proposed method is applied to this extended set, then the histogram for the Serbian nationality ends up being the second most likely outlier just after the whole Jasenovac list with  $d' = 40.82$ . The Romani

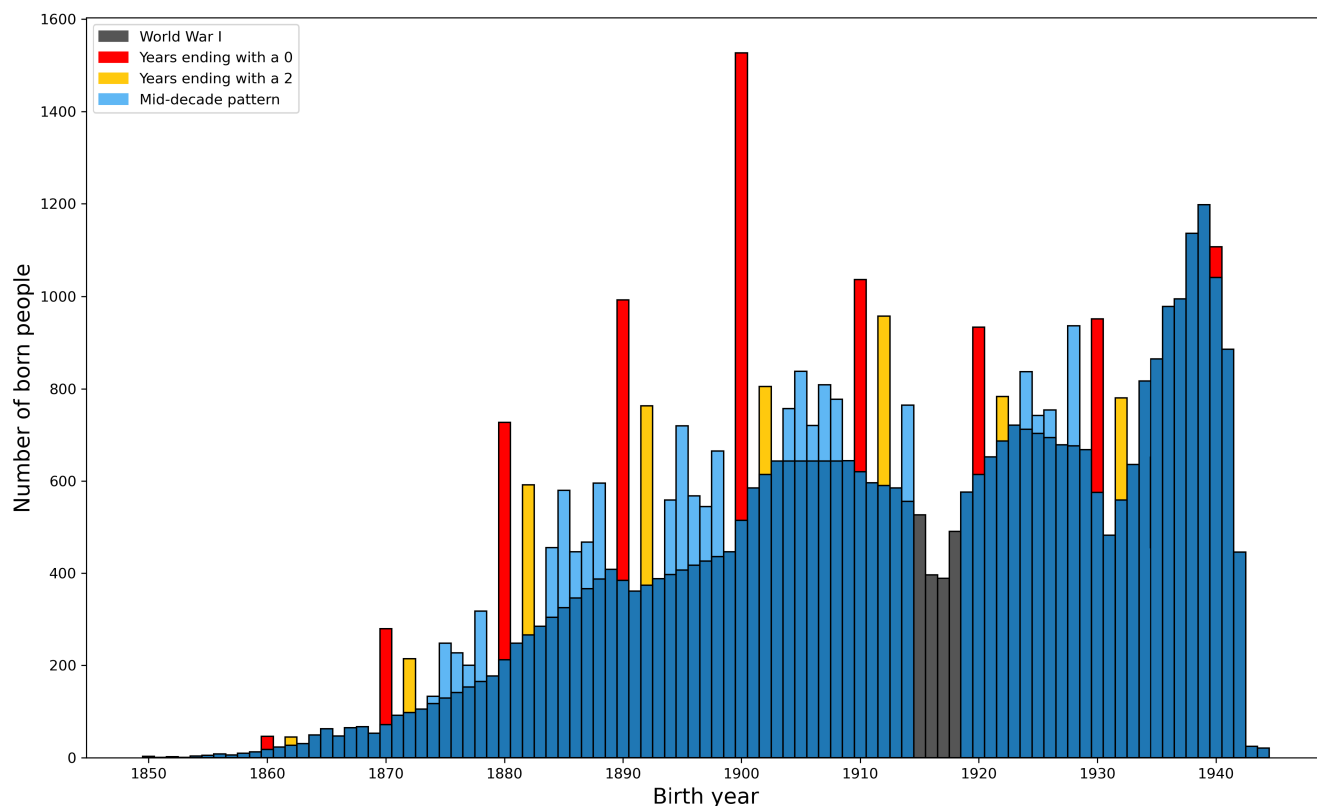


FIGURE 25. Birth year histogram of Jasenovac inmates of Serbian nationality with same markings for age heaping as in Fig. 20.

nationality histogram ends up on the 21st place with  $d' = 15.12$ , the Jewish nationality histogram ends up on the 66th place with  $d' = 9.08$ , while other histograms are not inside the 100 most likely outliers. This shows how the proposed method can also be used to detect the potentially problematic parts of a sample, which in the case of the Jasenovac list lies in the birth years of Serbian inmates.

Additionally, there is another thing to be observed here. Namely, while Figs. 20 and 25 seem to be very similar, the score  $d'$  for the histogram of the birth years of the Serbian inmates was nevertheless smaller than the one for the whole Jasenovac list. This has to do with the fact that the sample with birth years of Serbian inmates has fewer values than the whole Jasenovac list, i.e. it makes up roughly 57% of the 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  $d'$ .

## ACKNOWLEDGMENT

(Nikola Banić and Neven Elezović contributed equally to this work.) The authors would like to thank the reviewers for their constructive and useful feedback, which significantly helped in improving the paper. Additionally, they would like to thank Prof. Branko Jeren for the discussions, advice for a clearer presentation, and his networking efforts, Dr. Juraj Radić for the significant help on the initial theoretical derivation of the used statistical model, Dr. Mladen Koić for motivating to provide better visualization and explanations, Dr. Josip Pečarić and Dr. Josip Stjepandić

for their role in motivating the research, Dr. Tomislav Petković for his useful advice on making the paper more readable, Dr. Vuko Brigljević for his advice on language improvement, Dr. Stjepan Šterc for his advice on some demographic topics, Dr. Viktoria Oliver for her proof-reading of the paper and suggestions on English style improvement as a native speaker, Dr. Ivan Hrvoić and Ivana Kovačević Mandac for connecting the authors with Dr. Oliver. Finally, the authors would like to thank Dr. Julio Guijarro Garcia, Dr. Josep Peguera Poch, Dr. Jorge Ramos, and Dr. Jure Bogdan for their kind support.

## REFERENCES

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Outlier detection: A survey," *ACM Comput. Surv.*, vol. 14, Aug. 2007, p. 15.
- [2] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.
- [3] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.
- [4] K. Pearson, "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material," *Philos. Trans. Roy. Soc. London A*, vol. 186, no. 186, pp. 343–414, 1895.
- [5] H. A. Sturges, "The choice of a class interval," *J. Amer. Stat. Assoc.*, vol. 21, no. 153, pp. 65–66, Mar. 1926.
- [6] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.
- [7] D. Freedman and P. Diaconis, "On the histogram as a density estimator:  $\ell_2$  theory," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.
- [8] H. Shimazaki and S. Shinomoto, "A method for selecting the bin size of a time histogram," *Neural Comput.*, vol. 19, no. 6, pp. 1503–1527, Jun. 2007.



- [9] M. Goldstein and A. Dengel, "Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm," in *Proc. 35th German Conf. Artif. Intell.*, 2012, pp. 59–63.
- [10] K. Pearson, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, Jul. 1900.
- [11] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, vol. 162. Hoboken, NJ, USA: Wiley, 2009.
- [12] F. C. Porter, "Testing consistency of two histograms," 2008, *arXiv:0804.0380*. [Online]. Available: <http://arxiv.org/abs/0804.0380>
- [13] S. Bityukov, N. Krasnikov, A. Nikitenko, and V. Smirnova, "A method for statistical comparison of histograms," *Vestnik Rossijskogo Universiteta Družby Narodov. Serija, Matematika, Informatika, Fizika*, vol. 2, no. 2, pp. 324–330, 2014.
- [14] S. I. Bityukov, A. V. Maksimushkina, and V. V. Smirnova, "Comparison of histograms in physical research," *Nucl. Energy Technol.*, vol. 2, no. 2, pp. 108–113, Jun. 2016.
- [15] N. D. Gaganashvili, "Tests for comparing weighted histograms. Review and improvements," *Eur. Phys. J. Plus*, vol. 132, no. 5, p. 196, May 2017.
- [16] G. Caselli, J. Vallin, and G. Wunsch, *Demography: Analysis and Synthesis, Four Volume Set: A Treatise in Population*. Amsterdam, The Netherlands: Elsevier, 2005.
- [17] S. Mallat, *A Wavelet Tour Signal Processing: The Sparse Way*. Amsterdam, The Netherlands: Elsevier, 2008.
- [18] S. Heymann, M. Latapy, and C. Magnien, "Outskewer: Using skewness to spot outliers in samples and time series," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 527–534.
- [19] H. B. Ahmed, D. Dare, and A. O. Boudraa, "Graph signals classification using total variation and graph energy informations," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2017, pp. 667–671.
- [20] K. Gopalakrishnan, M. Z. Li, and H. Balakrishnan, "Identification of outliers in graph signals," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Dec. 2019, pp. 4769–4776.
- [21] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.
- [22] Y.-M. Huang, M. K. Ng, and Y.-W. Wen, "A new total variation method for multiplicative noise removal," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 20–40, Jan. 2009.
- [23] R. W. Liu, L. Shi, W. Huang, J. Xu, S. C. H. Yu, and D. Wang, "Generalized total variation-based MRI rician denoising model with spatially adaptive regularization parameters," *Magn. Reson. Imag.*, vol. 32, no. 6, pp. 702–720, Jul. 2014.
- [24] R. W. Liu, L. Shi, S. C. H. Yu, and D. Wang, "A two-step optimization approach for nonlocal total variation-based rician noise reduction in magnetic resonance images," *Med. Phys.*, vol. 42, no. 9, pp. 5167–5187, Aug. 2015.
- [25] L. I. Rudin and S. Osher, "Total variation based image restoration with free local constraints," in *Proc. 1st Int. Conf. Image Process.*, vol. 1, 1994, pp. 31–35.
- [26] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, Jan. 2005.
- [27] Y. Wang, J. Yang, W. Yin, and Y. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Imag. Sci.*, vol. 1, no. 3, pp. 248–272, Jan. 2008.
- [28] Z. Jia, M. K. Ng, and W. Wang, "Color image restoration by saturation-value total variation," *SIAM J. Imag. Sci.*, vol. 12, no. 2, pp. 972–1000, Jan. 2019.
- [29] H. A. Aly and E. Dubois, "Image up-sampling using total-variation regularization with a new observation model," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1647–1659, Oct. 2005.
- [30] M. K. Ng, H. Shen, E. Y. Lam, and L. Zhang, "A total variation regularization based super-resolution reconstruction algorithm for digital video," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, Dec. 2007, Art. no. 074585.
- [31] T. F. Chan, S. H. Kang, and J. Shen, "Total variation denoising and enhancement of color images based on the CB and HSV color models," *J. Vis. Commun. Image Represent.*, vol. 12, no. 4, pp. 422–435, Dec. 2001.
- [32] F. Pierre, J.-F. Aujol, A. Bugeau, G. Steidl, and V.-T. Ta, "Variational contrast enhancement of gray-scale and RGB images," *J. Math. Imag. Vis.*, vol. 57, no. 1, pp. 99–116, Jan. 2017.
- [33] C. Li, "An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing," Ph.D. dissertation, Dept. Comput. Appl. Math., Rice Univ., Houston, TX, USA, 2010.
- [34] X.-Z. Jian, R.-Z. Lu, Q. Guo, and G.-P. Wang, "Single image non-uniformity correction using compressive sensing," *Infr. Phys. Technol.*, vol. 76, pp. 360–364, May 2016.
- [35] I. Ihrke, X. Granier, G. Guennebaud, L. Jacques, and B. Goldluecke, "An introduction to optimization techniques in computer graphics," in *Proc. Eurographics Tuts.*, 2014. [Online]. Available: <https://diglib.org/handle/10.2312/egt.20141019.t9>, doi: [10.2312/egt.20141019](https://doi.org/10.2312/egt.20141019).
- [36] D. J. Garling, *Inequalities: A Journey Into Linear Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
- [37] B. Datta, *Numerical Methods for Linear Control Systems*, vol. 1. New York, NY, USA: Academic, 2004.
- [38] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [39] B. Everitt and A. Skrondal, *The Cambridge Dictionary of Statistics*, vol. 106. Cambridge, U.K.: Cambridge Univ. Press, 2002.
- [40] "Die Bevölkerung Nach Geburtsjahren und Familienstand am 17. Mai 1939," in *Statistisches Jahrbuch F r Das Deutsche Reich*. Berlin, Germany: Statistisches Reichsamt, 1939.
- [41] (1939). *Statistisches Jahrbuch für das Deutsche Reich*. [Online]. Available: [https://www.digizeitschriften.de/dms/toc/?PID=PPN514401303\\_1939](https://www.digizeitschriften.de/dms/toc/?PID=PPN514401303_1939)
- [42] (2020). *Holocaust Survivors and Victims Database—List of Individual Victims of Jasenovac Concentration Camp*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=45409](https://www.ushmm.org/online/hsv/source_view.php?SourceId=45409)
- [43] (2020). *Holocaust Survivors and Victims Database—Extraordinary Commission Data*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=20781](https://www.ushmm.org/online/hsv/source_view.php?SourceId=20781)
- [44] (2020). *Holocaust Survivors and Victims Database—Franzstrasse Nr. 38*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=38665](https://www.ushmm.org/online/hsv/source_view.php?SourceId=38665)
- [45] (2020). *Holocaust Survivors and Victims Database—[RG-75.002, Registration Cards of Jewish Refugees in Tashkent, Uzbekistan During WWII]*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=20492](https://www.ushmm.org/online/hsv/source_view.php?SourceId=20492)
- [46] H. Shryock, J. Siegel, E. Larmon, and U. S. B. of the Census, *The Methods and Materials of Demography* (The Methods and Materials of Demography), no. 1. Washington, DC, USA: U.S. Department of Commerce, Bureau of the Census, 1980.
- [47] J. Siegel, D. Swanson, and H. Shryock, *The Methods and Materials of Demography*. Amsterdam, The Netherlands: Elsevier, 2004.
- [48] (2020). *Holocaust Survivors and Victims Database—Search for Lists*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_advance\\_search.php](https://www.ushmm.org/online/hsv/source_advance_search.php)
- [49] Z. Wang, Y. Zeng, B. Jeune, and J. W. Vaupel, "Age validation of Han Chinese centenarians," *Genus*, vol. 54, pp. 123–141, Jan. 1998.
- [50] T. Spoorenberg, "Is the Whipple's index really a fair and reliable measure of the quality of age reporting? An analysis of 234 censuses from 145 countries," in *Proc. 24th Int. Population Conf.*, Marrakesh, Morocco, 2009. [Online]. Available: <https://iussp2009.princeton.edu/papers/90551>
- [51] (2020). *Holocaust Survivors and Victims Database—[Index to the Visual History Archive of Holocaust oral Testimonies from the USC Shoah Foundation Institute]*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=25016](https://www.ushmm.org/online/hsv/source_view.php?SourceId=25016)
- [52] (2020). *Holocaust Survivors and Victims Database—Stanko Nick*. [Online]. Available: [https://www.ushmm.org/online/hsv/person\\_view.php?PersonId=5003266](https://www.ushmm.org/online/hsv/person_view.php?PersonId=5003266)
- [53] (2020). *Holocaust Survivors and Victims Database—STANKO NICK*. [Online]. Available: [https://www.ushmm.org/online/hsv/person\\_view.php?PersonId=7554720](https://www.ushmm.org/online/hsv/person_view.php?PersonId=7554720)
- [54] (2020). *Stanko Nick—Wikipedia*. [Online]. Available: [https://en.wikipedia.org/wiki/Stanko\\_Nick](https://en.wikipedia.org/wiki/Stanko_Nick)
- [55] (2020). *Holocaust Survivors and Victims Database—[Prisoner Registration Forms from Auschwitz]*. [Online]. Available: [https://www.ushmm.org/online/hsv/source\\_view.php?SourceId=21303](https://www.ushmm.org/online/hsv/source_view.php?SourceId=21303)
- [56] (2020). *Holocaust Survivors and Victims Database—Josip Stern*. [Online]. Available: [https://www.ushmm.org/online/hsv/person\\_view.php?PersonId=4891947](https://www.ushmm.org/online/hsv/person_view.php?PersonId=4891947)
- [57] (2020). *Holocaust Survivors and Victims Database—JOSIP STERN*. [Online]. Available: [https://www.ushmm.org/online/hsv/person\\_view.php?PersonId=7527187](https://www.ushmm.org/online/hsv/person_view.php?PersonId=7527187)
- [58] (2020). *Holocaust Survivors and Victims Database—STOJAN RAŽOKRAK*. [Online]. Available: [https://www.ushmm.org/online/hsv/person\\_view.php?PersonId=7546608](https://www.ushmm.org/online/hsv/person_view.php?PersonId=7546608)



**NIKOLA BANIĆ** received the B.Sc., M.Sc., and Ph.D. degrees in computer science in 2011, 2013, and 2016, respectively. He is currently working as a Senior Computer Vision Engineer with Gideon Brothers, Croatia. He has worked in real-time image enhancement for embedded systems, digital signature recognition, people tracking and counting, and image processing for stereo vision. His research interests include image enhancement, color constancy, image processing for stereo vision, and tone mapping.



**NEVEN ELEZOVIĆ** received the B.Sc., M.Sc., and Ph.D. degrees in mathematics in 1977, 1981, and 1985, respectively. He is currently a Full Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb, and the Founder of the Publishing House Element. He has published numerous publications in high-ranking journals and books for college mathematics. His research interests include mathematical analysis and probability theory.

...