stats

MDPI

Article

Empirical Inferences Under Bayesian Framework to Identify Cellwise Outliers

Luca Sartore [1,2,*], Lu Chen [1,2] and Valbona Bejleri [2]

[1] National Institute of Statistical Sciences, P.O. Box 33762, Washington, DC, 20033, USA; lchen@niss.org
[2] United States Department of Agriculture, National Agriculture Statistics Service, Washington, DC, 20250, USA; valbona.bejleri@usda.gov
* Correspondence: lsartore@niss.org

Abstract: Outliers are typically identified using frequentist methods. The data are classified as "outliers" or "not outliers" based on a test statistic that measures the magnitude of the difference between a value and the majority part of the data. The threshold for a data value to be an outlier is typically defined by the user. However, a subjective choice of the threshold increases the uncertainty associated with outlier status for each data value. A cellwise outlier detection algorithm named FuzzyHRT is used to automate the editing process in repeated surveys. This algorithm uses Bienaymé–Chebyshev's inequality and fuzzy logic to detect four different types of outliers resulting from format inconsistencies, historical, tail, and relational anomalies. However, fuzzy logic is not suited for probabilistic reasoning behind the identification of anomalous cells. Bayesian methods are well suited for quantifying the uncertainty associated with the identification of outliers. Although, as suggested by the literature, there exist well-developed Bayesian methods for record-level outlier detection, Bayesian methods for identifying outliers within individual records (i.e., at the cell level) remain unexplored. This paper presents two approaches from the Bayesian perspective to study the uncertainty associated with identifying outliers. A Bayesian bootstrap approach is explored to study the uncertainty associated with the output scores from the FuzzyHRT algorithm. Empirical likelihoods in a Bayesian setting are also considered for probabilistic reasoning behind the identification of anomalous cells. NASS survey data for livestock and major crop yield (such as corn) are considered for comparing the performances of the two proposed approaches with recent cellwise outlier methods.

Keywords: anomaly identification; Bayesian bootstrap; empirical likelihood; fuzzy logic; predictive distribution; uncertainty

Citation: Sartore, L.; Chen, L.; Bejleri, V. Empirical Inferences Under Bayesian Framework to Identify Cellwise Outliers. *Stats* **2024**, *7*, 1244–1258. https://doi.org/10.3390/stats7040073

Academic Editor: Wei Zhu and Wenhao Gui
Received: 14 August 2024
Revised: 16 October 2024
Accepted: 18 October 2024
Published: 19 October 2024

1. Introduction

Agricultural data acquired through surveys are inherently complex, often characterized by skewed distributions, multivariate relationships and spatio-temporal dynamics. This complexity in the nature of agricultural data along with imperfect data collection processes can lead to anomalies at the entry (cell) level data. The presence of anomalous values in a dataset can affect the analyses based on these data, e.g., introducing bias to the modeled estimates or forecasts. The United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) has implemented a semi-automated revision process to improve the accuracy of the data acquired through surveys. The current semi-automated system detects cellwise anomalies (Figure 1, graph on the right) that are subsequently corrected manually. Even though the system uses automated decision rules based on if-else conditions designed by experts in agriculture, it still requires human intervention at several levels. Recently, NASS has been investigating alternative approaches to modernize its anomaly detection system and is extending its traditional editing techniques [1] with novel, more accurate, flexible and objective methodologies.

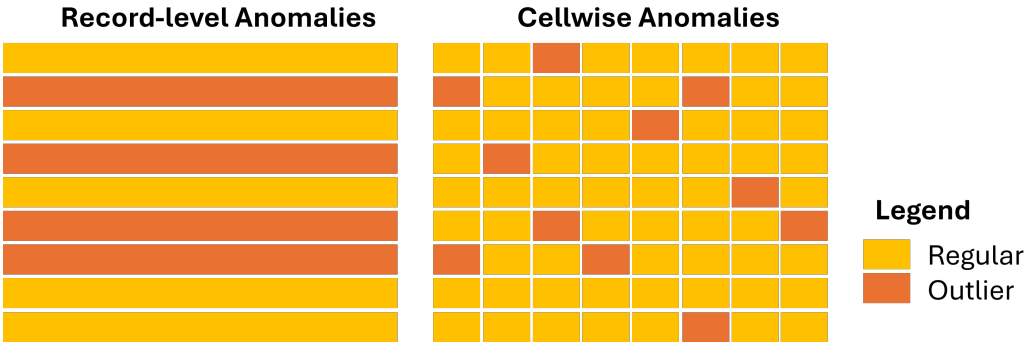

Figure 1. Graphical illustration of two datasets with anomalies. On the left, anomalies at the record-level are represented by rows colored in red, and on the right, anomalies at the data-entry level (that could occur one or several times within a multivariate record) are represented by cells colored in red.

Historically, researchers have addressed outliers from two perspectives. From the first perspective, an outlier is typically defined as an observation generated by a mechanism different from the one that produced the majority of observations in a dataset. While this assumption was previously used by [2–4] to specify a model for generating outliers, it was Freeman who presented in 1980 the definition of outlier as "an observation that has not been generated by the mechanism that generated the majority of observations in the dataset" [5]. Alternatively, from the second perspective, all data are considered observations generated by a single mechanism. A (statistical) model is assumed to fit the observed data and the outliers arise from the model outputs. From this perspective, the "outlier" status of the data is typically determined through inferences on standardized residuals under the assumed model. After the records are investigated for influential points, they are classified either as outliers or regular cases based on a score variable. However, algorithms for detecting cellwise outliers (i.e., anomalous cells within a record) have been overlooked in the literature from both perspectives until recent developments [6–8].

In addition, primarily focusing on record-level anomalies, most outlier detection methods are typically developed under the frequentist framework. A variety of algorithms developed under the frequentist framework generate scores for each individual record and use these scores to detect anomalies at the record-level [9–13]. The Detect-Deviating-Cells (DDC) method [8] is known as the first method developed to detect cellwise outliers in multivariate datasets by accounting for the correlations among variables. This method assumes normality but does not consider stratification or historical information. Due to the advances in computational power and tools, more meaningful criteria for outlier detection in complex multivariate data situations have been developed. For example, it is more intuitive to apply a degree of belief that a cell in a multidimensional dataset is an outlier rather than simply classifying that cell as an outlier. This could be achieved through Bayesian methods where similar to the frequentist framework, outliers are typically identified by analyzing random errors under a given probabilistic model. However, under the Bayesian paradigm, outliers are detected through the analysis of posterior distributions.

The interest in Bayesian procedures for outlier detection has increased in recent decades. A review of the probabilistic methods for record-level outlier detection in a Bayesian setting under a linear model is given in [14]. The authors classify these methods into two groups: (1) methods that investigate the predictive density of the response variable and (2) methods that investigate the posterior (predictive) probabilities of unobserved residuals. In [15], the outlier scores are transformed into probabilities, and two approaches are presented. The first considers a logistic sigmoid as a posterior distribution and estimates its parameters using the outlier score data. The second derives the posterior probabilities by assuming a mixture of exponential and Gaussian distributions for the score data. For completeness, we also mention the following Bayesian contributions on record-level outlier detection. Considering a univariate linear model in [16], the authors propose using the posterior distribution of the squared norm of the realized errors to identify anomalous

records. Later, in [17], the approach is extended to a multivariate linear model using Bayes factors to detect outliers. In [18], general measures (i.e., Kullback–Leibler (KL), chi-square, L1-divergence) are used to study, via simulations, how individual observations can affect the distribution of the response given the covariates. The authors found that KL and chi-square measures were monotonic on L1 divergence. They recommended L1 divergence as a diagnostic measure due to the easiness of its interpretation [18]. Geisser's work on the predictive density of one observation given the rest of the data is worth mentioning here as well [19–22]. However, none of these previous works has addressed the Bayesian identification of anomalous data entries within a record (i.e., cellwise outliers).

Despite the extensive literature on outlier detection, publications for cellwise outlier detection are quite sparse [6–8]. While most of the algorithms are developed to detect an anomaly at a record level, classical record-level assumptions are seldom satisfied when an anomaly in data collected through surveys occurs at one (or several) component(s) of a record (as illustrated in Figure 1). Also, most of the existing algorithms proposed for identifying cellwise outliers are often designed without considering characteristics of the data, i.e., missing values, skewed marginal distributions, and spatio-temporal dynamics. Sartore et al. [23] introduced an outlier detection approach based on robust estimation methods and fuzzy logic. This approach is applicable to sparse or stratified datasets and can leverage additional information when historical data are available. Implementation of the FuzzyHRT algorithm has automated the editing process for repeated surveys at NASS.

The FuzzyHRT algorithm uses the Bienaymé–Chebyshev's inequality [24] which is suitable for nonparametric and empirical inferences. The only assumption made about the distribution of the data is that the first and second moments exist, which could be estimated empirically. Then, for a random variable X with distribution function $F_X(x)$ and a location parameter estimator

$$\hat{\mu}_\delta = \arg\min_{\mu \in \mathbb{R}} \int |X - \mu|^\delta \mathrm{d}F_X,$$

the Bienaymé–Chebyshev's inequality is defined for any $\delta \geq 1$ and $\varepsilon \in \mathbb{R}$ as

$$\Pr(|X - \hat{\mu}_\delta| \geq |\varepsilon|) \leq \min\left\{1, \frac{\mathsf{E}\left[|X - \hat{\mu}_\delta|^\delta\right]}{|\varepsilon|^\delta}\right\}. \tag{1}$$

In the special case of $\delta = 2$, Equation (1) produces the classical Chebyshev's inequality,

$$\Pr(|X - \hat{\mu}_2| \geq |\varepsilon|) \leq \min\left\{1, \frac{\mathsf{Var}[X]}{|\varepsilon|^2}\right\},$$

where $\hat{\mu}_2 = \mathsf{E}[X]$. Fuzzy logic is successively applied to detect four different types of cellwise outliers resulting from format inconsistencies and historical, tail, and relational anomalies. Nonetheless, fuzzy logic is not suited for the probabilistic reasoning behind the identification of anomalous cells. Therefore, Bayesian methods are preferred as better suited methods for quantifying the uncertainty associated with the identification of cellwise outliers. In this paper, two novel Bayesian methods for detecting outliers at the cell level are proposed. Both methods are based on the theory of empirical likelihoods [25], which provides a nonparametric framework for inference without making assumptions on the distribution of the data. First, a Bayesian bootstrap approach (hereafter referred to as bootstrap approach) is proposed to mitigate the uncertainty associated with the output scores from the FuzzyHRT algorithm [23]. Second, empirical likelihood under a Bayesian framework [26] (hereafter referred to as the empirical likelihood approach) provides a probabilistic reasoning for identifying outliers at the cell level.

The rest of the paper is structured as follows. In Section 2, we give a brief review of the FuzzyHRT algorithm and set up the problem. The bootstrap approach is presented in Section 3. The empirical likelihood approach is introduced in Section 4. Results from a controlled simulation study using NASS survey data for livestock and crops to assess the

performances of the two proposed approaches are presented in Section 5. Our concluding remarks are presented in Section 6.

2. Materials and Methods

2.1. Brief Background on the FuzzyHRT Algorithm

The FuzzyHRT algorithm [23] identifies four types of anomalies occurring at a data entry (or cell), each resulting in a specific type of outlier. The first anomaly type consists of erroneous data with format inconsistencies (B), also known as bit-flip errors. The second type refers to historical anomalies (H), i.e., data entries that are large deviations from previously reported data. The third type is traditionally known as a distribution-tail anomaly (T), i.e., univariate outliers. The fourth type of cell anomaly detected by the FuzzyHRT algorithm involves the breaking of linear relationships (R) among multiple variables, i.e., deviations from typical multivariate relationships. In this paper, we concentrate on historical, tail, and relational anomalies and disregard bit-flip errors. In fact, these errors are often corrected automatically by modern hardware [27,28], and thus, they are quite rare in practice.

It is possible that a historical anomaly can be a tail anomaly, or vice versa. It is also possible that a relational anomaly can be also a tail or historical, or both tail and historical anomaly. However, while an anomaly could be of historical, tail, relational, or a combination of the three, each type of anomaly is identified using a specific approach. The FuzzyHRT algorithm utilizes regression-like methodologies including time series, linear regression, and median estimation procedures at the cell level to produce standardized residuals associated with historical (H), tail (T), and relational (R) anomalies for each cell (i, j). First, in [23], an $\mathsf{ARIMA}(1,0,0)$ provides a prediction $\hat{x}_{ij}^{[H]}$ that minimizes the mean absolute error (MAE) with respect to x_{ij}. However, the authors also discussed extensions based on more sophisticated models that are better suited for longitudinal analyses or other types of temporal data. Second, the stratum median of each variable j is used to compute prediction $\hat{x}_{ij}^{[T]}$. This prediction is constant within a stratum but can change across the sample. Third, a linear regression model is fitted to compute a prediction $\hat{x}_{ij}^{[R]}$ using the other ($\ell \neq j$) normalized variables as predictors. Sartore et al. [23] also discusses the use of monotone link functions to transform skewed data before performing regression analyses. In this paper, residuals are denoted by

$$\varepsilon_{ij}^{[k]} = x_{ij} - \hat{x}_{ij}^{[k]},$$

where $k \in \{H, T, R\}$ for historical, tail, and relational type anomalies.

The Chebyshev's inequality and its robust extension, i.e., Bienaymé–Chebyshev's inequality, is used in [23] to score the standardized residuals without imposing distributional assumptions. In particular, the inequality for a random variable $X_{ij} - \hat{x}_{ij}^{[k]}$ is expressed in terms of a transformed realization of errors, $\varepsilon_{ij}^{[k]}$,

$$\Pr\left(\left|X_{ij} - \hat{x}_{ij}^{[k]}\right| > g\left(\varepsilon_{ij}^{[k]}\right)\right) < \min\left\{1,\ g\left(\varepsilon_{ij}^{[k]}\right)^{-\delta} \mathsf{E}\left[\left|X_{ij} - \hat{x}_{ij}^{[k]}\right|^{\delta}\right]\right\},$$

where $g : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$ is a generic function of $\varepsilon_{ij}^{[k]}$ with nonnegative codomain, and the scalar $\delta \geq 1$ represents the order of the absolute moment. In [23], $\delta = 1$ when the least absolute residuals are used for computing the predictions, and $\delta = 2$ when the least squared errors are used instead. For the specific choice of $g(\varepsilon) = |\varepsilon|$, the score $s_{ij}^{[k]}$ is defined as

$$s_{ij}^{[k]} := \min\left\{1,\ \left|\varepsilon_{ij}^{[k]}\right|^{-1} \mathsf{E}\left[\left|X_{ij} - \hat{x}_{ij}^{[k]}\right|^{\delta}\right]^{1/\delta}\right\}.$$

In this paper, the three separate scores are computed based on the standardized residuals, i.e., $s_{ij}^{[\mathrm{H}]}$ for historical anomalies, $s_{ij}^{[\mathrm{T}]}$ for distribution-tail anomalies, and $s_{ij}^{[\mathrm{R}]}$ for relational anomalies.

The scores under consideration (i.e., associated with the historical, tail, and relational anomalies) are combined into one final score. The concept of triangular norms (or t-norms) from the fuzzy logic literature is used to derive the matrix of final scores,

$$\mathbf{S}^* = \{s_{ij}^*, \forall i = 1, \ldots, n, \forall j = 1, \ldots, p\}, \tag{2}$$

where s_{ij}^* is the score corresponding to the data entry for variable j in record i and takes values in $[0, 1]$. In general, each final score summarizes the evidence for a data entry to be regular (and thus, the larger the final scores the more regular the data entry is).

One can classify the observed data values as outliers or not using the 100θ empirical percentile of the final scores s_{ij}^*, $\forall i = 1, \ldots, n$, $\forall j = 1, \ldots, p$ produced through the FuzzyHRT algorithm as a threshold; θ is a user-provided level of contamination in the observed dataset $\mathbf{X}$. The subjective choice of the data-contamination level, $\theta \in (0, 1)$, increases the uncertainty associated with the status of data entries to be either regular or outliers. Typically, a degree of uncertainty is associated with each individual score, which contributes through a fuzzy logic process (used by the FuzzyHRT algorithm) to the uncertainty associated with the final score s_{ij}^*. These uncertainties further translate into the uncertainty associated with the outlier status of its respective value x_{ij}.

Remark 1. *T-norm functions*

A t-norm function satisfies the commutative, monotonic, associative, and identity properties. A general t-norm function maps the unit square in a Cartesian plane to the closed unit interval $[0, 1]$ *in* $\mathbb{R}$*, i.e.,* $d : [0, 1] \times [0, 1] \to [0, 1]$*. T-norm functions are used recursively to combine the three anomaly scores into a final one. In this paper, we use two t-norm functions, the product t-norm denoted by* d_1 *(i.e.,* $d = d_1$*) and the minimum t-norm denoted by* d_2 *(i.e.,* $d = d_2$*). FuzzyHRT uses the product t-norm function,*

$$d_1(z_1, z_2) = z_1 z_2, \ \forall z_1, z_2 \in [0, 1].$$

The Bayesian approach based on the empirical likelihood presented in Section 4 uses the minimum t-norm function,

$$d_2(z_1, z_2) = \min(z_1, z_2), \ \forall z_1, z_2 \in [0, 1].$$

The final score s_{ij}^* *for the* $(i, j)^{th}$ *entry is computed based on historical, relational and tail anomaly scores as*

$$s_{ij}^* = d\left(s_{ij}^{[\mathrm{H}]}, d\left(s_{ij}^{[\mathrm{R}]}, s_{ij}^{[\mathrm{T}]}\right)\right),$$

where $d = d_1$ *for the FuzzyHRT algorithm and* $d = d_2$ *for the empirical likelihood approach. More details on fuzzy logic and probabilistic inequalities are given in Appendix A of [23].*

2.2. Problem Setup

Let $\mathbf{X}$ denote an observed dataset with outliers, consisting of n records and p variables, and let $\mathbf{s}^* = \mathrm{vec}(\mathbf{S}^*)$, where $\mathbf{S}^*$ is defined in (2) as the matrix of scores produced from $\mathbf{X}$ using the FuzzyHRT algorithm [23]. Let $\hat{Q}_\theta$ denote the 100θ empirical percentile of the scores $\mathbf{s}^*$, where θ is a user-provided level of contamination in $\mathbf{X}$. $\hat{Q}_\theta$ partitions the scores $\mathbf{s}^*$ into two groups corresponding to regular data and outliers, and it is used as a cut-off when determining the outlier status of a cell x_{ij} in $\mathbf{X}$. A subjective choice of a contamination level θ increases the uncertainty associated with the identification of cellwise outliers. Two Bayesian approaches are considered to mitigate this uncertainty.

The goal is to classify the data entry, x_{ij}, into regular data or outlier and to study the uncertainty associated with this process. First, a bootstrap approach is used to obtain the conditional distribution of $\hat{Q}_\theta$ given the scores. Second, the empirical likelihood approach at

the cell level is considered under the Bayesian perspective for probabilistic reasoning behind the identification of anomalous cells. In the second approach, the unknown contamination parameter $\theta \in (0,1)$ is treated as a random variable. Furthermore, in the latter approach, posterior inferences are conducted based on the residuals from the original data at each individual cell, x_{ij}.

3. A Bayesian Bootstrap Approach

Let H_{ij} denote an indicator random variable for an outlier cell (i,j),

$$\begin{aligned} H_{ij} = 1 &\iff x_{ij} \text{ is outlier} \\ H_{ij} = 0 &\iff x_{ij} \text{ is regular.} \end{aligned} \tag{3}$$

In this section, we focus on inferences about the empirical distribution of the scores produced by FuzzyHRT, $\mathbf{S}^*$, and explore the uncertainty associated with their 100θ empirical percentile $\hat{Q}_\theta$. Hereafter, in this section, the scores $\mathbf{S}^*$ are thought of as a vector $\mathbf{s}^* = \text{vec}(\mathbf{S}^*)$ with length $N = n \times p$. Then, conditions in (3) for partitioning the data take the form,

$$\begin{aligned} H_{ij} = 1 &\iff 0 \le s^*_{ij} \le \hat{Q}_\theta; \\ H_{ij} = 0 &\iff \hat{Q}_\theta \le s^*_{ij} \le 1. \end{aligned} \tag{4}$$

To obtain the empirical distribution of $\mathbf{s}^*$, let $\boldsymbol{\omega}$ denote a parameter vector of length M whose components are probabilities associated with M distinct values of the scores $\mathbf{s}^*$,

$$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)^\top \in \mathbb{S}_{M-1},$$

where $M \le N$ and $\mathbb{S}_{M-1}$ denotes the $M-1$ dimensional set of unit probability simplex for $\mathbf{s}^*$ [25]. These probabilities satisfy the two following constraints:

$$\omega_m \ge 0, \ \forall m = 1, \ldots, M, \text{ and } \sum_{m=1}^{M} \omega_m = 1.$$

A noninformative Dirichlet prior is selected for $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)^\top$,

$$\boldsymbol{\omega} \sim \mathbb{1}_{\mathbb{S}_{M-1}}(\boldsymbol{\omega}) \times \prod_{m=1}^{M} (\omega_m)^0 \times c = \mathbb{1}_{\mathbb{S}_{M-1}}(\boldsymbol{\omega}) \times c, \tag{5}$$

where c is the normalizing constant. Assuming that the $n_m \in \{0, 1, \ldots, N\}$ scores produced from FuzzyHRT are associated with their respective probability ω_m, for $m = 1, \ldots, M$, one can write the likelihood as

$$\mathcal{L}(\boldsymbol{\omega}) = \Pr(\mathbf{s}^* | \boldsymbol{\omega}) \propto \prod_{m=1}^{M} (\omega_m)^{n_m}.$$

Incorporating the information provided from the final scores $\mathbf{s}^*$ defined in (2), we obtain the posterior distribution of $\boldsymbol{\omega}$ given $\mathbf{s}^*$,

$$\boldsymbol{\omega} | \mathbf{s}^* \sim \mathbb{1}_{\mathbb{S}_{M-1}}(\boldsymbol{\omega}) \times \prod_{m=1}^{M} (\omega_m)^{n_m} \times c, \tag{6}$$

which is a Dirichlet$(n_1, \ldots, n_M)$.

Remark 2. *If* $n_m = 1, \forall m \in \{1, \ldots, M\}$, *then* $M = N$.

The Bayesian bootstrap for the 100θ percentile is based on a resampling scheme performed B times as follows. To obtain a sample for the random variable $\boldsymbol{\omega}|\mathbf{s}^*$ based on

its empirical posterior distribution, first, a sample of size M is drawn from the uniform distribution with support in $(0,1)$, i.e., $u_m \sim \text{Uniform}(0,1)$, for any $m = 1,\ldots,M$. Then, for all $m \in \{1,\ldots,M\}$, the $\omega_m|\mathbf{s}^*$ is computed,

$$\omega_m|\mathbf{s}^* = \frac{\log(u_m)}{\sum_{h=1}^{M}\log(u_h)}, \tag{7}$$

resulting in a realized sample $(\omega_1|\mathbf{s}^*,\ldots,\omega_M|\mathbf{s}^*)$. This sample provides probabilities for the vector of final scores $\mathbf{s}^*$. For practical reasons, without loss of generality, one can assume that the condition of Remark 2 is true, i.e., $n_m = 1$ (and hence, $M = N$), to construct a Monte-Carlo sample $\boldsymbol{\omega}|\mathbf{s}^*$ of size N.

After sorting the scores, the cumulative distribution of $\mathbf{s}^*|\boldsymbol{\omega}$ is empirically evaluated at $s^*_{(h)}$, for all $h = 1,\ldots,N$, as the sum of the posterior probabilities in (7) that are associated with all scores $s^*_{(m)}$ such that $m \leq h$,

$$\hat{F}_{\mathbf{s}^*|\boldsymbol{\omega}}(s^*_{(h)}|\boldsymbol{\omega}) = \sum_{m\leq h}\omega_m|\mathbf{s}^*.$$

Then, the value of $\hat{Q}_\theta^{(b)}$, for all $b \in \{1,\ldots,B\}$, is obtained by solving the following equation:

$$\hat{F}_{\mathbf{s}^*|\boldsymbol{\omega}}\left(\hat{Q}_\theta^{(b)}|\boldsymbol{\omega}\right) = \theta,$$

or similarly

$$\hat{Q}_\theta^{(b)} = \hat{F}^{-1}_{\mathbf{s}^*|\boldsymbol{\omega}}(\theta). \tag{8}$$

This results in a sample of size $B = 1000$ constructed using the bootstrapped statistics in (8). The mean (and other summary statistics) of the $\hat{Q}_\theta^{(b)}$ were considered when determining the outliers based on the scores produced by the FuzzyHRT algorithm.

4. A Bayesian Testing Approach Based on Empirical Likelihoods

In this approach, empirical likelihoods at the cell level are used to study the posterior distribution of the indicator random variable H_{ij} defined in (3) given the data x_{ij}. H_{ij} describes the outlier status of x_{ij} (i.e., cell (i,j)), where $j = 1,\ldots,p$ and $i = 1,\ldots,n$. Specifically, we construct the distribution of the random variable $H_{ij}|\boldsymbol{\varepsilon}_{ij}$, where

$$\boldsymbol{\varepsilon}_{ij} = \left(\varepsilon_{ij}^{[\mathrm{H}]},\varepsilon_{ij}^{[\mathrm{R}]},\varepsilon_{ij}^{[\mathrm{T}]}\right)^\top \tag{9}$$

and $\varepsilon_{ij}^{[k]}$ are standardized residuals provided by the FuzzyHRT algorithm for each anomaly type $k \in \{H,R,T\}$. The distribution of $\boldsymbol{\varepsilon}_{ij}|\theta,H_{ij}$ is used to evaluate $\Pr(H_{ij}|\boldsymbol{\varepsilon}_{ij})$.

In this setting, different from the bootstrap approach, the level $\theta \in (0,1)$ of data contamination is unknown and treated as a random variable. It is reasonable to specify the conditional distribution for θ given H_{ij},

$$\theta|H_{ij} \sim \text{Uniform}\left(\frac{1-H_{ij}}{2},\frac{2-H_{ij}}{2}\right). \tag{10}$$

This specific choice of distribution for $\theta|H_{ij}$ makes it possible to simplify the successive calculations of the posterior in (12). A noninformative (Bernoulli) hyperprior is adopted for the outlier status random variable H_{ij},

$$\Pr(H_{ij}=0) = \Pr(H_{ij}=1) = 0.5, \tag{11}$$

at each cell (i,j).

One can easily obtain the posterior distribution of $H_{ij}|\boldsymbol{\varepsilon}_{ij}$,

$$\Pr(H_{ij}|\boldsymbol{\varepsilon_{ij}}) \propto \Pr(H_{ij}) \int \Pr(\boldsymbol{\varepsilon_{ij}}|\theta, H_{ij}) \Pr(\theta|H_{ij}) \mathrm{d}\theta, \tag{12}$$

where θ is integrated out. Incorporating the hyperprior (11) and Equation (10) in (12), one obtains the conditional distribution of $H_{ij}|\boldsymbol{\varepsilon}_{ij}$,

$$\Pr(H_{ij}|\boldsymbol{\varepsilon}_{ij}) \propto \int_{(1-H_{ij})/2}^{(2-H_{ij})/2} \Pr(\boldsymbol{\varepsilon}_{ij}|\theta, H_{ij}) \mathrm{d}\theta, \tag{13}$$

where the vector $\boldsymbol{\varepsilon}_{ij}$ is defined in (9). This distribution is of an unknown form. Therefore, $\Pr(\boldsymbol{\varepsilon}_{ij}|\theta, H_{ij})$ is evaluated using the empirical likelihood [25] defined as follows:

$$\begin{aligned} \mathcal{L}(\theta) := \Pr(\boldsymbol{\varepsilon}_{ij}|\theta, H_{ij}) = \max_{w_{ij} \in \mathbb{R}^3} \Bigg\{ 3^3 \prod_{k \in \{\mathrm{H,R,T}\}} w_{ij}^{[k]} : & \sum_{k \in \{\mathrm{H,R,T}\}} w_{ij}^{[k]} q\left(\varepsilon_{ij}^{[k]}, \theta\right) = 0, \\ & \theta \in \left[\frac{1-H_{ij}}{2}, \frac{2-H_{ij}}{2}\right], \\ & w_{ij}^{[k]} \geq 0, \ \forall k \in \{\mathrm{H,R,T}\}, \\ & \sum_{k \in \{\mathrm{H,R,T}\}} w_{ij}^{[k]} = 1 \Bigg\}, \end{aligned} \tag{14}$$

where the terms $q\left(\varepsilon_{ij}^{[k]}, \theta\right)$, for $k \in \{\mathrm{H,R,T}\}$, $i = 1, \ldots, n$, and $j = 1, \ldots, p$, control the shape of the estimating equation.

Because the empirical likelihood in (14) does not provide a closed-form analytical expression for any given values of θ, its evaluation is performed by solving an optimization problem. However, the direct optimization of (14) is computationally challenging, and hence, the theory of duality [29,30] is used to simplify the optimization problem. Therefore, the quantification of $\Pr(\boldsymbol{\varepsilon}_{ij}|\theta, H_{ij})$ is achieved by maximizing the following function:

$$G(w_{ij}, \lambda_1, \lambda_2) = \sum_{k \in \{\mathrm{H,R,T}\}} \log w_{ij}^{[k]} + \lambda_1 - \lambda_1 \sum_{k \in \{\mathrm{H,R,T}\}} w_{ij}^{[k]} + \lambda_2 \log \theta - \lambda_2 w_{ij}^{\left[\hat{\kappa}_{ij}\right]} \log \hat{\zeta}_{ij}^{[\hat{\kappa}_{ij}]} \tag{15}$$

with respect to $w_{ij} \in \mathbb{R}^3$, where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the Lagrange multipliers. The optimization is achieved under the assumption that

$$\sum_{k \in \{\mathrm{H,R,T}\}} w_{ij}^{[k]} q\left(\varepsilon_{ij}^{[k]}, \theta\right) = w_{ij}^{\left[\hat{\kappa}_{ij}\right]} \log \hat{\zeta}_{ij}^{[\hat{\kappa}_{ij}]} - \log \theta,$$

where

$$\hat{\kappa}_{ij} = \arg \min_{k \in \{\mathrm{H,R,T}\}} \log \hat{\zeta}_{ij}^{[k]},$$

and

$$\hat{\zeta}_{ij}^{[k]} = \frac{\Phi^{-1}(0.995)}{\Phi^{-1}(0.995) + \left|\varepsilon_{ij}^{[k]}\right|},$$

for any $k \in \{\mathrm{H,R,T}\}$, where the notation $\Phi^{-1}(0.995)$ denotes the 99.5% quantile from a cumulative distribution function from the standard normal. By setting the first derivative

of (15) with respect $w_{ij}^{[k]}$ to zero, one obtains a closed-form solution of the weights as a function on the Lagrange multipliers:

$$\hat{w}_{ij}^{[k]} = \frac{1}{\lambda_1 - \lambda_2 \log\theta + \lambda_2 w_{ij}^{[\hat{\kappa}_{ij}]} \log \hat{\zeta}_{ij}^{[\hat{\kappa}_{ij}]}}, \ \forall k \in \{\mathrm{H}, \mathrm{R}, \mathrm{T}\}.$$

The iterative algorithm to identify the optimal value of an empirical likelihood for a given θ, uses $\lambda_1 = 3$, and $\lambda_2 = 0$ as initial values. After the weights are updated, the Lagrange multipliers are updated using the Gauss–Newton method:

$$\lambda_1 \leftarrow \lambda_1 - 1 + \sum_{k \in \{H,R,T\}} w_k$$
$$\lambda_2 \leftarrow \lambda_2 - \log\theta + w_{ij}^{[\hat{\kappa}_{ij}]} \log \hat{\zeta}_{ij}^{[\hat{\kappa}_{ij}]}.$$

This iterative algorithm stops when either convergence is reached or a maximum number of iterations has been performed. Finally, the likelihood is computed based on (14),

$$\Pr(\boldsymbol{\varepsilon}_{ij} | \theta, H_{ij}) = 3^3 \prod_{k \in \{\mathrm{H},\mathrm{R},\mathrm{T}\}} \hat{w}_{ij}^{[k]}.$$

The integral in (13) is evaluated using classical quadrature methods. In this paper, the definition of Riemann's integral is used to compute the posterior probabilities.

5. Simulation Study

As in [23], a controlled simulation study using four national surveys administered by NASS is conducted to assess the performances of the proposed outlier detection approaches. These national surveys provide a wide range of different agricultural scenarios (see Table 1 for a short summary). The first two surveys have been conducted for sheep-and-goat and cattle inventories, and the last two on row-crop yields and cranberry production. Usually, livestock surveys focus on the herd composition, while crop surveys collect information on production and yields. Surveys under consideration were administered between 2021 and 2022, and have sample sizes ranging between 218 and 21,154 and a number of nonnegative continuous variables ranging between 3 and 49.

Table 1. Description of surveys used to evaluate the proposed methodology.

Survey	Number of Resposes	Number of Variables	Major Inquiries	Scenario
Sheep and Goat Inventory	10,090	49	Sheep and/or goat herd composition (ewes, rams, lambs, billies, nannies, kids, etc.)	Many records; two distinct inventoriess of aggregated part
Cattle Inventory	21,154	18	Cattle herd composition (cows, bulls, calves, etc.)	Many records; one inventory of aggregated parts
Agricultural Yield	1762	54	Expected yield and acres of small grain crops (i.e., barley, wheat, oats)	Fewer records; multiple crops
Cranberry Production	218	3	Cranberry acres	Very few records; single crop

Because the four datasets shown in Table 1 have been manually edited, there are no ground-truth labels on the anomaly status of each cell. Thus, cellwise outliers have

been synthetically introduced in the four datasets to test the robustness of the proposed algorithms across various scenarios. These scenarios have been developed to include different types of outliers to better reflect the anomalies observed in the raw data. This approach allowed us to flag and track anomalous values when evaluating the proposed detection algorithms across a wide variety of potential data conditions.

The generative algorithm used in [23] has been applied to each dataset in Table 1 by randomly replacing a few cells with anomalous values. Three specific modules synthesizing historical, tail, and "relational" anomalies, respectively, are created. Each module randomly selects 5% of the item responses. Half of these are replaced by multiplying the current values by random factors in $(0,1)$, and the other half using random factors in $(1,3)$. The random factors are generated from uniform distributions in intervals shown in Table 2 (more specifically in columns 2 and 3). Shrinking and expansion ranges are randomly selected with equal probability for each combination of anomaly type and dissimilarity level. Therefore, two distinct datasets have been created for each survey. The datasets marked as "high" contain anomalous cells that are more likely to be identified. On the other hand, the datasets marked as "low" contain anomalous cells that are more difficult to detect. The "high" and "low" distinctions describe the level of dissimilarity between artificial anomalies and regular data.

In detail, historical anomalies are introduced by replacing a current value with its historical one multiplied by a random factor. Tail and "relational" anomalies are produced by multiplying original values with their respective random factors. "Relational" anomalies are introduced only for the variables with stronger linear relationships (i.e., having a correlation coefficient larger than 0.8). Hence, every record is equally likely to receive a historical, tail, or "relational" anomaly for one or several of its item responses.

Table 2. Ranges of the multiplicative factors used to alter the original data for each module of the generative algorithm for both higher (more obvious) and lower (less obvious) anomalies. Ranges for up and down multipliers are randomly selected with equal probability.

Anomaly Type-Level	Down Multiplier Range	Up Multiplier Range
Tail-low	0.90–1.00	1.0-1.1
Historic-low	0.30–0.60	1.3—2.0
Relational-low	0.30–0.60	1.3–2.0
Tail-high	0.20–0.30	2.0–3.0
Historic-high	0.01–0.05	2.0–3.0
Relational-high	0.01–0.05	2.0–3.0

For the relational outliers, the dataset is organized such that each row represents a record, and each column represents a field (item response). The dataset matrix may have many missing data, resulting in a sparse matrix. It is natural to have a sparse matrix of data collected from surveys, especially in the Agricultural Yield Survey, which includes multiple crops. The respondents in different states and different strata would only have certain types of crops but not all. Therefore, about 92% of the values in the Agricultural Yield data matrix are missing.

5.1. Evaluations

Several accuracy measures have been computed according to the standards found in the literature [31]. The confusion matrix is constructed by comparing the classification results to the ground-truth labels as in a binary classification problem (where the two classes are outliers and nonoutliers). This 2×2 matrix contains the counts of True Positives (TP) and True Negatives (TN) in the main diagonal, and False Positives (FP) and False Negatives (FN) in the off diagonal. TP refers to the number of true outliers correctly classified as such. TN refers to the number of true regular data (nonoutliers) correctly classified as such. FP refers to the number of regular data (nonoutliers) incorrectly classified as outliers. FN refers to the number of outliers incorrectly classified as nonoutliers. The overall accuracy is

computed as the ratio between the number of correct identifications divided by the total number of units:

$$\text{Overall accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

The recall statistics are based on the ratios computed by conditioning on the ground truth labels (for true outliers and truly regular data):

$$\text{Recall}_{\text{out}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ (Sensitivity)},$$
$$\text{Recall}_{\text{reg}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \text{ (Specificity)}.$$

The precision statistics are based on ratios computed by conditioning on the labels provided by two approaches proposed in Section 3 and 4. It shows the fraction of outlier identifications that are truly outliers:

$$\text{Precision}_{\text{out}} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

or the fraction of regular identifications that are truly regular:

$$\text{Precision}_{\text{reg}} = \frac{\text{TN}}{\text{TN} + \text{FN}}.$$

The proposed outlier detection methodologies have been evaluated for accuracy at the item response level.

Table 3 shows the results based on the Bayesian bootstrap approach setting the threshold as $\theta = 0.08$. The overall accuracy ranges between 87% and 94%, the precision for detected outliers ranges between 21% and 70%, and the recall of outliers ranges between 21% and 63%. Table 4 shows the results based on the Bayesian empirical likelihood approach without setting the threshold. At the item response level, the overall accuracy ranges between 83% and 94%, the precision for detected outliers ranges between 10% and 71%, and the recall of outliers ranges between 3% and 73%. Furthermore, the precision for regular cells (i.e., for data entries that are not cellwise outliers) is larger than 87% for both approaches. The recall for regular cells is larger than 93%. The change in contamination level (from high to low) substantially affects the precision and recall of outliers, with drops of 31–90%; however, the precision and recall for regular cells is quite stable with differences of 1–6%. The overall accuracy has also shown a similar behavior with differences of 2–8%. In general, both proposed methods perform better on datasets with higher contamination levels.

Table 3. The cell-level overall accuracy, precision and recall for two labels (i.e., outliers and nonoutliers) are computed on several synthetic datasets with two contamination settings and threshold $\theta = 0.08$ in Bayesian bootstrap approach. Sensitivity varies between 21% and 63%, and specificity varies between 93% and 98%.

Survey	Level	Precision Regular	Precision Outlier	Recall Regular	Recall Outlier	Overall Accuracy
Cranberry	Low	0.893	0.483	0.962	0.233	0.867
Cranberry	High	0.934	0.700	0.978	0.429	0.919
Cattle	Low	0.932	0.278	0.938	0.260	0.880
Cattle	High	0.955	0.551	0.961	0.513	0.923
Ag. Yield	Low	0.963	0.327	0.944	0.433	0.914
Ag. Yield.	High	0.976	0.478	0.957	0.625	0.937
Sheep/Goats	Low	0.930	0.209	0.932	0.208	0.873
Sheep/Goats	High	0.951	0.420	0.951	0.421	0.910

Table 4. The cell-level overall accuracy, precision and recall for two labels (i.e., outliers and nonoutliers) are computed on several synthetic datasets with two contamination settings in Bayesian empirical likelihood approach. Sensitivity varies between 3% and 73%, and specificity varies between 93% and 98%.

Survey	Level	Precision Regular	Precision Outlier	Recall Regular	Recall Outlier	Overall Accuracy
Cranberry	Low	0.867	0.095	0.952	0.033	0.832
Cranberry	High	0.925	0.567	0.968	0.347	0.902
Cattle	Low	0.924	0.322	0.973	0.138	0.902
Cattle	High	0.957	0.708	0.980	0.522	0.941
Ag. Yield	Low	0.965	0.291	0.929	0.461	0.901
Ag. Yield	High	0.982	0.437	0.940	0.731	0.928
Sheep/Goats	Low	0.925	0.177	0.947	0.130	0.881
Sheep/Goats	High	0.961	0.510	0.956	0.539	0.924

5.2. Comparisons

In this section, the performances of the four approaches are compared on two datasets discussed above for the Agricultural Yields and Cattle Inventory. The first approach described in Section 2.1 is the FuzzyHRT algorithm. Given a user-provided degree of contamination $\theta = 0.08$, one can classify the observed data values as outliers or not by using the 100θ empirical percentile of the final scores produced through the FuzzyHRT algorithm. The DDC method [8] is used as a second application with the same contamination threshold set at $\theta = 0.08$. DDC was the first method proposed in the literature to detect cellwise outliers in multivariate datasets by accounting for the correlations among variables. The third method applied is the Bayesian Bootstrap (with a contamination threshold set to $\theta = 0.08$) and the fourth one is the Bayesian approach based on empirical likelihoods.

Figure 2 shows the overall accuracies of the four methods for each available state. All datasets are split by states because the DDC drops all variables with more than 50% of missing values by default, and it processes only the few that remain. However, the other three methods are better suited for sparse matrices and use all available data entries. Therefore, the accuracies of the four methods are compared at the state level. When applied to the Agriculture Yield dataset, the DDC algorithm does not provide the overall accuracies for six states due to the high level of sparseness. When applied to the Cattle Inventory dataset, the DDC did not produce results for one state. In contrast, the other three methods have identified anomalies in all states. Therefore, the results for the states where the DDC can not detect outliers are excluded from the comparisons and are not shown in Figure 2. The upper left panel (a) is based on the "low" Cattle Inventory dataset, and the upper right panel (b) is based on the "high" one. The lower left panel (c) is based on the "low" Agriculture yield dataset, and the lower right panel (d) is based on the "high" one.

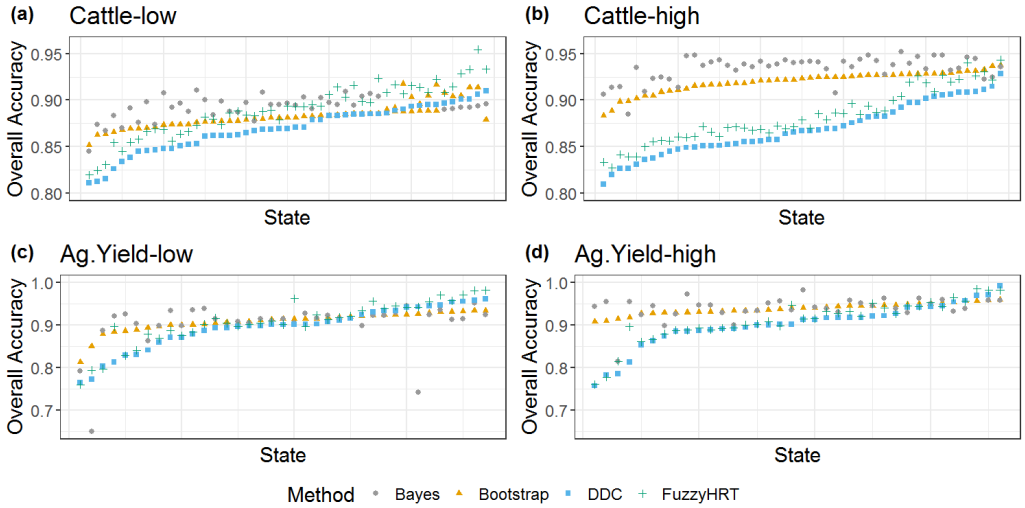

Figure 2. The state-level overall accuracies obtained from four methods applied to four synthetic datasets.

As shown by the graphs, the Bayesian bootstrap approach and the Bayesian likelihood approach have correctly detected more outliers on "high" Cattle Inventory dataset and have provided generally higher accuracies for all states than the DDC and FuzzyHRT methods. The Bayesian likelihood approach has the highest accuracies in 45 out of 50 states among the four methods. As shown in the previous section, similar results are obtained when comparing the bootstrap approach with DDC and FuzzyHRT methods. There are mixed accuracy results throughout states for both Cattle and Agriculture Yield datasets with a "low" level of contamination. Specifically, the Bayesian empirical likelihood approach has performed better in 28 out of 50 states in the "low" Cattle dataset, while FuzzyHRT has performed better in 15 out of 50 states. Similarly, the Bayesian empirical likelihood approach performed better in 18 out of 50 states in the "low" Agriculture Yield dataset, while FuzzyHRT performed better in 15 out of 50 states. These results are reasonable because the percentage of historical outliers in the Cattle Inventory datasets is larger than the percentage in the Agricultural Yield datasets.

6. Conclusions

In many applications, it is more helpful to check for anomalies at the data-entry level rather than at the record level. The FuzzyHRT algorithm uses the Bienaymé–Chebyshev's inequality and fuzzy logic, along with a user-provided level of contamination to detect four different types of outliers resulting from format inconsistencies, historical, tail, and relational anomalies within a record. The user-provided level of contamination contributes to the uncertainty associated with the outlier detection. In addition, fuzzy logic is not suited for the probabilistic reasoning behind the identification of anomalous cells.

The novelty of this work stands on mitigating the uncertainty associated with the scores produced by the FuzzyHRT algorithm [23]. Two methods are developed under a Bayesian framework. The proposed bootstrap approach explores the uncertainty associated with the output scores. The empirical likelihoods approach provides a probabilistic reasoning behind the detection of anomalous cells. The new algorithm based on empirical likelihoods at the cell level are developed as a better alternative to FuzzyHRT.

Furthermore, the new and improved algorithms can be applied to datasets that potentially suffer from the presence of cellwise anomalies, skewed distributions (with positive support), missing values, and multivariate relationships. The new algorithms do effectively cope with sparse and missing data by accounting for zero inflation without removing entire records and/or variables with missing values.

The performance of the proposed algorithms has been illustrated using NASS livestock and crop survey data with randomly generated anomalies. Our simulation study

considered four different datasets, where the proposed algorithms detected the cellwise outliers with high accuracy and robustness. Moreover, comparisons using real data with previously reported data illustrate that the proposed approaches have generally higher overall accuracy than the DDC method. When previously reported data are not available, the proposed algorithms are comparable (or even equivalent) to the DDC method (as is the case for the FuzzyHRT algorithm). However, as an advantage, our algorithms are designed to identify cellwise outliers without dropping records or variables with many missing values (as is the case for the DDC algorithm). Lastly, as future work, the algorithm could be updated to produce a candidate (prediction) value for each detected cell anomaly by leveraging administrative, structured, or unstructured data available for the whole or a subset of the surveyed records.

Author Contributions: Conceptualization, V.B. and L.S.; Methodology, L.S. and V.B.; Software, L.S. and L.C.; Validation, L.S.; Formal Analysis, L.C. and L.S.; Investigation, V.B., L.S. and L.C.; Writing—Original Draft Preparation, V.B., L.S. and L.C.; Writing—Review and Editing, V.B., L.S. and L.C.; Visualizations, L.C.; Supervision, V.B. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the US Department of Agriculture's National Agricultural Statistics Service.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to the NASS Confidentiality Pledge, the data underlying this article cannot be shared publicly. Secure access of NASS data may be obtained by agreement and sworn status only; restrictions apply. More information is available at https://www.nass.usda.gov/Data_and_Statistics/Special_Tabulations/index.php.

Acknowledgments: The findings and conclusions in this paper are those of the authors and should not be construed to represent any official USDA, or US Government determination or policy. The authors would like to thank the editors and the reviewers for providing comments that improved this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

US	United States
USDA	United States Department of Agriculture
NASS	National Agricultural Statistics Service
DDC	Detect-Deviating-Cells
KL	Kullback–Leibler
H	Historical
R	Relational
T	Tail

References

1. Fellegi, I.P.; Holt, D. A systematic approach to automatic edit and imputation. *J. Am. Stat. Assoc.* **1976**, *71*, 17–35.
2. Box, G.E.; Tiao, G.C. A Bayesian approach to some outlier problems. *Biometrika* **1968**, *55*, 119–129.
3. Guttman, I.; Dutter, R.; Freeman, P.R. Care and handling of univariate outliers in the general linear model to detect spuriosity—A Bayesian approach. *Technometrics* **1978**, *20*, 187–193.
4. Abraham, B.; Box, G.E. Linear models and spurious observations. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1978**, *27*, 131–138.
5. Freeman, P.R. On the number of outliers in data from a linear model. *Trab. Estadística Investig. Oper.* **1980**, *31*, 349–365.
6. Alqallaf, F.; Van Aelst, S.; Yohai, V.J.; Zamar, R.H. Propagation of outliers in multivariate data. *Ann. Stat.* **2009**, *37*, 311–331.
7. Agostinelli, C.; Leung, A.; Yohai, V.J.; Zamar, R.H. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* **2015**, *24*, 441–461.

8. Rousseeuw, P.J.; Van den Bossche, W. Detecting Deviating Data Cells. *Technometrics* **2018**, *60*, 135–145. https://doi.org/10.1080/00401706.2017.1340909.
9. Knorr, E.M.; Ng, R.T.; Tucakov, V. Distance-based outliers: Algorithms and applications. *VLDB J.* **2000**, *8*, 237–253.
10. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
11. Savitsky, T.D. Scalable approximate Bayesian inference for outlier detection under informative sampling. *J. Mach. Learn. Res.* **2016**, *17*, 1–49.
12. Smiti, A. A critical overview of outlier detection methods. *Comput. Sci. Rev.* **2020**, *38*, 100306.
13. Boukerche, A.; Zheng, L.; Alfandi, O. Outlier detection: Methods, models, and classification. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–37.
14. Peña, D.; Guttman, I. Comparing Probabilistic Methods for Outlier Detection in Linear Models. *Biometrika* **1993**, *80*, 603–610.
15. Gao, J.; Tan, P.N. Converting output scores from outlier detection algorithms into probability estimates. In Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Hong Kong, China, 18–22 December 2006; IEEE: Piscataway, NJ, USA, 2006; pp. 212–221.
16. Chaloner, K.; Brant, R. A Bayesian approach to outlier detection and residual analysis. *Biometrika* **1988**, *75*, 651–659.
17. Varbanov, A. Bayesian approach to outlier detection in multivariate normal samples and linear models. *Commun. Stat. Theory Methods* **1998**, *27*, 547–557.
18. Peng, F.; Dey, D.K. Bayesian analysis of outlier problems using divergence measures. *Can. J. Stat.* **1995**, *23*, 199–213.
19. Geisser, S. Discussion of a paper by G. E. P. Box. *J. R. Statist. Soc. A* **1980**, *143*, 416–417.
20. Geisser, S. Influential observations, diagnostics and discovery tests. *J. Appl. Stat.* **1987**, *14*, 133–142.
21. Geisser, S. *Predictive Approaches to Discordancy Testing*; Technical Report; University of Minnesota: Minneapolis, MN, USA, 1987.
22. Geisser, S. *Diagnostics, Divergences and Perturbation Analysis*; Technical Report; University of Minnesota: Minneapolis, MN, USA, 1989.
23. Sartore, L.; Chen, L.; van Wart, J.; Dau, A.; Bejleri, V. Identifying Anomalous Data Entries in Repeated Surveys. *J. Data Sci.* **2024**, *22*, 436–455. https://doi.org/10.6339/24-JDS1136.
24. Zwillinger, D. *Standard Mathematical Tables and Formulas*; CRC Press: Boca Raton, FL, USA, 2018.
25. Owen, A.B. *Empirical Likelihood*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2001.
26. Lazar, N.A. Bayesian empirical likelihood. *Biometrika* **2003**, *90*, 319–326.
27. Kolditz, T.; Kissinger, T.; Schlegel, B.; Habich, D.; Lehner, W. Online bit flip detection for in-memory b-trees on unreliable hardware. In Proceedings of the Tenth International Workshop on Data Management on New Hardware, Snowbird, UT, USA, 23 June 2014; pp. 1–9.
28. Das, S.; Chatterjee, A.; Ghosh, S. Investigating impact of bit-flip errors in control electronics on quantum computation. *arXiv* **2024** arXiv:2405.05511.
29. Hanson, M. Duality and self-duality in mathematical programming. *J. Soc. Ind. Appl. Math.* **1964**, *12*, 446–449.
30. Walk, M. *Theory of Duality in Mathematical Programming*; Walter de Gruyter GmbH & Co KG: Berlin, Germany, 2022; Volume 72.
31. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-label confusion matrix. *IEEE Access* **2022**, *10*, 19083–19095.