Journal of Data Science 22 (3), 436–455
July 2024

DOI: 10.6339/24-JDS1136
Statistical Data Science

# Identifying Anomalous Data Entries in Repeated Surveys[☆]

Luca Sartore[1,2,*], Lu Chen[1,2], Justin van Wart[2], Andrew Dau[2], and Valbona Bejleri[2]

[1]*National Institute of Statistical Sciences, 1750 K Street NW Suite 1100, Washington DC, 20006, USA*
[2]*United States Department of Agriculture, National Agriculture Statistics Service, 1400 Independence Avenue SW, Washington DC, 20250, USA*

**Abstract**

The presence of outliers in a dataset can substantially bias the results of statistical analyses. In general, micro edits are often performed manually on all records to correct for outliers. A set of constraints and decision rules is used to simplify the editing process. However, agricultural data collected through repeated surveys are characterized by complex relationships that make revision and vetting challenging. Therefore, maintaining high data-quality standards is not sustainable in short timeframes. The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) has partially automated its editing process to improve the accuracy of final estimates. NASS has investigated several methods to modernize its anomaly detection system because simple decision rules may not detect anomalies that break linear relationships. In this article, a computationally efficient method that identifies format-inconsistent, historical, tail, and relational anomalies at the data-entry level is introduced. Four separate scores (i.e., one for each anomaly type) are computed for all nonmissing values in a dataset. A distribution-free method motivated by the Bienaymé-Chebyshev's inequality is used for scoring the data entries. Fuzzy logic is then considered for combining four individual scores into one final score to determine the outliers. The performance of the proposed approach is illustrated with an application to NASS survey data.

**Keywords** *agricultural data; Bienaymé-Chebyshev's inequality; cellwise outliers; fuzzy logic; outlier detection; statistical analysis*

## 1 Introduction

Statistical analyses of a dataset with outliers can be biased. There is an extensive literature on how to mitigate outliers when conducting data analyses. For instance, developing robust estimators can down-weight the contribution of outliers on final estimates (Huber and Ronchetti, 1981). Alternatively, identification of anomalous records before the start of statistical analyses can lead to either their removal (Stigler, 1973) or their correction through editing procedures (De Waal et al., 2011). In general, correcting anomalous values can improve the overall accuracy and precision of final estimates.

To maintain high-quality data standards, micro edits are manually performed for each

[☆]The findings and conclusions in this article are those of the authors and should not be construed to represent any official USDA or US Government determination or policy. This research was supported in part by the intramural research program of the US Department of Agriculture, National Agriculture Statistics Service.

[*]Corresponding author. Email: luca.sartore@usda.gov or lsartore@niss.org.


Received November 30, 2023; Accepted April 16, 2024

record using auxiliary information. These operations can be lengthy even when using automated screening procedures based on simple constraints and decision rules. Further challenges are also presented by the nature of the data. For example, agricultural data acquired through repeated surveys are often characterized by complex relationships, which limit the efficacy of rule-based systems.

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS) has partially automated the review and vetting of the information acquired through surveys. The current system quickly detects anomalies using decision rules designed by agricultural experts. However, these rules are static and tend to point out extreme values and ignore potential relationships with other variables. Therefore, NASS has investigated the use of a data-driven methodology to improve its anomaly detection system.

The classical view on outlier detection is based on a contamination model at the record level (Huber and Ronchetti, 1981). A new perspective on outliers explores anomalies in the cells (entries) of the data matrix (or tabular dataset). Cellwise outliers occur when an individual cell substantially deviates from its "standard" behavior. This perspective implies that the contamination rate can exceed the classical 0.5 breakdown point, which is the proportion of anomalous values that can invalidate the analysis if introduced in the dataset (Alqallaf et al., 2009). Traditional estimators cannot be used when the contamination happens within records (i.e., at the entry/cell level; Agostinelli et al., 2015). Overall, cellwise-outlier detection methods are more informative than traditional record-level algorithms.

Most cellwise detection procedures have been developed under the assumption of multivariate normality (e.g., see Raymaekers and Rousseeuw, 2019), which is unsuitable for agricultural survey data. The literature suggests a few outlier-detection techniques for data that are not normally distributed. For example, Filzmoser and Gregorich (2020) proposed a multivariate method for spatial compositional data. However, this approach does not account for stratified samples and is not applicable to generic NASS data (Miller et al., 2010). In general, distribution-free methods should be preferred for avoiding too stringent assumptions.

NASS collects nonnegative data often characterized by zero-inflation. Observations of zero are often associated with farms that do not produce specific commodities. Also, missing values due to item nonresponse are common in NASS survey data. Agostinelli et al. (2015) overlooked the mechanisms of data missingness when identifying cellwise outliers. These mechanisms are of three different types, i.e., missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) as suggested by Rubin (1976). The type of missingness is important when conducting any inference from the data. Raymaekers and Rousseeuw (2019) proposed a detection-imputation approach like the Estimation-Maximization (EM) algorithm, which handles missing values when correcting the anomalies. This approach, however, fails to provide suitable results when more than 50% of the entries are missing for either a record or a variable.

The `cellHandler` method (Raymaekers and Rousseeuw, 2019) and the Detect-Deviating-Cells (DDC) method (Rousseeuw and Van den Bossche, 2018) have been developed to account for the correlation structures of high-dimensional genetic datasets. The DDC approach is one of the first methods devoted to the detection of cellwise outliers. In particular, this approach starts by standardizing the data and then flags the anomalous cells using the information available in individual columns (i.e., via univariate distribution-tail analysis). Each data cell of a given record is then predicted based on the other unflagged cells that are correlated with the cell in question. Finally, a cell is considered an outlier if its observed value (assumed to follow a Gaussian distribution) deviates significantly from its predicted value.

To the best of the authors' knowledge, state-of-the-art methods on cellwise outlier detection do not take advantage of Previously Reported Data (PRD) to assess irregular departures from historical trends. PRD can enhance the outlier detection on data collected through longitudinal studies or repeated surveys, such as those conducted by government agencies or other national and international institutions. Furthermore, the current statistical literature does not include a distribution-free method to detect cellwise outliers.

In this article, a new method of identifying cellwise outliers is proposed. As is the case with the DDC, the approach accounts for the correlation structure arising from high-dimensional data, but it is applied to the full dataset, even when the correlation structure is sparse. A distribution-free approach is developed, and PRD is used, when available.

In addition, in the new approach, fuzzy logic is used to detect cellwise outliers resulting from different types of anomalies. Four types of cellwise contamination are typically observed in NASS survey data: 1) bit-flip errors (i.e., changes of a few binary values from zero to one or vice versa due to cosmic rays; O'Gorman, 1994), 2) historical anomalies (i.e., large deviations from PRD), 3) distribution-tail anomalies (i.e., univariate outliers), and 4) relational anomalies (i.e., deviations from typical multivariate relationships). These four types have been considered when developing the proposed algorithm because the DDC does not check for bit-flip errors. Four separate scores (i.e., one for each contamination type) are computed for all nonmissing values in a dataset. Chebyshev's inequality (Tchebichef, 1867) and its robust extension (Bienaymé, 1867) are used for scoring the data entries without imposing distributional assumptions. The four scores are then combined into one final score to determine the anomalous entries.

The remaining sections are organized as follows. Section 2 describes the four types of contamination and introduces the methodological background to identify cellwise outliers. Numerical aspects to achieve high-performance computing are addressed in Section 3. Section 4 provides a simulation study where the anomalies are randomly introduced on NASS survey data. Section 5 further illustrates the application of the proposed methodology using ground-truth labels from metadata on manual edits. The results from this application illustrate the improvements of the proposed method when compared to the DDC method implemented in the R-package *cellWise* (Raymaekers et al., 2023). Concluding remarks are given in Section 6.

## 2 Methods

Consider a continuous random variable $X$ with an arbitrary distribution $F_X(\cdot)$. No assumption is made about analytical expression of $F_X(\cdot)$. It is just assumed that $F_X(\cdot)$ has finite central absolute moments, symbolically represented by

$$\mathsf{E}\left[|X-\mu|^{\delta}\right] = \int_{\mathbb{R}} |x-\mu|^{\delta} \mathrm{d}F_X(x) < \infty, \tag{1}$$

where $\mu$ denotes the location parameter of $X$, and the scalar $\delta \geqslant 1$ represents the order of the moments. Location estimates, for a given $\delta$, are obtained by minimizing (1) with respect to $\mu$. The median of a data vector $\mathbf{x}$ is the solution denoted by $\hat{\mu}_1$ that minimizes the expectation in (1) when $\delta = 1$. Also, the mean of $\mathbf{x}$ is the solution denoted by $\hat{\mu}_2$ that minimizes the expectation in (1) when $\delta = 2$. The expectation in (1) is known as the mean absolute error (MAE) when $\delta = 1$, and the mean squared error (MSE) when $\delta = 2$.

For a given moment of order $\delta$ (i.e., $\hat{\mu}_\delta$) and a given threshold, $\epsilon > 0$, one can use the Bienaymé-Chebyshev's inequality (Chepulis and Shevlyakov, 2020) to compute an upper

bound (in probability) for deviations of the random variable $X$ from its central moments. The Bienaymé-Chebyshev's inequality is formulated as

$$\Pr\left(|X - \hat{\mu}_\delta| \geqslant \epsilon\right) \leqslant \epsilon^{-\delta} \mathsf{E}\left[\left|X - \hat{\mu}_\delta\right|^\delta\right], \tag{2}$$

where the right-hand side is finite for any $\delta \in \{1, 2\}$. For $\delta = 2$, this inequality is equivalent to the classical Chebyshev's inequality (Zwillinger, 2018), where

$$\mathsf{E}\left[\left|X - \hat{\mu}_\delta\right|^\delta\right] = \mathsf{Var}[X].$$

Typical outlier detection methods based on the Chebyshev's inequality produce a confidence interval around the mean. In fact, a value is likely classified as an outlier if the probability of the value being far from the mean more than $\epsilon$ does not exceed $\epsilon^{-2}\mathsf{Var}[X]$. It is also trivial to show that the probability on the left side of (2) goes to zero as $\epsilon \to \infty$ at a faster rate than $o(\epsilon^{-\delta})$. If the value of $\epsilon$ is allowed to vary depending on a realization $x$ of the random variable $X$, namely $\epsilon = g(x)$, where $g : \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$ is a generic function with nonnegative codomain, the inequality (2) could be rewritten as

$$\Pr\left(|X - \hat{\mu}_\delta| \geqslant g(x)\right) \leqslant \min\left\{1, g(x)^{-\delta} \mathsf{E}\left[\left|X - \hat{\mu}_\delta\right|^\delta\right]\right\}. \tag{3}$$

The right side of (3) could be thought of as a measure of regularity (or score) for a given sample. For the specific choice of $g(x) = \left|x - \hat{\mu}_\delta\right|$, the score is defined as

$$s = \min\left\{1, |x - \hat{\mu}_\delta|^{-1} \mathsf{E}\left[\left|X - \hat{\mu}_\delta\right|^\delta\right]^{1/\delta}\right\} \in [0, 1], \tag{4}$$

where the power of $1/\delta$ is used to regularize the score across different orders of the moments (e.g., when identifying different anomaly types). Large values of $s$ indicate regular data, and small values indicate anomalous ones. Typically, probabilistic inequalities are not based on distributional assumptions and hence their application does not require a data transformation. Nonetheless, link functions can be considered in cases when the support of $X$ is a subset of $\mathbb{R}$. For example, the logarithm is used for random variables with positive support, or the logit function for random variables with $(0, 1)$ bounded support. Because NASS collects positive agricultural counts, the log transformation is considered for the rest of the paper. However, depending on the application at hand, other link functions might be more appropriate.

NASS often collects data using a stratified sampling design. The population of interest is divided into $K \in \mathbb{N}$ homogeneous groups called strata. Each stratum $k \in \{1, \ldots, K\}$ consists of units that share common attributes. A random sample of size $n_k \in \mathbb{N}$ is taken from each stratum $k$ with sample size proportional to the stratum size (computed with respect to the whole population size). These subsets of the strata are then pooled to form a random sample. Notice that stratified sampling designs are complex survey designs, and so the sampling weights are usually unequal. However, in this article, our attention is primarily on the identification of anomalies from individual record responses, rather than at aggregated level when considering the weights. As a result, the consideration of sampling weights is not within the scope of this article. Stratification is not specifically discussed in the cellwise-outlier literature; however, inferences by conditioning on strata could result in improved outlier detection due to the data homogeneity within each stratum.

Statisticians at NASS encounter four types of cellwise anomalies. A specific algorithm based on equation (4) is developed to score the data (in a continuous scale from [0, 1]) for each type of cellwise anomaly. The score notation $s^{[B]}$ is used for bit-flip errors, $s^{[H]}$ for historical-anomalies, $s^{[T]}$ for distribution-tail anomalies, and $s^{[R]}$ for relational anomalies. The four different scores are combined through fuzzy logic to compute the final score $s^* \in [0, 1]$. In general, the proposed method operates with multivariate datasets where the number of records, $n$, is assumed to be larger than the number of variables, $p$.

## 2.1 Data-Format Anomalies

In sparse datasets, zeros and missing values are usually removed from storage files. Therefore, nonpositive values that appear in a database would be inconsistent with the data format in use. Although this type of anomaly is rare, it might occur due to bit-flip errors (O'Gorman, 1994), i.e., when binary corruptions occur in Random Access Memory (RAM). For instance, binary corruptions can erroneously alter a positive sign to a negative one. Data-format anomalies are highlighted with binary scores, $s_{ij}^{[B]}$, where subscript $i = 1, \ldots, n$ denotes records and subscript $j = 1, \ldots, p$ denotes variables. These scores are zero if nonpositive values are observed, and one otherwise. Bit-flip errors can also introduce anomalies that are consistent with the sparsity format, but these anomalies are identified by the algorithms discussed in the following subsections.

## 2.2 Historical Anomalies

The statistical literature suggests quantile-based outlier-detection methods for univariate time series (Hidiroglou and Berthelot, 1986; Sandqvist, 2016). These methods rely on the construction of a robust interval often defined by the interquartile range (IQR). However, these approaches do not perform as one would expect, especially when the differences over time have a leptokurtic (i.e., heavy-tailed) distribution. In these cases, when the IQR estimates are often zero, another solution is needed.

For repeated surveys, historical information is available for units that have participated and responded to surveys administered in the past. One can predict the current value using available information through forecasting or filtering. Therefore, at time $t \in \mathbb{Z}$, a positive response is an historical cellwise anomaly if the discrepancy between current, $x_{ijt}$, and predicted value, $\hat{x}_{ijt}$, is too large. The differences of current and predicted values on the log scale,

$$\Delta_{ijt} = \log x_{ijt} - \log \hat{x}_{ijt},$$

are compared for any $i = 1, \ldots, n$ and $j = 1, \ldots, p$. These differences are assumed to have zero median and a finite positive MAE, even under leptokurtic (or heavy-tailed) distributions. In this article, the predicted value, $\hat{x}_{ijt} = x_{ij(t-1)}$, is based on an autoregressive model of order one (which uses PRD) and with autoregressive parameter equal to one.

When considering PRD, missing values at time $t - 1$ may result from the sampling process because not all units in the current sample have participated in a previous survey. Either $x_{ijt}$ or the PRD $x_{ij(t-1)}$ would result in `NA` if a unit appears in two consecutive samples and produces different commodities at times $t$ and $t - 1$. Therefore, $\Delta_{ijt}$ will be undefined because one of its components is undefined (i.e., either $\log x_{ijt} =$ `NA` or $\log x_{ij(t-1)} =$ `NA`). Positive data entries at time $t$ associated with these two problematic cases are considered historical regularities with score one.

For historical anomalies, the MAE of the finite log-differences is computed under the assumptions that $\delta = 1$ and $\hat{\mu}_1 = 0$. However, an alternative measure of variability is preferred when the log-differences are also zero inflated. For example, for $m \in \mathbb{N}$ data entries that satisfy the inequality:

$$0 < |\Delta_{ijt}| < \infty,$$

the historical variability can be measured as

$$\hat{\sigma}^{[H]} = \sum_{ij} \frac{|\Delta_{ijt}|}{h}, \tag{5}$$

where $h < np$. Therefore, the historical regularity score, $s_{ij}^{[H]} \in [0, 1]$, is computed as

$$s_{ij}^{[H]} = \min\left\{1, \frac{\hat{\sigma}^{[H]}}{|\Delta_{ijt}|}\right\}.$$

Leptokurtic (i.e., heavy-tailed) or zero-inflated distributions with positive kurtosis may not always characterize the log-differences between two repeated surveys. In the cases when relatively larger variations around the median have been observed over time, such as in thin tail distributions with positive kurtosis (i.e., platykurtic), the estimator in (5) should be replaced with the MAE. Furthermore, the MAE is typically estimated for each distinct variable by accounting for the stratification and including the zeros. As a matter of fact, this approach is more appropriate for longitudinal studies where the sampling units do not change over time. However, it is impractical in repeated surveys where consecutive appearances of a record are often limited by the sampling design. In general, the selection of an estimator for $\sigma^{[H]}$ depends on the number of usable data entries needed to produce reliable results.

### 2.3 Distribution-Tail Anomalies

Survey data within a stratum share similar features that one could use in subsequent analyses to effectively detect distribution-tail anomalies. Even if the data suffer from excessive skewness, nonlinear transformations can highlight extreme values on both ends of the distribution. When studying the anomalies on the tail of a distribution, one can also encounter missing values. However, the mechanisms producing these types of anomalies are different from those of the historical anomalies. For example, not all farms produce specific commodities, and the missing commodities are true zeros, which lead to nonfinite values when transformed. All nonpositive entries, i.e. $x_{ijt} \leqslant 0$, are disregarded from the stratum-level analysis, due to their incompatibility with the log transformation. These cases are automatically labeled as "regular", and a score of one is assigned for compatibility with successive calculations.

Distribution-tail anomalies are identified using an alternative regularity score (yet very similar to (4) with $\delta = 1$). For each stratum $k \in \{1, \ldots, K\}$, and any variable $j \in \{1, \ldots, p\}$, robust estimates of the location and scale parameters are computed using $h_{jk} \in \mathbb{N}$ usable values (where $h_{jk} \leqslant n_k$). Here, the notation $\hat{\mu}_{1jk}$ represents the median of variable $j$ within stratum $k$. The following scale estimator was proposed by Hampel (1974) as a robust replacement for the MAE:

$$\hat{\sigma}_{jk}^{[T]} = \text{median}|\log x_{ijt} - \log \hat{\mu}_{1jk}|,$$

where the index $i \in \{1, \ldots, n_k\}$ is used for records within stratum $k$ at time $t$, and $n_k$ denotes the size of the stratum $k$. Finally, the distribution-tail regularity score, $s_{ij}^{[T]} \in [0, 1]$, is computed

as

$$s_{ij}^{[T]} = \min\left\{1, \frac{\hat{\sigma}_{jk}^{[T]}}{|\log x_{ijt} - \log \hat{\mu}_{1jk}|}\right\}.$$

## 2.4 "Relational" Anomalies

Units with multiple item responses could be subject to several relationships among the variables in a dataset. When these relationships are broken, the items providing invalid information are considered anomalous. In this article, the term "relational" refers exclusively to the anomalies that violate linear dependencies between two or among several variables. In the presence of linear dependency, values of a variable can be estimated as outputs from a linear model where the values of the other variables are used as inputs. This type of anomaly is usually identified by robust multivariate standardizations.

When scoring for a "relational" anomaly, missing values need to be imputed. Unlike previous anomaly types that disregard missing values, relational scores are computed using all entries in the dataset because matrix-algebra routines require all input data to be finite. Therefore, the following standardized values at the stratum level are considered:

$$y_{ij} = \begin{cases} (\log x_{ijt} - \log \hat{\mu}_{1jk})/\hat{\sigma}_{jk}^{[T]}, & \text{if } x_{ijt} > 0, \\ 0, & \text{if } x_{ijt} \leqslant 0 \text{ or missing,} \end{cases}$$

for all $i = 1, \ldots, n$, and $j = 1, \ldots, p$. This approach coincides with the replacement of missing values with the most suitable stratum-level medians computed before applying any transformation and standardization. Afterwards, these transformed data entries are organized in a matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$, where the columns represent $p$ noncollinear variables. This assumption allows for the estimation of linear models formulated as

$$y_{ij} = \sum_{\ell \neq j} \beta_{j\ell} y_{i\ell} + \varepsilon_{ij},$$

where $\beta_{j\ell}$ represents the $\ell$-th coefficient of model $j \in \{1, \ldots, p\}$, and $\varepsilon_{ij}$ denotes the error of model $j$ for the record $i \in \{1, \ldots, n\}$.

Rather than estimating the variance-covariance matrix of $\mathbf{Y}$ to rotate the data (as proposed by Rousseeuw and Van den Bossche, 2018), ordinary least squares (OLS) are performed for scoring "relational" anomalies based on the MSE criteria ($\delta = 2$). In fact, the OLS are used to estimate a matrix of model parameters $\mathbf{B} \in \mathbb{R}^{p \times p}$. Furthermore, the diagonal components of this matrix are assumed to be fixed at zeros across the regression process, i.e. $\beta_{jj} = 0$, for all $j = 1, \ldots, p$. By introducing an identity matrix $\mathbf{I}_p$ of size $p \times p$, the residual matrix is formulated as $\mathbf{E} = \mathbf{Y}(\mathbf{I}_p - \mathbf{B})$. Parameter estimates $\hat{\beta}_{j\ell}$, for any $\ell \neq j$, are obtained by minimizing the MSE of column $j$ in $\mathbf{E}$, for all $j = 1, \ldots, p$. Therefore, optimal residuals are

$$\hat{\varepsilon}_{ij} = y_{ij} - \sum_{\ell \neq j} \hat{\beta}_{j\ell} y_{i\ell},$$

for all $i = 1, \ldots, n$, and $j = 1, \ldots, p$. Finally, the "relational" scores are computed as

$$s_{ij}^{[R]} = \min\left\{1, \frac{1}{|\hat{\varepsilon}_{ij}|}\sqrt{\frac{\sum_{\ell=1}^{n} \hat{\varepsilon}_{\ell j}^2}{n-1}}\right\}.$$

### 2.5 Anomaly Score Based on Fuzzy Logic

The four scores (i.e., one for each anomaly type) are combined into a final one to identify the anomalous cells. The concept of triangular norms (or t-norms) is borrowed from the fuzzy logic literature (Gupta and Qi, 1991) as a rigorous foundation for the proposed detection algorithm.

A t-norm is a function $T : [0, 1] \times [0, 1] \to [0, 1]$ that satisfies the following properties for any $z_1, z_2, z_3 \in [0, 1]$:

$$\begin{aligned} T(z_1, z_2) &= T(z_2, z_1) \text{ (commutativity)}, \\ T(z_1, z_2) &\leqslant T(z_1, z_3), \text{ if } z_2 \leqslant z_3 \text{ (monotonicity)}, \\ T(z_1, T(z_2, z_3)) &= T(T(z_1, z_2), z_3) \text{ (associativity)}, \\ T(z_1, 1) &= z_1 \text{ (identity)}. \end{aligned}$$

The product t-norm, defined as $T(z_1, z_2) = z_1 z_2$, can be used recursively to combine the four anomaly scores into a final one, i.e.

$$\begin{aligned} s_{ij}^* &= T\left(s_{ij}^{[H]}, T\left(s_{ij}^{[R]}, T\left(s_{ij}^{[T]}, s_{ij}^{[B]}\right)\right)\right) \\ &= s_{ij}^{[H]} s_{ij}^{[R]} s_{ij}^{[T]} s_{ij}^{[B]}. \end{aligned}$$

**Remark 1.** The outliers resulting from any type of data anomaly can be thought of as a fuzzy set. Therefore, there is a degree of uncertainty or imprecision described by individual scores for each data entry to belong either in or out of any outlier-type set. This results in a combination of complex scenarios when evaluating the outlier status of a given data entry. A fuzzy logic system allows one to integrate and process uncertainties associated with each type of anomaly. More details on fuzzy logic and probabilistic inequalities are given in Appendix A. □

A single user-based threshold, $\theta \in (0, 1)$, is provided as the degree of contamination used to determine cellwise anomalies. Several contamination criteria can be considered to set a reasonable value of $\theta$. For example, $\theta$ can depend on either the user expectations or the maximum number of cells (or entries) that can be manually edited over a feasible timeframe. This user-based value is successively used to compute the $100\theta$ empirical percentile of the final scores, i.e. $\hat{Q}_\theta$. Lastly, the cellwise anomalies are identified if the inequality $s_{ij}^* < \hat{Q}_\theta$ is satisfied, for any $i = 1, \ldots, n$, and $j = 1, \ldots, p$.

## 3 Computational Aspects

Several algorithms have been developed or reimplemented to accelerate the detection algorithm. Although the main program can be executed in R, its core functionalities have been coded in C. Because R is an interpreted language, it is much slower than software developed with more traditional languages (such as FORTRAN, C, and C++). Unlike R, C programs are translated into a machine-readable format as a sequence of binary instructions. Furthermore, modern C compilers can reduce the number of instructions to execute, and they also allow for different types of parallelization. For example, single-instruction multiple-data (SIMD) operations (Flynn, 1966) can process several values at once within a single core, while multicore computations are achieved via the OpenMP library (Dagum and Menon, 1998). Therefore, the required classical routines for linear algebra and statistics have been reimplemented in C to achieve higher performances than the libraries provided by the R interface.

In our approach, quick computation of medians is required for every combination of variable and stratum. In general, this can be time-consuming especially when working with high-dimensional datasets. We have improved the computational performances of the median algorithm through an iterative procedure that requires $O(n)$ parallelizable operations. The algorithm starts by finding the minimum and maximum values of a variable. The range is then split into $A$ bins of equal width. After counting the number of observations in each bin, the algorithm selects the bin that is more likely to contain the median. At the next iteration, this bin is split further into $A$ smaller bins that are used for a new set of computations. The algorithm stops when the width of the bins approaches zero (or when a user-defined maximum number of iterations is reached). At the end, the median is approximated by the lower bound of the interval provided by the last selected bin. Instead of using sorting algorithms that often require $O(n \log n)$ operations (Sedgewick, 1978), this technique employs concepts borrowed from ordinary histograms to accelerate the estimation of the median (see Algorithm 1 in Appendix B).

To reduce the computational burden caused by standard routines adopted for linear regression, we have developed customized linear algebra functions. Each column of the data matrix $\mathbf{Y}$ is processed in parallel to compute the residual matrix $\mathbf{E}$. This is achieved using $\mathbf{y}_j$ (i.e., column $j$ of matrix $\mathbf{Y}$) as the vector associated to a response variable and $\mathbf{Y}_{-j}$ (i.e., the matrix resulting from removing column $j$ from matrix $\mathbf{Y}$) as the covariate matrix. The Gram-Schmidt decomposition (Daniel et al., 1976) of matrix $\mathbf{Y}_{-j}$ has been reimplemented, such that $\mathbf{Y}_{-j} = \mathbf{QR}$, where the matrix $\mathbf{Q}$ is orthonormal, and the matrix $\mathbf{R}$ is an upper triangular matrix. This approach avoids the estimation of the regression coefficients through the explicit/analytical formulation of a vector of residuals using the following equation:

$$\hat{\mathbf{e}}_j = \mathbf{Q}\mathbf{Q}^\top \mathbf{y}_j,$$

where $\hat{\mathbf{e}}_j$ corresponds to column $j$ of the residual matrix $\mathbf{E}$. In general, the combination of these computational techniques allows one to achieve high performances, especially on sophisticated environments (such as computational clusters or clouds).

**Remark 2.** We refrained from recalling functions found in external libraries developed by the scientific community. Thus, we implemented our own code for the Gram-Schmidt decomposition and embedded it into the function that computes the residuals. Thereby, we gained a substantial time-performance advantage. Only three C libraries were used. The libraries in Windows were `libc.dll`, for standard memory-allocation functions; `libm.dll`, for standard mathematical functions; and `libgomp.dll`, for multicore processing. The respective C libraries for Linux or Mac were named `libc.so`, `libm.so`, and `libgomp.so`. Once the C code was compiled, it could easily run within other software (such as R, python, or SAS). □

## 4 Simulation Study

To assess the performance of the proposed methodology, we conducted a controlled simulation study using four national surveys administered by NASS. 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. Data collected through surveys are first internally reviewed and vetted, and a Monte Carlo outlier contamination is performed

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

| Survey | Survey date | Total respondent | Major inquiries | Scenario |
|---|---|---|---|---|
| Sheep and Goat Inventory | 1/1/2021 | 10,090 | Sheep and/or goat herd composition (ewes, rams, lambs, billies, nannies, kids, etc.) | Many records; two distinct inventories of aggregated parts |
| Cattle Inventory | 1/1/2021 | 21,154 | Cattle herd composition (cows, bulls, calves, etc.) | Many records; one inventory of aggregated parts |
| Agricultural Yield | 7/1/2021 | 1,762 | Expected yield and acres of small grain crops (i.e., barley, wheat, oats) | Fewer records; multiple crops |
| Cranberry Production | 2/1/2021 | 218 | Cranberry acres | Very few records; single crop |

next. Therefore, these data are assumed to be complete and correct for each respondent. Surveys under consideration were administered between 2021 and 2022, and have sample sizes ranging between 218 and 21,154.

Ideally, the proposed algorithm would be applied to "untouched/raw" data that are being simultaneously vetted by the data analysts during the collection phase. In fact, the proposed algorithm identifies potential cell-wise outliers and does not provide predicted values to fully automate the editing process. In a production environment, data vetting and manual edits are performed after outlier flags are generated. These procedures often occur in a cycle where the proposed algorithm runs several times per hour to update the outlier flags using the latest and most accurate information. In this case, human intervention is meant as a controlling mechanism to avoid unintended edits that a software would automatically operate on false positive cells. However, the accuracy of the proposed methodology cannot be fully evaluated using invalidated raw data. Thus, some anomalous cells have been synthetically introduced in the four datasets shown in Table 1. This approach allowed us to flag and track the anomalies for the evaluation of the proposed detection algorithm across different databases using $\theta = 0.08$.

A generative algorithm has been applied to each dataset in Table 1 by randomly replacing a few cells with anomalous values. Therefore, two distinct datasets have been created for each survey. The datasets marked as "high" contain anomalous cells that were more likely to be identified. On the other hand, the datasets marked as "low" contain anomalous cells that were more difficult to detect. The "high" and "low" distinctions describe the level of dissimilarity between artificial anomalies and regular data. These datasets are used for studying the ability of the proposed methodology to distinguish regular values from cellwise outliers.

The generative algorithm used to introduce anomalies in the datasets was composed of three specific modules synthesizing historical, tail, and "relational" anomalies, respectively. Each module randomly selects 5% of the item responses. Half of these were replaced by multiplying the current values by random factors in $(0, 1]$, and the other half using random factors greater than one. The random factors were generated from uniform distributions in intervals shown in Table 2 (more specifically in columns 2 and 3). Shrinking and expansion ranges were ran-

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 |

domly selected with equal probability for each combination of anomaly type and dissimilarity level.

Historical anomalies were introduced by replacing a current value with its historical one multiplied by a random factor. On the other hand, tail and "relational" anomalies were produced by multiplying original values with their respective random factors. "Relational" anomalies were introduced only for the variables with stronger linear relationships (i.e., having a correlation coefficient larger than 0.8). Even if this third module reduces the number of variables to contaminate, the 5% replacement rate has been kept at the same level of the other two modules. Hence, every record was equally likely to receive a historical, tail, or "relational" anomaly for one or several of its item responses. Furthermore, bit-flip errors were introduced by the simulation mechanism in 0.03% of the positive cells by setting the values to zero.

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 that 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.

Several accuracy measures have been computed according to the standards found in the literature (Heydarian et al., 2022). 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 $2 \times 2$ 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 was 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 were based on the ratios computed by conditioning on the ground truth

labels (for truly 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 the fuzzy logic system proposed in Section 2. 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 methodology has been evaluated for accuracy at the record level and at the item response level. Table 3 shows the results at the record level, i.e., after the statistical units have been flagged as outliers for having at least one item response identified as an anomaly. Table 4 shows the results at the item-response level, where the available data have been flagged as a cellwise anomaly.

Table 3: The record-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 $\theta = 0.08$. Sensitivity varies between 33% and 70%, and specificity varies between 69% and 96%.

| Survey | Level | Precision Regular | Precision Outlier | Recall Regular | Recall Outlier | Overall Accuracy |
|---|---|---|---|---|---|---|
| Cranberry | Low | 0.800 | 0.546 | 0.908 | 0.327 | 0.762 |
| Cranberry | High | 0.896 | 0.800 | 0.959 | 0.596 | 0.881 |
| Cattle | Low | 0.652 | 0.591 | 0.732 | 0.498 | 0.629 |
| Cattle | High | 0.766 | 0.768 | 0.839 | 0.675 | 0.767 |
| Ag. Yield | Low | 0.821 | 0.399 | 0.855 | 0.340 | 0.742 |
| Ag. Yield. | High | 0.875 | 0.696 | 0.924 | 0.567 | 0.841 |
| Sheep/Goats | Low | 0.605 | 0.618 | 0.693 | 0.523 | 0.610 |
| Sheep/Goats | High | 0.755 | 0.775 | 0.817 | 0.705 | 0.764 |

At the record level, the overall accuracy ranges between 0.61 and 0.88, the precision for detected outliers ranges between 0.40 and 0.80, and the recall of outliers ranges between 0.33 and 0.71. At the item response level, the overall accuracy ranges between 0.87 and 0.93, the precision for detected outliers ranges between 0.21 and 0.73, and the recall of outliers ranges between 0.21 and 0.57. Generally, the proposed methodology identifies outliers better at the record level for the surveys with smaller sample sizes. However, at the item level, the sample size does not appear to affect the performances, which have been generally better than those achieved at the record level. Furthermore, the precision for regular statistical units (i.e., for the records without cellwise outliers) has been larger than 0.6 at the record level, and larger than 0.9

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 and threshold $\theta = 0.08$. Sensitivity varies between 21% and 57%, and specificity varies between 93% and 98%.

| Survey | Level | Precision Regular | Precision Outlier | Recall Regular | Recall Outlier | Overall Accuracy |
|---|---|---|---|---|---|---|
| Cranberry | Low | 0.900 | 0.487 | 0.952 | 0.300 | 0.867 |
| Cranberry | High | 0.948 | 0.730 | 0.976 | 0.551 | 0.930 |
| Cattle | Low | 0.932 | 0.276 | 0.937 | 0.262 | 0.880 |
| Cattle | High | 0.961 | 0.609 | 0.966 | 0.574 | 0.933 |
| Ag. Yield | Low | 0.928 | 0.272 | 0.936 | 0.248 | 0.876 |
| Ag. Yield | High | 0.952 | 0.591 | 0.964 | 0.519 | 0.923 |
| Sheep/Goats | Low | 0.931 | 0.214 | 0.932 | 0.212 | 0.874 |
| Sheep/Goats | High | 0.962 | 0.537 | 0.960 | 0.551 | 0.928 |

at the item response level. The recall for regular statistical units has been larger than 0.73 at the record level, and larger than 0.93 at the item response level. The change in contamination level (from high to low) has affected more substantially the precision and recall of outliers with a drop of 20–35%; however, the precision and recall for regular units has remained quite stable with differences of 5–10%. The overall accuracy has also shown a similar behavior. In fact, the proposed method has been overall more accurate on datasets with higher contamination levels.

### 4.1 Proposed Method Compared to DDC

The DDC (Rousseeuw and Van den Bossche, 2018) was the first method developed to detect cellwise outliers in multivariate datasets by accounting for the correlations among variables. However, this method does not consider PRD. In fact, the use of these data is a key difference in the proposed approach. The performances of the two algorithms were compared on two datasets discussed above for the Agricultural Yields and Cattle Inventory.

Figure 1 shows the overall accuracies of both 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 proposed method is better suited for sparse matrices and uses all available data entries. The accuracy of the two methods were compared at the state level. In the Agriculture Yield dataset, the DDC algorithm has not provided the overall accuracies for six states due to the high level of sparseness. In the Cattle Inventory dataset, the DDC has not produced results for one state. In contrast, the proposed algorithm has identified anomalies in all states. Therefore, Figure 1 excludes the states where the DDC has not detected outliers. The upper left panel (a) was based on the "low" Cattle Inventory dataset, and the upper right panel (b) was based on the "high" one. The lower left panel (c) was based on the "low" Agriculture yield dataset, and the lower right panel (d) was based on the "high" one.

As shown by the graphs, the proposed method has correctly detected more outliers on both Cattle Inventory datasets and provided uniformly higher accuracies for all states than the DDC method. For the Agriculture Yield datasets, while both methods have similar accuracies in many states, the proposed method has outperformed the DDC in about 18 states. These results are

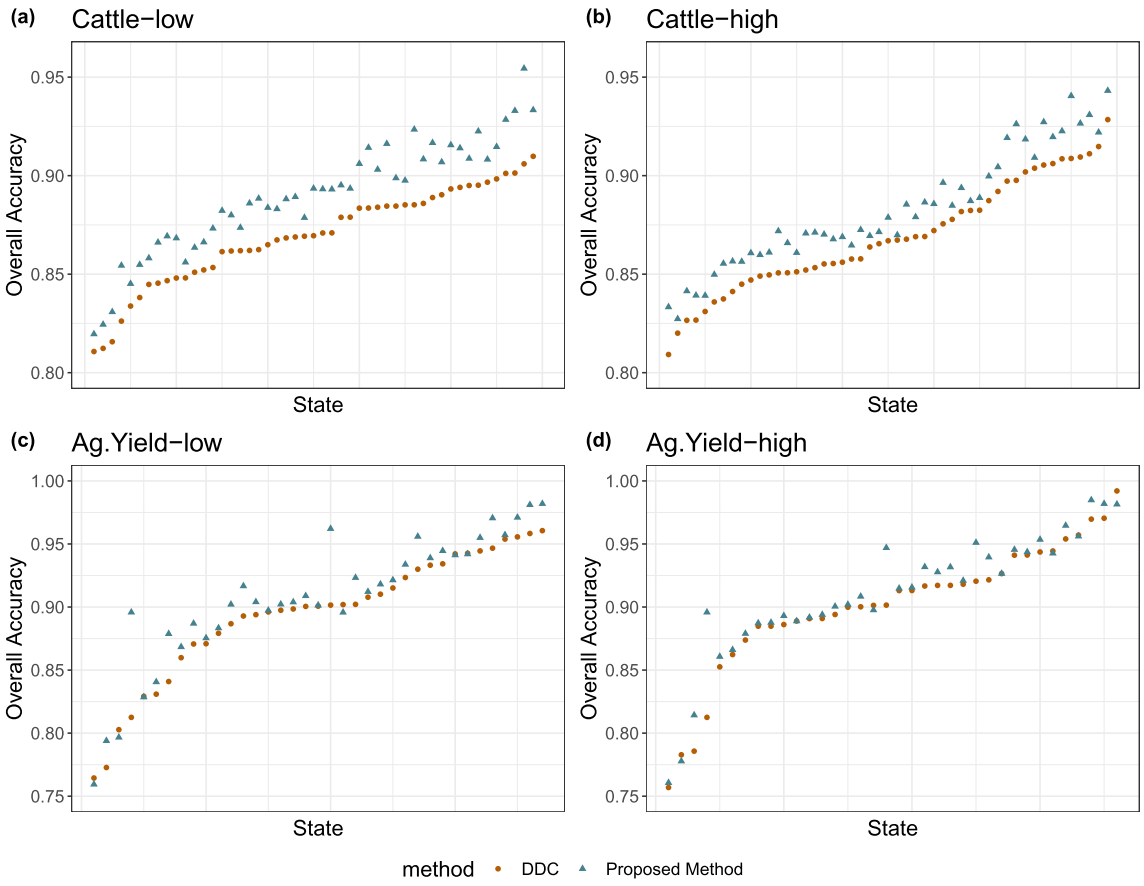


Figure 1: The state-level overall accuracy are obtained from the proposed method (in blue triangles) and the Detect-Deviating-Cells (DDC) method (in brown dots) on four synthetic datasets. The proposed method has correctly identified more cellwise outliers on the Cattle Inventory data due to a high availability of previously reported data (PRD). However, the two methods produced comparable results on the Agricultural Yields data due to a low availability of PRD. The states where the DDC method failed to provide results are excluded from the plots.

reasonable because the percentage of historical outliers in Cattle Inventory datasets is larger than the percentage in the Agricultural Yield datasets. Therefore, the proposed methodology detects more outliers because it uses additional information from the PRD.

In addition, the differences of the overall accuracies provided by the two methods in all states are higher in the "low" datasets than those in "high" datasets. The sum of the differences between the two methods in all available states for Cattle Inventory dataset is 0.985 for the "low" dataset and 0.678 for the "high" dataset. The sum of the differences in all available states for Agriculture Yield dataset is 0.486 for the "low" dataset and 0.339 for the "high" dataset. Recall that the "high" datasets have anomalous cells that are more likely identified as outliers and the "low" datasets have anomalous cells that are more difficult to detect. Therefore, the differences between "low" and "high" datasets demonstrate that the proposed methodology performs better than the DDC method on the datasets where the outliers are more difficult to detect.

## 5 Illustration on Real Data

The anomalous values in all datasets of the previous section have been synthetically generated. Although this approach has shown noticeable differences between the performances of the DDC method and the proposed algorithm, an open question remains. Can the proposed approach identify real anomalies on "raw" survey data acquired prior to the editing process? Therefore, we conducted further analyses with ground-truth anomalous data, as identified by agricultural experts through a manual revision process, to compare the two outlier-detection algorithms.

The data collected during 2022 for the Cattle Inventory have been considered for this illustration. These data coexist in a relational database with every other survey conducted by NASS. In this database, the original values from every respondent and all successive edits and updates are recorded with their respective timestamps. Therefore, it is possible to retrieve and compare both farmer-reported (pre-edit) and finalized (post-edit) values to identify the data entries that have been changed by manual revisions. The post-edit data for the 2021 Cattle Inventory are used as PRD when linked to the 2022 pre-edit data. Once the linkage between the datasets from these two years is completed, each state is separately processed to compare the proposed algorithm (with $\theta = 0.08$) to the DDC method.

Figure 2 shows the overall accuracies of both methods for each available state. The proposed method was substantially more accurate than the DDC method. In fact, it produced accuracies ranging from 0.76 to 0.91, whereas the DDC produced accuracies between 0.58 and 0.81. The overall accuracy of the proposed method in all states has been on average 21.9% higher than the overall accuracy of the DDC method. On average, the output of the proposed method coincides with the anomalies found through manual edits for 89% of all data entries. Based on the survey data considered in this study, the proposed method appears to be more accurate for the identification of cellwise outliers on real data.

From a computational point of view, the DDC is much faster than the proposed algorithm. However, it is important to notice that the DDC drops several variables and it does not

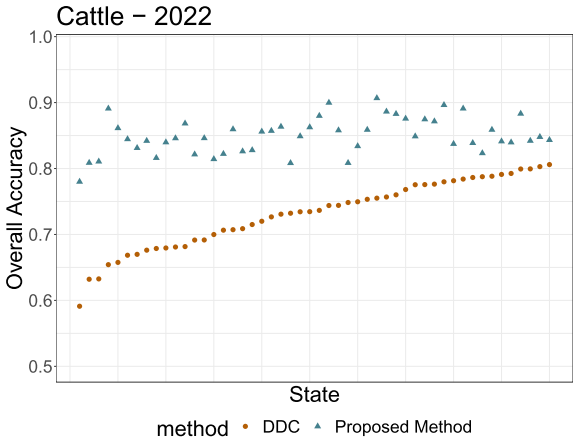


Figure 2: The state-level overall accuracy obtained from the proposed method (in blue triangles) and the Detect-Deviating-Cells (DDC) method (in brown dots). Ground-truth anomalies in the cattle data are identified based on nonzero differences between manually edited and nonedited cell values. The proposed approach has consistently identified correct anomalies with a higher accuracy than the DDC method.

account for different sources of data contamination. Both approaches have been implemented using parallel computing techniques, which are very useful in high-performance computational environments. In fact, the DDC method has processed each state in under 0.08 seconds and the proposed algorithm in less than 0.3 seconds. The proposed approach and the DDC method have, respectively, processed 1,702,069 cells in 4.2 and 1.26 seconds using 64 cores of an AMD EPYC 7V12 processor at 2.44 GHz and 440 GB of RAM on a Windows-Server-2019 virtual machine. It is worth mentioning that the code provided as an R package in the supplementary material can also run on a single-core CPU with less than 4GB of RAM without making further modifications.

## 6 Conclusion

A cellwise outlier detection technique based on fuzzy logic to identify four different types of anomalies was proposed. The first type of cellwise outliers considered in this article consisted of erroneous data with format inconsistencies. The second type referred to historical anomalies. The third type was traditionally known as a distribution-tail anomaly. The fourth type of cellwise outliers involved breaking of linear relationships among multiple variables. Furthermore, this article primarily focuses on identifying anomalies within an individual record rather than at an aggregated level, such as, county or state level. Therefore, sampling weights were not considered as they are not relevant for the developments in this article. The detection algorithm we developed for NASS agriculture survey data is general in nature (as described in Section 2) and can easily be extended to other survey data or even data collected through complete enumerations (such as a census).

The proposed algorithm is based on a distribution-free approach and the assumption that the first two central absolute moments exist. It can be applied to datasets that potentially suffer from the presence of cellwise anomalies, skewed distributions (with positive support), missing values, and multivariate relationships. It effectively copes with sparse and missing data by accounting for zero inflation without removing entire records and/or variables with missing values. Typically, the algorithm provides informative outputs comprising of 1) a numerical score for each available entry and 2) a binary identifier (or flag) based on a user-defined percentage of contamination. Alternative procedures to the percentage of contamination, $\theta$, can also be valid (as discussed in Appendix A). However, it is common to set $\theta$ based on historical observations or reasonable assumptions made by the analysts. For instance, the fifth percentile of the final scores can be used as a cut-off value (or threshold) that can separate the outliers from regular observations. Cells with their corresponding final scores smaller than the $100\theta$ percentile are classified as cellwise outliers. This approach allows for the identification of anomalous entries even in more extreme scenarios (e.g., when $\theta = 0.005$).

The performance of the proposed algorithm has been illustrated using NASS livestock and crop survey data with randomly generated anomalies. Our simulation study considered four different datasets and showed that the algorithm provides accurate and robust results when detecting cellwise outliers. Moreover, comparisons using real data with PRD illustrate that the proposed approach has generally higher overall accuracy than the DDC method. Relatively large datasets (with over 20,000 statistical units) have been processed within 5 seconds using 64 cores of an AMD EPYC 7V12 Processor at 2.44 GHz and 440 GB of RAM on a virtual machine with Windows Server 2019 Datacenter. When PRD are not available, the proposed algorithm is comparable (or even equivalent) to the DDC method. However, as an advantage, our algorithm

is designed to identify cellwise outliers without dropping records or variables with many missing values (as it is the case for the DDC algorithm).

Lastly, the use of model-based predictions in lieu of PRD allows for the application of the proposed algorithm with data collected under different scenarios. For instance, when time series or longitudinal data are regularly acquired, or other data are collected through complete enumerations (such as two or more consecutive censuses). Even if this article focused on a simple time-series model, other models can be used to leverage administrative, structured, or unstructured data available for the whole or a subset of the surveyed records.

## Supplementary Material

## A Reasoning on Fuzzy Logic and Probabilistic Inequalities

In traditional logic, a statement can be either true or false. For detecting cellwise outliers, the statement $\mathrm{G}_{ij}$ = "the cell $(i, j)$ is regular" is true if all following statements are true: $\mathrm{E}_{ij}^{[B]}$ = "the cell $(i, j)$ is not a data-format (or bit-flip) outlier"; $\mathrm{E}_{ij}^{[H]}$ = "the cell $(i, j)$ is not an historical outlier"; $\mathrm{E}_{ij}^{[T]}$ = "the cell $(i, j)$ is not a distribution-tail outlier"; and $\mathrm{E}_{ij}^{[R]}$ = "the cell $(i, j)$ is not a relational outlier". Therefore, if at least one of the four statements above is false, the statement $\mathrm{G}_{ij}$ is false. The truthfulness of $\mathrm{G}_{ij}$ is determined from the truth table of the $\mathrm{E}_{ij}^{[\cdot]}$ based on the following logical expression:

$$\begin{aligned} \mathrm{G}_{ij} &= \mathrm{E}_{ij}^{[H]} \wedge \mathrm{E}_{ij}^{[R]} \wedge \mathrm{E}_{ij}^{[T]} \wedge \mathrm{E}_{ij}^{[B]} \\ &= \neg \left( \neg \mathrm{E}_{ij}^{[H]} \vee \neg \mathrm{E}_{ij}^{[R]} \vee \neg \mathrm{E}_{ij}^{[T]} \vee \neg \mathrm{E}_{ij}^{[B]} \right), \end{aligned} \tag{6}$$

where the symbols $\wedge$, $\vee$, $\neg$ represent the logical operators *And*, *Or*, and *Negation*, respectively. The scores $s_{ij}^{[\cdot]} \in [0, 1]$ provide a degree of truthfulness for the statements $\mathrm{E}_{ij}^{[\cdot]}$, where zero and one indicate if a statement is certainly false or true, respectively. Therefore, (6) is described using fuzzy logic operators (such as t-norm, t-conorms, and fuzzy complement for *And*, *Or*, and *Negation*, respectively) that satisfy De Morgan laws. For example, the fuzzy complement, $C(z) = 1-z$, and the product t-norm, $T(z_1, z_2) = z_1 z_2$, can be used to formulate the t-conorm, $S(z_1, z_2) = 1 - (1 - z_1)(1 - z_2)$. One can explore the use of different t-norms; however, the product t-norm consistently provided the most accurate results. The literature provides alternative formulations of t-norms and conorms that may not satisfy De Morgan laws (Gupta and Qi, 1991). These alternative formulations are disregarded for the purposes of this study.

Ideally, one would use the cumulative distribution function (CDF) of $X_{ijt}$, $F_{X_{ijt}}(\cdot)$, to quantify the degree of truthfulness of $\mathrm{G}_{ij}$ (see (6)). However, there are two issues with pursuing this approach. First, each individual cell in the dataset is the only realization available from the random process associated with that cell. Second, the CDF is unknown. Therefore, univariate or multivariate analyses of past observations and other records within the same stratum are needed to "standardize" each cell.

Typically, the Bienaymé-Chebyshev's inequality is used to identify probabilistic bounds for $X_{ijt}$ based on the following:

$$\Pr\left(|X_{ijt} - \mu_\delta| \geqslant \epsilon\right) = F_{X_{ijt}}(\mu_\delta - \epsilon) + 1 - F_{X_{ijt}}(\mu_\delta + \epsilon) \leqslant \epsilon^{-\delta} \mathsf{E}\left[|X_{ijt} - \mu_\delta|^\delta\right],$$

where the right-hand side, $\epsilon^{-\delta}\mathsf{E}[|X_{ijt} - \mu_\delta|^\delta]$, is a monotonically decreasing function on $\epsilon > 0$, for $\delta \geqslant 1$ given that $\mathsf{E}[|X_{ijt} - \mu_\delta|^\delta]$ is finite. However, to study the distribution of $X_{ijt}$, one would

need several realizations $x_{ijtu}$ from $U$ parallel universes, where $u = 1$, by convention, corresponds to a value in the dataset, and $u \in \{2, 3, \ldots\} \subset \mathbb{N}$ corresponds to non-accessible values from other universes. For the selection of $\epsilon = g(x)$, the probability $\Pr(|X_{ijt} - \mu_\delta| \geqslant g(x))$ depends on $g(x)^{-\delta}$.

If a user decides to operate with $g(x) = |x - \mu_\delta|\xi^{-1}$, where $\xi \in [1, +\infty)$ is often set to either 1.5 or 3, then the alternative score

$$s_{ijt} = \min\left\{1,\ \xi^{\delta}\ \frac{\mathsf{E}\left[|X_{ijt} - \mu_\delta|^{\delta}\right]}{|x_{ijt} - \mu_\delta|^{\delta}}\right\}^{1/\delta} \tag{7}$$

is more likely to be one for regular cells and less than one for outliers. When the expression in (7) is used to determine the truthfulness of the statements $\mathrm{E}_{ij}^{[\cdot]}$, the statement $\mathrm{G}_{ij}$ would be considered true if

$$T\left(s_{ijt}^{[H]}, T\left(s_{ijt}^{[R]}, T\left(s_{ijt}^{[T]}, s_{ijt}^{[B]}\right)\right)\right) = 1,$$

and false otherwise. In this case, any t-norm that satisfies the De Morgan laws could be used to determine the truthfulness of $\mathrm{G}_{ij}$.

## B Median Algorithm

The details of the median algorithm developed for implementing the cellwise outlier detection via fuzzy logic are provided as pseudo-code shown in Algorithm 1. The algorithm takes an input vector $\mathbf{v} \in \mathbb{R}^n$, and it returns an approximated median value when one of two stopping criteria (i.e, error tolerance, $\eta = 10^{-16}$, and maximum number of iterations, $M = 2{,}000$) are satisfied.

The algorithm computes the minimum and the maximum values the input vector $\mathbf{v}$. These two operations can be merged in a single loop iterating over the $n$ components of the input vector. The maximum and minimum of $\mathbf{v}$ partition the observed range of the data into $A$ equal-width bins, where $A \approx \sqrt{n}$. A bin is used to store the fraction of data falling within it, $f_\alpha \in [0, 1]$, for any $\alpha = 1, \ldots, A$. The algorithm iterates over the $n$ observations to compute the fraction of data falling in each bin (i.e., with time complexity $O(n)$). Then, it identifies the bin containing the median with at most $A$ iterations over the bins (i.e., with time complexity $O(\sqrt{n})$ in worst case scenario). These two loops are nested in an outer loop; thus, the time complexity is proportional to $O(n) + O(\sqrt{n}) = O(n)$. Furthermore, the algorithm converges to the solution at the rate $o(n^{-1/2})$, and hence, the outer loop requires less iterations as the sample size increases.

## Acknowledgement

The findings and conclusions in this article 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 Linda J. Young, the associate editor, and two anonymous reviewers for providing comments that improved this article.

## Funding

This research was supported by the intramural research program of the US Department of Agriculture, National Agricultural Statistics Service (NASS).

**Algorithm 1** Pseudo-code of the proposed median algorithm.

$\eta \leftarrow 10^{-16}$ ▷ Error tolerance (stopping criterion 1)
$M \leftarrow 2000$ ▷ Maximum number of iterations (stopping criterion 2)
**function** MEDIAN($n \in \mathbb{N}$, $\mathbf{v} \in \mathbb{R}^n$)
  **if** $n < 1$ **then return** NAN
  **else if** $n = 1$ **then return** $v_1$
  **else if** $n = 2$ **then return** $\frac{1}{2}(v_1 + v_2)$
  **end if**
  $A \leftarrow 2^{\lceil \frac{1}{2} \log_2 n \rceil}$ ▷ Compute $A$ histogram bins such that $A \approx \sqrt{n}$
  $u_1 \leftarrow \min(\mathbf{v})$
  $u_2 \leftarrow \max(\mathbf{v})$
  $m \leftarrow 0$
  **repeat**
    $\rho \leftarrow A/(u_2 - u_1)$ ▷ Compute the inverted width of each bin
    $\mathbf{f} \leftarrow \mathbf{0}$ ▷ Initialize a frequency vector $\mathbf{f} = (f_1, \ldots, f_A)^\top$ to zero
    **for all** $i \in \{1, 2, \ldots, n\}$ **do** ▷ Use atomic instructions if this loop is made parallel
      $\alpha \leftarrow \lceil \rho(v_i - u_1) \mathbb{1}\{v_i \geqslant u_1\} \rceil$ ▷ Find bin to update
      $\alpha \leftarrow \alpha + \mathbb{1}\{\alpha = 0\}$ ▷ Fix zero values of $\alpha$
      $f_\alpha \leftarrow f_\alpha + \frac{1}{n}$ ▷ Update the histogram frequencies
    **end for**
    $\alpha \leftarrow 2$ ▷ Search the bin containing the median
    **while** $f_1 < \frac{1}{2}$ and $\alpha \leqslant A$ **do**
      $f_1 \leftarrow f_1 + f_\alpha$ ▷ Compute cumulative frequencies until they exceed $\frac{1}{2}$
      $\alpha \leftarrow \alpha + 1$ ▷ Track the position of the next bin to process
    **end while**
    $\alpha \leftarrow \alpha - 2$
    $\rho \leftarrow 1/\rho$ ▷ Compute the width of the selected bin
    $u_1 \leftarrow u_1 + \alpha\rho$ ▷ Update the lower bound of the interval containing the median
    $u_2 \leftarrow u_1 + \rho$ ▷ Update the upper bound of the interval containing the median
    $m \leftarrow m + 1$ ▷ Update the number of iterations performed
  **until** $\rho \leqslant \eta$ or $m \geqslant M$ ▷ Stop the loop if either condition is satisfied
  **return** $u_1$
**end function**

## References

Agostinelli C, Leung A, Yohai VJ, Zamar RH (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24(3): 441–461. https://doi.org/10.1007/s11749-015-0450-6

Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1): 311–331. https://doi.org/10.1214/07-AOS588

Bienaymé IJ (1867). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Journal de Mathématiques Pures et Appliquées*, 2(12): 158–176.

Chepulis MA, Shevlyakov G (2020). On outlier detection with the Chebyshev type inequalities. *Journal of the Belarusian State University. Mathematics and Informatics*, 3: 28–35.

https://doi.org/10.33581/2520-6508-2020-3-28-35

Dagum L, Menon R (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Computational Science and Engineering*, 5(1): 46–55. https://doi.org/10.1109/99.660313

Daniel JW, Gragg WB, Kaufman L, Stewart GW (1976). Reorthogonalization and stable algorithms for updating the Gram-Schmidt $QR$ factorization. *Mathematics of Computation*, 30(136): 772–795. https://doi.org/10.1090/S0025-5718-1976-0431641-8

De Waal T, Pannekoek J, Scholtus S (2011). *Handbook of Statistical Data Editing and Imputation*, volume 563. John Wiley & Sons.

Filzmoser P, Gregorich M (2020). Multivariate outlier detection in applied data analysis: Global, local, compositional and cellwise outliers. *Mathematical Geosciences*, 52(8): 1049–1066. https://doi.org/10.1007/s11004-020-09861-6

Flynn M (1966). Very high-speed computing systems. *Proceedings of the IEEE*, 54(12): 1901–1909. https://doi.org/10.1109/PROC.1966.5273

Gupta MM, Qi J (1991). Theory of t-norms and fuzzy inference methods. *Fuzzy Sets and Systems*, 40(3): 431–450. https://doi.org/10.1016/0165-0114(91)90171-L

Hampel FR (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346): 383–393. https://doi.org/10.1080/01621459.1974.10482962

Heydarian M, Doyle TE, Samavi R (2022). MLCM: Multi-label confusion matrix. *IEEE Access*, 10: 19083–19095. https://doi.org/10.1109/ACCESS.2022.3151048

Hidiroglou MA, Berthelot JM (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12(1): 73–83.

Huber PJ, Ronchetti EM (1981). *Robust Statistics*. John Wiley & Sons, New York.

Miller D, Robbins M, Habiger J (2010). Examining the challenges of missing data analysis in phase three of the agricultural resource management survey. *JSM Proceedings. American Statistical Association Section on Survey Research Methods.*

O'Gorman TJ (1994). The effect of cosmic rays on the soft error rate of a DRAM at ground level. *IEEE Transactions on Electron Devices*, 41(4): 553–557. https://doi.org/10.1109/16.278509

Raymaekers J, Rousseeuw PJ (2019). Handling cellwise outliers by sparse regression and robust covariance. arXiv preprint: https://arxiv.org/abs/1912.12446.

Raymaekers J, Rousseeuw PJ, Van den Bossche W, Hubert M (2023). `cellWise`*: Analyzing data with cellwise outliers*. CRAN, R package version 2.5.2.

Rousseeuw PJ, Van den Bossche W (2018). Detecting deviating data cells. *Technometrics*, 60(2): 135–145. https://doi.org/10.1080/00401706.2017.1340909

Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3): 581–592. https://doi.org/10.1093/biomet/63.3.581

Sandqvist AP (2016). Identifizierung von Ausreissern in eindimensionalen gewichteten Umfragedaten. *KOF Analysen*, 2016(2): 45–56.

Sedgewick R (1978). Implementing quicksort programs. *Communications of the ACM*, 21(10): 847–857. https://doi.org/10.1145/359619.359631

Stigler SM (1973). The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3): 472–477.

Tchebichef P (1867). Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, 2(12): 177–184.

Zwillinger D (2018). *Standard Mathematical Tables and Formulas*. CRC Press.