Knowl Inf Syst
DOI 10.1007/s10115-014-0760-0

REGULAR PAPER

A masking index for quantifying hidden glitches

Laure Berti-Équille · Ji Meng Loh · Tamraparni Dasu

Received: 31 December 2013 / Revised: 2 May 2014 / Accepted: 18 May 2014

Abstract Data glitches are errors in a dataset. They are complex entities that often span multiple attributes and records. When they co-occur in data, the presence of one type of glitch can hinder the detection of another type of glitch. This phenomenon is called *masking*. In this paper, we define two important types of masking and propose a novel, statistically rigorous indicator called *masking index* for quantifying the hidden glitches. We outline four cases of masking: outliers masked by missing values, outliers masked by duplicates, duplicates masked by missing values, and duplicates masked by outliers. The masking index is critical for data quality profiling and data exploration. It enables a user to measure the extent of masking and hence the confidence in the data. In this sense, it is a valuable data quality index for choosing an anomaly detection method that is best suited for the glitches that are present in any given dataset. We demonstrate the utility and effectiveness of the masking index by intensive experiments on synthetic and real-world datasets.

Keywords Anomaly detection · Masking · Duplicate record identification · Missing values · Outlier detection

L. Berti-Équille
IRD ESPACE DEV, 500, rue J.F. Breton, Montpellier, France

L. Berti-Équille (✉)
Qatar Computing Research Institute, Tornado Tower, 18th Floor, West Bay, Doha, Qatar
e-mail: lberti@qf.org.qa

J. M. Loh
Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, USA

T. Dasu
AT&T Labs-Research, Bedminster, NJ, USA

Published online: 01 July 2014

1 Introduction

Data glitches are errors in the data that can significantly impact the analysis, and conclusions drawn from the data. They occur in a wide variety of ways, ranging from human error (e.g., typos and duplicate entries), to software and hardware problems (e.g., missing values due to transmission failure). As data become more structurally complex and heterogeneous, the gathering, storing and monitoring of data become dependent on intricate systems of interconnected hardware and software. There are numerous opportunities for data to get corrupted at each of these stages, introducing a daunting variety and quantity of glitches into the data in complex and interrelated patterns.

Financial data streams, communication network data, social data, and scientific data, almost all real-world data suffer from missing values, incomplete and distorted values, inconsistent values, duplicate records, and outliers, to mention just a few possible ways that data can go bad. These glitches do not occur randomly or in small proportions. They often touch very specific sections of the data, introducing biases into the analysis of the remaining data. The glitches also occur in patterns, as in overloaded network devices with extremely high (outlying) loads that result in outages (missing values). Previous work of [4] has addressed and formalized the definition of complex glitches and glitch patterns. Sometimes, one type of glitch makes the other undetectable. For instance, when there are missing values, we might not be able to detect duplicates or true outliers.

When the presence of one type of glitch *masks* another type of glitch and impedes its detection, it can have far-reaching consequences. Masking could result in underestimating the number of glitches and consequently the cost of cleaning the data. It could also give a false confidence in the results of data analysis. For instance, if the masked glitches occur non-randomly in a systematic pattern, they could seriously bias the conclusions drawn from the analysis.

In the past, masking was discussed specifically in the context of outlier detection. Masking, along with the related notion of *swamping* where outliers are duplicated to such an extent that they dominate the distribution and "normal" values become outliers, have been proposed in [3]. In [1], the authors give an intuitive understanding of these effects. Additional references include [2,7,9,10].

In our work, we focus on masking but generalize the definition to apply widely to any type of glitch, considerably expanding the scope of previous work beyond outlier detection. We provide a mathematical definition of masking and propose a statistically rigorous method for quantifying masking through a *masking index*.

The masking index is a critical tool for data quality profiling, data exploration, preparation, and mining. It serves two important purposes.

- It enables us to quantify the "hidden" glitches in data and estimate the confidence we have in the analysis results derived from the data.
- It enables us to empirically choose a best glitch detection method when there are multiple glitches in a dataset, which is frequently the case in real-world data.

An interesting consequence of masking is that cleaning one set of glitches can reveal (or "unmask") other glitches. For instance, imputing missing values can create "new" duplicate records and outlying values. Or, removing outliers can introduce new duplicates. In such a setting, where there is a need for an iterative approach to data cleaning, the masking index plays an important role in determining the cleanliness of the data, determining the best strategies for cleaning and preparing the data, and serves as an objective stopping criterion for iterative cleaning.

Our original contributions can be summarized as follows:

- We propose a general definition of masking that applies widely to any glitch type, including multi-record glitches such as outliers and duplicate records, and univariate glitches such as missing values;
- We define two distinct notions of masking, *inner*, and *outer* masking;
- We define a novel, statistically rigorous indicator called *masking index* for quantifying the extent of masking in the dataset;
- We provide a framework for empirically evaluating the masking index in four cases of masking;
- We propose a method for estimating the masking index in two scenarios: (1) when the ground truth is known, based on synthetic data, and (2) the case of real-world data where the ground truth is not available.

Our framework for computing the masking index scales effectively to big datasets.

The rest of the paper is organized as follows. In Sect. 2, we introduce an illustrative example to explain masking. In Sect. 3, we define the problem of masking, introduce the notations, and describe the main characteristics of masking. In addition, we present two different types of masking. In Sect. 4, we formally define the masking index. We also instantiate the theoretical and conceptual formulation of the index in four cases of masking. In Sect. 5, we demonstrate the validity of our approach on synthetic datasets where we control the occurrence of glitches. This allows us to empirically estimate the masking index and study its canonical behavior. In Sect. 6, we discuss computing the masking index in real-life scenarios where the ground truth is not known. The theoretical formulations of Sect. 4 combined with re-sampling from clean parts of the real-world data allow us to compute the masking index. We demonstrate the utility of our contributions on publicly available datasets. Finally, in Sect. 7, we discuss existing literature that is relevant to this paper. In Sect. 8, we summarize the salient points of the paper and outline future work.

2 An illustrative example

To illustrate the complexity of data glitches and the problem of *masking*, consider the dataset available at OpenData by Socrata[1] which contains information about Canadian unclaimed bank accounts at branches in the Edmonton area or registered to addresses in the Edmonton area. The Web site claims that the total of such abandoned accounts amounts to more than 7 million Canadian dollars.

Here, we consider a subset of 20 records (see Table 1). Each record contains the following information about the banking accounts: business or last name (B/LN), client first name (FN), balance in Canadian dollars (BL), address (AD), city (CY), last transaction date (LT), and bank name (BN).

A simple sum of the unclaimed money in these 20 accounts is 15,542.10 Canadian dollars. However, notice that there are missing values encoded with various notations (e.g., NULL, _, UNKNOWN, ??, and blanks), duplicate records, and outlying or suspicious values in these 20 records. For example, records $x_{18.}$ and $x_{20.}$ appear to be duplicates, with the same transaction dates and very similar locations. The sum of money involved in these two records are identical except for a missing decimal point in $x_{18.}$. This in turn results in a large value of 10, 712 for the balance in cell $x_{18,3}$. Hence, what may be true outliers $x_{4,3} = 1{,}675.07$ and $x_{14,3} = 1{,}627.5$

[1] Data set from OpenData by Socrata retrieved on March 26, 2013: https://opendata.socrata.com/Government/Unclaimed-bank-accounts/.

Table 1 A subset of 20 records taken from the "Unclaimed Bank Accounts" Data Set from OpenData by Socrata[a] with examples of missing values, duplicates and outlying values

x_{ij}	$x_{.1}$ B/LN	$x_{.2}$ FN	$x_{.3}$ BL	$x_{.4}$ AD	$x_{.5}$ CY	$x_{.6}$ LT	$x_{.7}$ BN
$x_{1.}$	Canadian	Bernd/Jane	30.00	Box 36 Site 6 RR 2	Thorsby AB	10/22/1994	Bank of Montreal
$x_{2.}$	Canadian	Bernd/Jane	30.00	Box 36 Site 6 RR 2	Thorsby AB	10/19/1994	Bank of Montreal
$x_{3.}$	Trust Ac		1,675.00	McKenney	St. Albert	11/30/1993	
$x_{4.}$	Trust Am		1,675.07	ST Albert Trail	St. Albert	11/30/1993	Toronto-Dom Bank
$x_{5.}$	Bruno	Dakota H	5.02	_	_	_	Bank of Nova Scotia
$x_{6.}$	Bruno	Daniel S	5.02		M1M 1M1	10/30/1992	Bank of Nova Scotia
$x_{7.}$	Bruno	Grant C	5.02	UNKNOWN	UNKNOWN	10/30/1992	Bank of Nova Scotia
$x_{8.}$	Broderick	Margaret	122.91	20 OAK ST	Sherwood Park	12/21/1995	Imp Bank of Com
$x_{9.}$	Broderick	Margaret	107.88	20 OAK ST	Sherwood Park	12/22/1995	Imp Bank of Com
$x_{10.}$	Murphy	Doyle	0.01	34 Woodvale	AB	10/07/1992	Bank of Nova Scotia
$x_{11.}$	Murphy	Megan	0.01	34 Woodvale	AB	10/07/1992	Bank of Nova Scotia
$x_{12.}$	Quintal	Dani	0.01	RR 1	Calahoo AB	05/09/1991	Bank of Nova Scotia
$x_{13.}$	Quintal	Megan	0.01	165 Woodbuffalo way	Ft McMurray AB	05/09/1991	Bank of Nova Scotia
$x_{14.}$	Young	Musicians Academy	1,627.5	??	??	01/03/1975	Royal Bank of Can
$x_{15.}$	Young	Musicians Academy	76.06	NULL	NULL	01/04/1975	Royal Bank of Can
$x_{16.}$	Zittlaw	Edward	410.27			07/02/1988	Royal Bank of Can
$x_{17.}$	Zittlaw	Edward	341.53			02/07/1988	Royal Bank of Can
$x_{18.}$		Bush*W*J	10,712.00	Jasper Ave NW	Edmonton	04/18/1986	Toronto-Dom Bank
$x_{19.}$	Bush	William	8.66			08/16/1986	Montreal Trust
$x_{20.}$	William	*Bush*Jack	107.12	10230 Jasper Ave	Edmonton	04/18/1986	Toronto-Dom Bank
SUM			15,542.10				

[a] see Footnote 1

become masked, while the possibly legitimate values of 0.01 in $x_{10,3}$, $x_{11,3}$, $x_{12,3}$, and $x_{13,3}$ become swamped. Other glitches may be more subtle. For example, `St Albert Trail` and `McKenney` are actually the same location and may mask the fact that $x_{3,}$ and $x_{4,}$ are duplicate accounts.

The process of cleaning the data and computing a total balance is complex. Depending on how duplicates and outliers are treated, e.g., taking the mean account balance or taking the most recent transaction of duplicate accounts, different total balances may be obtained. Figuring out the exact strategy to clean this dataset and obtain a realistic total sum of money is beyond the scope of this paper and is part of our future work on iterative cleaning. This example clearly motivates the need for estimating the number of hidden glitches and understanding how glitch detection can be affected by the masking phenomenon.

3 The masking problem

Suppose that the dataset X is an $N \times V$ matrix, with N records and V variables and that there are K different types of glitches of interest. In the rest of the paper, we consider three types of glitch, namely: missing values, outlying values, and duplicate records.

For each $x_{ij} \in X$, we define a glitch vector $\mathbf{g}_{ij}$ with K elements, where each element $g_{ijk}(k = 1, \ldots, K)$ is 1 if x_{ij} belongs to a glitch of type k, and 0 otherwise. Hence, $\mathcal{G}$ is a $N \times V \times K$ array. The array $\mathcal{G}$ represents the true occurrence of glitches in the data X, and we refer it as the *ground truth*.

Further, we define $\mathcal{G}'$ to be the array that results from applying glitch detection methods to X. In a world with perfect detection methods, $\mathcal{G} = \mathcal{G}'$, but in reality, the matrix of comparisons between real and detected glitches, $\mathcal{G}' - \mathcal{G}$, contains elements of 0, 1 and -1. An element of 0 means a correct detection (true positive) or a correct non-detection (true negative). An element of 1 means there is a false detection (false positive). An element of -1 means that there is a false non-detection (false negative).

Masking arises when we are not able to detect a glitch due to the presence of another. Non-detection can happen in three ways. First, non-detection may be due to a lack of statistical power of a detection method. Second, the power of the detection method of glitch type k may be reduced by the presence of glitch type k'. The size of both these effects depend on the specific detection method used. The third possible cause is a direct effect of glitch type k' on glitch type k, independent of the detection method used. We describe these effects in the following subsections.

3.1 Power of detection

Glitch detection methods vary in their ability to detect glitches. One measure of performance of a detection method m is its *statistical power*, $\pi_m \in [0, 1]$. Traditionally, power is defined in the context of clean data. Suppose that the data is completely clean except for a single glitch of type k in x_{ij} and that we use method m to detect it.

Definition 3.1 The statistical power of method m for detecting glitch type k is the probability that the glitch is detected. It is given by:

$$\pi_{m,k} = P\left(\mathcal{G}'_{ijk} = 1 | \mathcal{G}_{ijk} = 1\right).$$

Except for cases where the detection is absolute (e.g., the detection of missing values), a detection method generally has power $0 < \pi_{m,k} < 1$ even when applied to clean data due to

random error. However, the presence of other types of glitches may interact with the detection method, resulting in a change in power. Let $\epsilon_{m,k,k'}$ be the change in the power of method m under the influence of other glitches of type k'. We will assume that $0 \leq \epsilon_{m,k,k'} \leq \pi_{m,k}$ though it might be possible for one type of glitch to improve power of detection of another type of glitch. Intuitively, a detection method that is not affected or only slightly affected by the presence of other data glitches, i.e., $\epsilon_{m,k,k'} \approx 0$, is said to be robust to data glitches of type k'. The altered power is given by:

$$\begin{aligned}\pi_{m,k,k'} &= P\left(\mathcal{G}'_{ijk} = 1 | \mathcal{G}_{i'j'k'} = 1 \wedge \mathcal{G}_{ijk} = 1, i \neq i'\right) \\ &= \pi_{m,k} - \epsilon_{m,k,k'}.\end{aligned}$$

We formalize this notion below.

Definition 3.2 The robustness $\rho_{m,k,k'}$ of method m in detecting glitches of type k in the presence of glitches of type k' is the ratio of the altered power $\pi_{m,k,k'}$ in the presence of glitches of type k' to $\pi_{m,k}$, the power of detection without glitches of type k'. That is,

$$\rho_{m,k,k'} = \frac{\pi_{m,k,k'}}{\pi_{m,k}}.$$

The closer $\rho_{m,k,k'}$ is to 1, the more robust m is in detecting glitches of type k in the presence of glitches of type k'. To simplify notation, we will henceforth drop the subscript m in all expressions without losing clarity or generality and refer to these quantities as π_k, $\epsilon_{k,k'}$, and $\rho_{k,k'}$ instead, for a given method.

3.2 Inner and outer masking

There is a fundamental connection between masking and power. Power is a measure of the ability to detect, while masking is exactly the opposite. The connection between the two can be expressed as:

$$\Pi = 1 - \mathcal{M} \tag{1}$$

where Π is power of detection and $\mathcal{M}$ the probability of masking. We explain further in the following sections.

First, it is useful to make a distinction between two basic types of masking. A glitch of type k' affecting the data cell x_{ij} could mask a glitch of type k in the same record i, or in a different record i'. This leads to the notion of two types of masking, *inner* and *outer masking*.

Before defining them, we illustrate them by considering the schematic depiction in Fig. 1. First, consider the case where only glitches of type k (i.e., outliers represented as circles) are present. Most of them are detected in the absence of glitches of type k' (e.g., missing values represented as diamonds or duplicate records represented as gray rectangles) as indicated by the large brace on the left. Furthermore, there are some glitches of type k that are not detected if the detection method's power is less than one.

Now suppose that we introduce glitches of type k' (diamonds and rectangles for missing values and duplicates). The method fails to detect some of the glitches of type k (black circles) that it could detect earlier. The braces on the left highlight the two possibilities. When glitches of type k' are present in the same record i as glitch k, and the method is unable to detect a glitch of type k that it could detect earlier, we call this phenomenon *inner masking*. When the glitch k' is in another record, and the method still fails to detect glitch k, it is called *outer masking*. We now define these two notions of masking below.

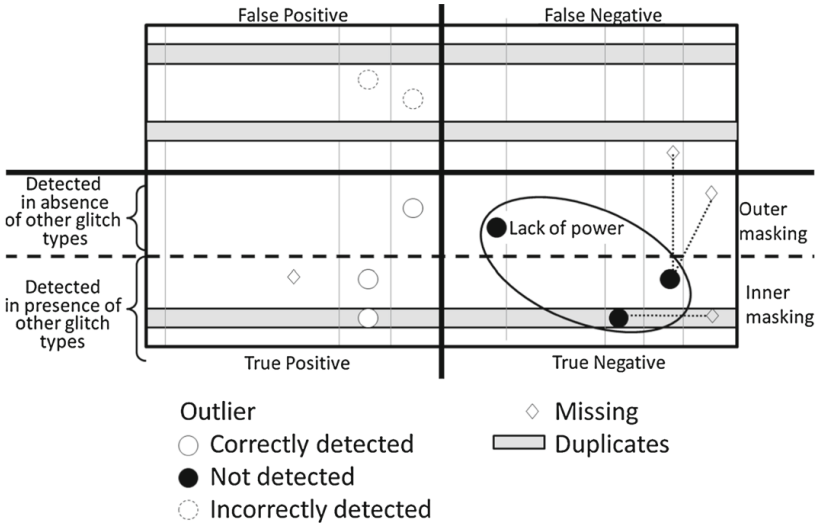


Fig. 1 Inner and outer masking: outliers (*circles*) that are detectable (*clear*) become undetectable (*black circles*) with the introduction of missing values (*diamonds*) and duplicate records (*gray rectangles*). If different glitch types are in the same record, this phenomenon is called *inner masking*, and if they are in different records it is called *outer masking*. In addition, a detection method would also result in true negatives (shown inside *ellipse* on *top left* of the figure) and false positives (*dashed circles*)

Definition 3.3 *Inner masking* is the phenomenon where the detection of a glitch of type k' prevents the detection of a glitch of type k in the same cell or record. That is, $\mathcal{G}'_{ijk'} = 1 \implies \mathcal{G}'_{ijk} = 0$ even though $\mathcal{G}_{ijk} = 1$. The glitches could be of the same type or of different types. The extent of inner masking is given by:

$$\mathcal{M}(k/k')|_{\text{inner}} = P\left(\mathcal{G}'_{ijk} = 0 | \mathcal{G}'_{ij'k'} = 1 \wedge \mathcal{G}_{ijk} = 1\right). \tag{2}$$

In the probability statement above, we made a simplifying assumption:

Assumption 1 If glitches of type k and k' are present on the same record i, then k is always masked.

For example, outliers will always be masked if there is a missing value present in the cell. This suggests the notion of *glitch dominance*.

Definition 3.4 Glitch type k' dominates glitch type k if

$$\mathcal{G}'_{ijk'} = 1 \implies \mathcal{G}'_{ij'k} = 0, \quad \forall i, j, j',$$

so that

$$\mathrm{P}(\mathcal{G}'_{ijk} = 1 | \mathcal{G}'_{ij'k'} = 1 \wedge \mathcal{G}_{ijk} = 1) = 0, \quad \forall i, j, j'.$$

Relaxing Assumption 1 requires splitting the expression in Eq. 2 by the conditional probability of detecting k in the presence of k' in the same record.

In addition to affecting glitch detection in the same record, glitches could interfere with the detection of glitches in other records. We define the notion of *outer masking* below.

Definition 3.5 *Outer masking* is the phenomenon where the detection of glitches of type k in one record is hindered by the occurrence of a glitch of type k' in other records. The glitches could be of the same type or of different types.

$$\mathcal{M}(k/k')|_{\text{outer}} = P\left(\mathcal{G}'_{ijk} = 0 | \mathcal{G}_{i'j'k'} = 1 \wedge \mathcal{G}_{ijk} = 1\right) \tag{3}$$

with $i \neq i'$.

Again, the probability statement involves a simplifying assumption:

Assumption 2 If a glitch of type k' is present, we will always detect it. That is, there exists a method m such that $\pi_{m,k'} = 1$ and hence we can assume that $\mathcal{G}'_{ijk'} = \mathcal{G}_{ijk'}, \forall i, j$ for method m.

Missing values are an excellent example of glitches that could be detected with certainty. Relaxing Assumption 2 requires an additional step in Eq. 3 of conditioning on the power of detection k'. Note that we do not require these assumptions to do empirical studies and estimation.

Outer masking is related to the change in power of detecting glitches of type k due to the presence of glitches of type k' in other records. It can be motivated using the concept of *robustness* of Definition 3.2 as follows. Suppose that we use a method with power π_k of detecting glitches of type k in the absence of glitches of type k'.

Suppose further that when glitches of type k' are introduced, the power changes. Some glitches of type k that occur on the same record as glitches of type k' disappear due to inner masking. Others of type k disappear due to outer masking by glitches of type k'. From Definition 3.2 the changed power is:

$$\pi_{k,k'} = \left(\pi_k - \epsilon_{k,k'}\right) = \rho_{k,k'}\pi_k.$$

Therefore, the probability that a glitch of type k will be outer masked (i.e., not detected), using the fundamental relationship in Eq. 1, is given by:

$$\begin{aligned} \mathcal{M}(k/k')|_{\text{outer}} &= 1 - \pi_{k,k'} \\ &= 1 - \rho_{k,k'}\pi_k. \end{aligned} \tag{4}$$

4 The masking index

In this section, we construct an index to quantify the masking effect of glitches in data.

Definition 4.1 The masking index of glitch type k with respect to glitch type k' is defined as the probability that the presence of a glitch of type k is masked by the presence of glitches of type k':

$$\mathcal{M}_{k/k'} = P\left(\mathcal{G}'_{ijk} = 0 | \mathcal{G}_{ijk} = 1 \wedge \mathcal{G}_{i'j'k'} \neq 0\right).$$

We can formulate the masking index in an alternate way to aid computation. A glitch is detected (not masked) if it is not inner masked and not outer masked, i.e.,

$$1 - \mathcal{M}_{k/k'} = A \times B,$$

where from Eqs. 2 and 3,

$$A = 1 - \mathcal{M}(k/k')|_{\text{inner}}$$

and

$$B = 1 - \mathcal{M}(k/k')|_{\text{outer}}$$

and therefore the masking index is

$$\mathcal{M}_{k/k'} = 1 - A \times B.$$

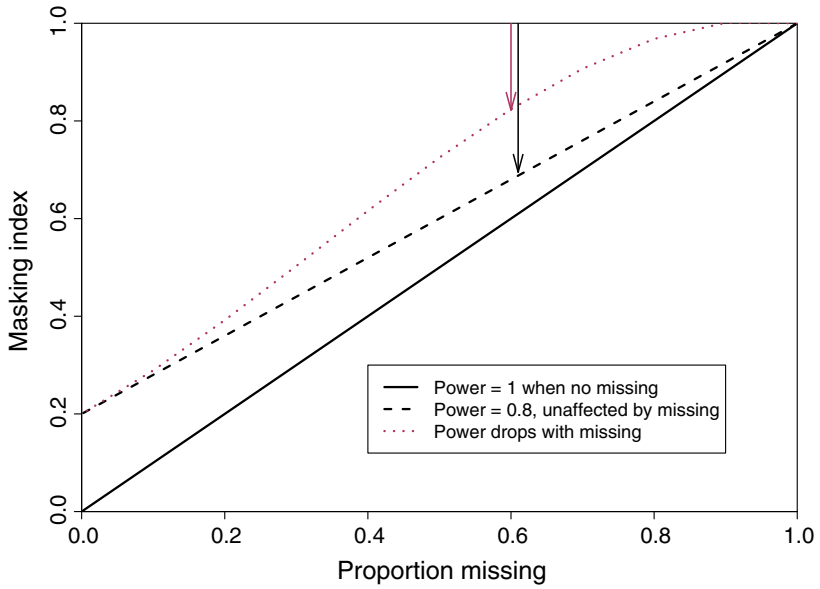


Fig. 2 Masking index: the X-axis depicts the proportion of missing values. The Y-axis represents the masking index or the probability of non-detection of outliers due to the presence of missing values. Each *plot line* corresponds to the masking index of a given method. The *solid line* represents a detection method with power 1, and the *dashed line* represents a detection method with power less than 1 but robustness 1 (not outer masked by missing). The difference between the *dotted* and *dashed lines* represents the additional effect due to outer masking

The above discussion can be represented by means of Fig. 2. For the purpose of illustration alone, assume that if k and k' are present in the same cell, k can never be detected. The X-axis depicts the proportion of glitch type k', in this case, the proportion of missing values. The Y-axis represents the masking index or the probability of non-detection of glitch of type k, (e.g., outliers), due to the presence of glitches of type k' (e.g., missing values). Each plotted line corresponds to the masking index of a given method.

In this figure, the solid line represents a detection method of glitch type k with power 1. Without glitches of k', the method detects glitches of type k with certainty. Here, the method's inability to detect glitches of type k is determined solely by the likelihood that they are replaced by a glitch of type k'. In other words, the masking index of this method for glitches of type k is the same as the proportion of glitches of type k'.

The dashed line represents the case where the detection method has power less than 1 (we used 0.8 in the figure), but whose performance is unaffected by the presence of type k' glitches, i.e., its robustness is $\rho_{k,k'} = 1$. This method detects glitches of type k with probability 0.8, and the loss in power is the same as the probability of replacement with a glitch of type k'.

The dotted line represents a method whose performance in detecting type k glitches is affected (impaired) by the presence of type k' glitches in other locations, with $\rho_{k,k'} < 1$. Therefore, according to Eq. 1, if we drop a perpendicular line from the top of the plot in Fig. 2 to any of the curves (as depicted by arrows in the figure), the length of that line would represent the power of the method. The difference between the dotted and dashed lines shows the size of the effect of type k' glitches on the performance of the detection method.

Note that the non-robust method loses more power as more glitches get masked compared with the other two robust methods. We study specific cases of masking in the next section where we instantiate A and B from the previous definition.

4.1 Specific cases of masking

We now discuss the conceptual formulations of the terms A and B for different instances of masking. The formulation depends on the nature of interaction between the two types of glitches considered. The conceptual formulations are important for computing the masking index in real-world datasets where the ground truth is not known.

When the ground truth is known, the quantities A (inner masking) and B (outer masking) can be estimated empirically. Glitches of type k are flagged, separately in the absence and presence of glitches of type k' and then compared with the ground truth. By controlling the proportion of glitches of type k and k', we can understand the behavior of the masking index $\mathcal{M}_{k/k'}$.

We defer a detailed discussion of the experiments and data simulation to Sects. 5 and 6.

4.2 Masking of numeric outliers by missing values

Let the outlier detection method O have power π_O and robustness $\rho_{O/M}$ to missing values. Suppose that the proportion p_M of missing values is scattered randomly throughout the dataset, so that the probability that an outlier and missing value occur together leading to masking of the outlier is given by p_M. Then, the masking index is given by

$$\mathcal{M}_{O/M} = 1 - (1 - p_M)\pi_O \rho_{O/M}. \tag{5}$$

Here, $A = 1 - p_M$ represents the probability that an outlier is not inner masked by a missing value. This probability is the same as that of any specific cell x_{ij} in the dataset being missing. The term $B = \pi_O \rho_{O/M}$ represents the altered power of the outlier detection method in the presence of missing values.

4.3 Masking of numeric outliers by duplicates

An outlier can be masked by duplicates if it is duplicated so many times that the outlier value becomes part of the normal portion of the data distribution as determined by the detection method.

Suppose records are randomly duplicated (either exactly or with slight errors so they become approximate duplicates). Suppose further that there is a process that generates these duplicate records, such that there is a probability p_d that the record is duplicated d times ($d = 1, 2, \dots$). Thus, $\sum_d p_d = 1$.

Clearly, for a moderately sized dataset, if an outlying value is duplicated two times, say, it is unlikely to affect the distribution of values much. Conceptually, there is some threshold K such that if the outlying value is duplicated at least K times, these K values overwhelm the rest of the distribution. The probability of this NOT happening (and hence not inner masked) is $\sum_{d<K} p_d$. Even if $d < K$, the power of the detection method may still be affected, with the new power given by $\pi_O \rho_{O/D}$, where $\rho_{O/D}$ is the robustness. This is related to outer masking. This suggests that a masking index for the masking of outliers by duplicates is of the form

$$\mathcal{M}_{O/D} = 1 - \left(\sum_{d<K} p_d \right) \pi_O \rho_{O/D}. \tag{6}$$

The difficulty in applying this expression, however, is that K and the p_d's are unknown and not easily inferred.

4.4 Masking of duplicates by missing values

Suppose that a record $x_{i,.}$ is duplicated d times, with duplicate records denoted by $x^1_{i,.}, x^2_{i,.}, \ldots x^d_{i,.}$. (Therefore, there are a total of $d+1$ duplicate records). The record $x_{i,.}$ is not identified as a duplicate if $x_{i,.}$ contains missing values, or if each of $x^1_{i,.}, x^2_{i,.}, \ldots x^d_{i,.}$ contains missing values. Hence, the probability that x_i is not inner masked is $(1 - p_M)(1 - p^d_M)$. With the power and robustness given by π_{Dd} and $\rho_{Dd/M}$, respectively (the subscript Dd represents duplication with d additional records), the masking index conditioned on a record with d duplicates is

$$\mathcal{M}_{Dd/M} = 1 - (1 - p_M)(1 - p^d_M)\pi_{Dd}\rho_{Dd/M}. \tag{7}$$

Note that both π_{Dd} and $\rho_{Dd/M}$ may depend on the actual number of duplicate records d, since duplicate detection might be easier with more duplicate records, for example.

The masking index for any duplicate record (duplicated any number of times) is then

$$\mathcal{M}_{D/M} = 1 - \sum_d p_d(1 - p_M)(1 - p^d_M)\pi_{Dd}\rho_{Dd/M} \tag{8}$$

where the second term is a weighted sum with weights given by p_d, the probability that a record is duplicated d times and p^d_M, the probability that d duplicate records contain missing values.

4.5 Masking of duplicates by numeric outliers

Similarly, outliers can mask duplicates in the same way as missing values, making the records different enough to be undetected as a duplicate. The masking index is given by

$$\mathcal{M}_{D/O} = 1 - \sum_d p_d(1 - p_O)(1 - p^d_O)\pi_{Dd}\rho_{Dd/O} \tag{9}$$

where p_O is the probability that any value x_{ij} is an outlier, and π_{Dd} is the power of the duplicate detection method for detecting d duplicates and $\rho_{Dd/O}$ the robustness to outliers.

In the rest of this paper, we study the masking index empirically, first using simulations with synthetically generated data (Sect. 5) and second with real-world datasets (Sect. 6). With synthetically generated datasets, the occurrence and amount of glitches can be controlled. Knowing the ground truth allows us to unambiguously quantify the amount of masking as well as the different contributions of inner and outer masking under various scenarios. With the real-world data sets, we quantify the degree of masking that exists by estimating the masking index in the absence of the ground truth. Since the masking index is computed with respect to a specific method, we also illustrate how we might use the masking index to choose a detection method that is least affected by masking from a set of candidate detection methods.

5 Masking index through simulations

The masking index is defined with respect to a pair of glitches of types k and k', as seen from Definitions 3.3 and 3.5. We ran experiments to study four cases: (1) masking of outliers (type k) by missing values (type k'), (2) masking of outliers by duplicates, (3) masking of duplicates by missing values, and (4) masking of duplicates by outliers. Simulations provide a controlled

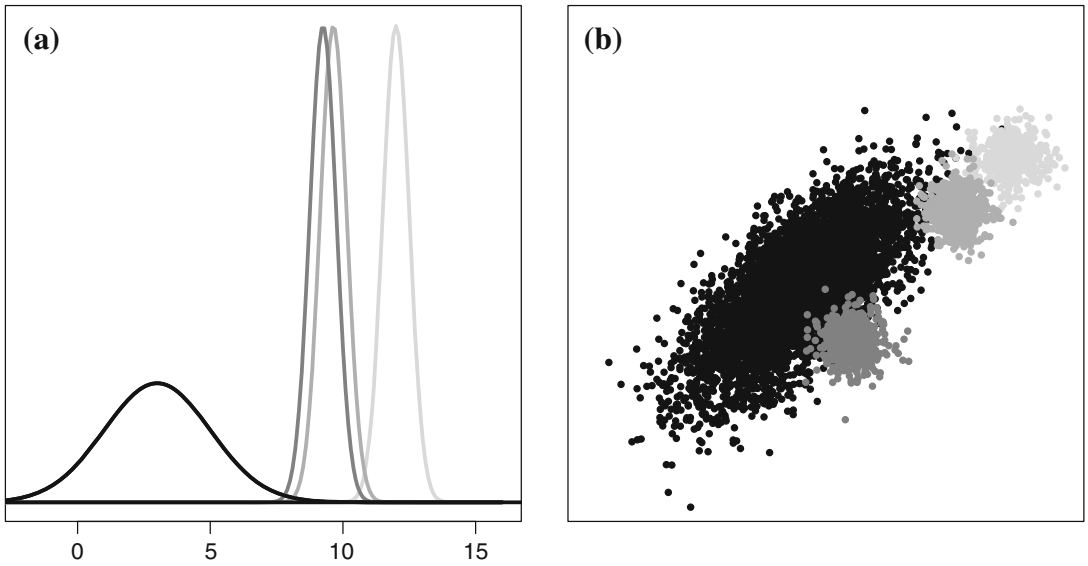


Fig. 3 Baseline distributions (*black*) and corrupting outlying distributions: **a** univariate $N(3, 4)$ baseline, with outlier distributions of means of 12 (*light gray*), 9.64 (*medium gray*) and 9.26 (*dark gray*), and variance 0.25; **b** bivariate normal (baseline in *black*) with corrupting distributions of independent bivariate normals with variance 0.1 and means (4, 2.5), (3, 1.5), and $(1, -1)$ in *light*, *medium*, and *dark gray*, respectively

environment to study the canonical behavior of the masking index. All the experiments were conducted using the R statistical package for data generation, outlier and duplicate detection, and for computing the masking index.

We started by creating a *baseline data set* of 5,000 records, each with four variables—a random string, a univariate normal variable and the two components of a bivariate normal variable. The univariate normal was $N(3, 4)$. The bivariate normal was specified by:

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.7\\0.7 & 1\end{pmatrix}\right).$$

5.1 Masking of outliers by missing values

To study the masking of outliers by missing values, we generated *corrupt data sets* by injecting outliers and missing values in varying proportions as illustrated in Fig. 3. Figure 3 shows plots of the univariate population and outlier densities, and sample realizations from the bivariate distributions we used. The baseline distribution is shown in black, with the corrupting distributions shown in dark, medium, and light gray.

Generating outliers The outlier values for the univariate normal variable in the baseline dataset were generated by drawing from a different univariate distribution, and similarly, from a separate bivariate normal distribution for corrupting the bivariate normal variable.

- For the univariate normal, the outlier distributions used were normal with variance 0.25, and means of (light gray), 9.64 (medium gray), and 9.26 (dark gray). The outlier distributions are chosen to be increasingly difficult to detect, resulting in a decrease in the power of the detection methods.
- For the bivariate normal variables, the outlier distributions were independent bivariate normals with variance 0.1 and means (4, 2.5) (light gray), (3, 1.5) (medium gray), and $(1, -1)$ (dark gray). These outliers distributions are detected with decreasing power

by any specific detection method as they become increasingly similar to the baseline distribution.
– We injected 5 and 10 % of outlier values, from each outlier distribution in turn.

Generating missing values To create missing values, we removed data at random, with the proportion of missing values ranging from 0.1 to 0.8. The original baseline dataset corresponds to missing value proportion = 0, and at the other hypothetical extreme, all the data are missing when missing value proportion = 1.

Outlier detection For outlier detection, we used the following methods defined in [2]:

– Four well-known univariate methods: (1) z-score with $p = .01$ and (2) z-score with $p = .05$, (3) inner ($Q1 \pm 1.5 * IQR$) and outer fences ($Q3 \pm 1.5 * IQR$) denoted by I/O fences method, and (4) 3σ on the continuous variables;
– Two multivariate methods based on: (1) Mahalanobis distance (see [12] for details) and (2) Jackknife distance (see [8] for details).

Computing the masking index $\mathcal{M}_{O/M}$ We computed the masking index very simply: By counting how many outliers were detected before and after the injection of missing values. Note that while it is obvious that some outliers are explicitly knocked off by the missing values (inner masking), other values that were outliers in the baseline data were not flagged as outliers in the corrupted dataset even though there was no missing value present at the record containing the former outlier. This is an example of outer masking.

Figures 4, 5, and 6 show the masking index for outlier detection in the presence of missing values, $\mathcal{M}_{O/M}$ for the cases of univariate and multivariate outlier detection methods. The X-axis denotes proportion of missing values. The Y-axis is the masking index of Eq. 5, estimated empirically from the simulated datasets where the ground truth is known. Figures 4 and 5 correspond to the univariate outlier detection methods using the inner and outer fences, z-score with $p = .01$, z-score with $p = .05$, and 3 Sigma methods; Fig. 6 corresponds to bivariate outlier detection with the Mahalanobis method. Since Jackknife bivariate outlier detection method behaves very similarly to Mahalanobis method in both cases of outlier injection, we did not report the figures. For each line of the figures, in the panels on the left side—panels (a), (c), and (e) in Fig. 4a in Figs. 5 and 6—the baseline dataset was contaminated with 5 % data from an outlying distribution shown in Fig. 3a. In the panels on the right side of each line—panels (b), (d), and (f) in Fig. 4b in Fig. 5, 10 % of the data are outliers from the univariate outlier distribution. In panel (a) of Fig. 6, 5 % of the data are injected with values from the outlying bivariate distributions represented in Fig. 3b and in panel (b), 10 % of the data are outliers.

The linear trend in the masking index is due to inner masking (dashed line). Each solid curve shows the masking index corresponding to a different outlying distribution as previously mentioned with variance 0.25, and means of 12 (light gray line), 9.64 (medium gray line), and 9.26 (dark gray line). According to the fundamental relationship between the masking index and power defined in Eq. 1, the power can be read from the top of the plot down to the curves. Although the same detection method is used within each panel in Fig. 4, the power is different due to the different alternatives (different means of the outlying distributions). As expected, the masking effect on outlier detection increases as the proportion of missing values increases. This is a consequence of inner masking. The amount of inner masking is almost the same for all the outlying distributions and in each panel; we show it for one distribution (medium gray line with mean 9.64) using a dashed line.

The result confirms our intuition that inner masking is linear in the proportion of missing values. This is because missing values *dominate* outliers. Note also the difference in the

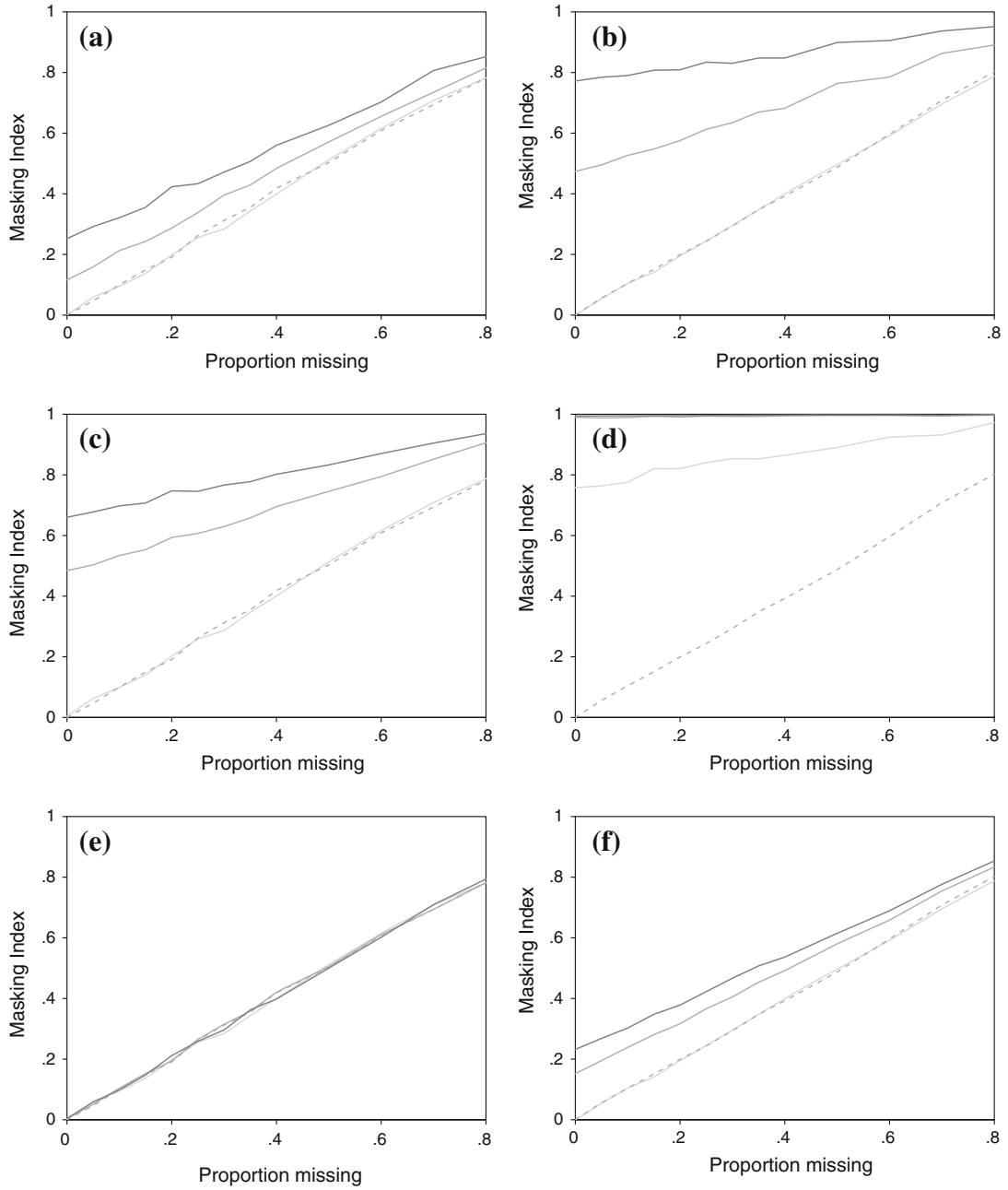


Fig. 4 Masking index for outlier detection in the presence of missing values $\mathcal{M}_{O/M}$—Case of Univariate Outlier Detection Methods. The X-axis denotes the proportion of missing values. The Y-axis is the masking index of Eq. 5, estimated empirically from the simulated datasets where the ground truth is known. The linear trend in the masking index is due to inner masking (*dashed line*). The *color* (*dark, medium, and light gray*) correspond to the outlier distributions used for corrupting the original dataset. **a** I/O fences 5 % univ., **b** I/O fences 10 % univ., **c** Z-score.01 5 % univ., **d** Z-score.01 10 % univ., **e** Z-score.05 5 % univ., **f** Z-score.05 10 % univ.

masking index (and hence the power) between panels (a)–(c)–(e) of Fig. 4 and panels (b)–(d)–(f) of Fig. 4 caused by the different amount of outlying values (from 5 to 10 %). The roughly straight lines in Fig. 4 suggest that the methods are not unduly affected by the amount of missing values. Similar behavior is shown for 3 sigma method in Fig. 5 and for Mahalanobis multivariate method in Fig. 6. As the amount of inner masking increases with

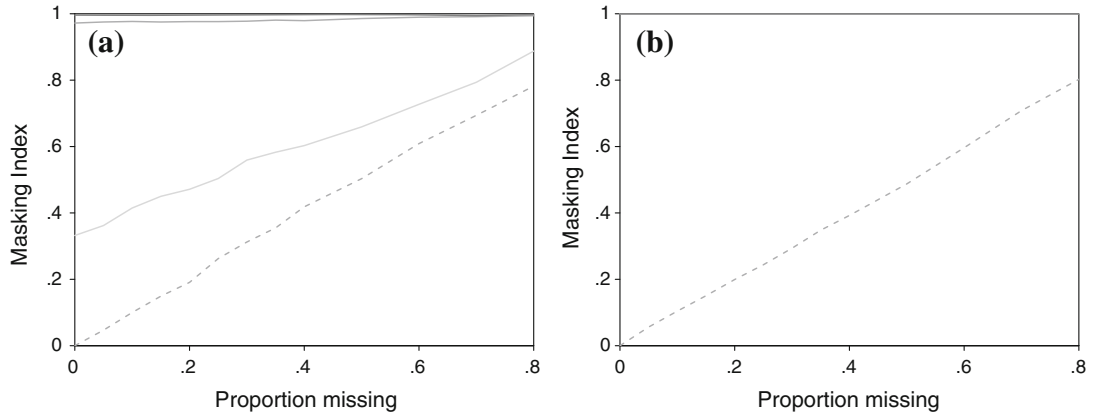


Fig. 5 Masking index for outlier detection in the presence of missing values $\mathcal{M}_{O/M}$—Case of Univariate Outlier Detection Methods (Continued). **a** 3 Sigma 5 % univ., **b** 3 Sigma 10 % univ.

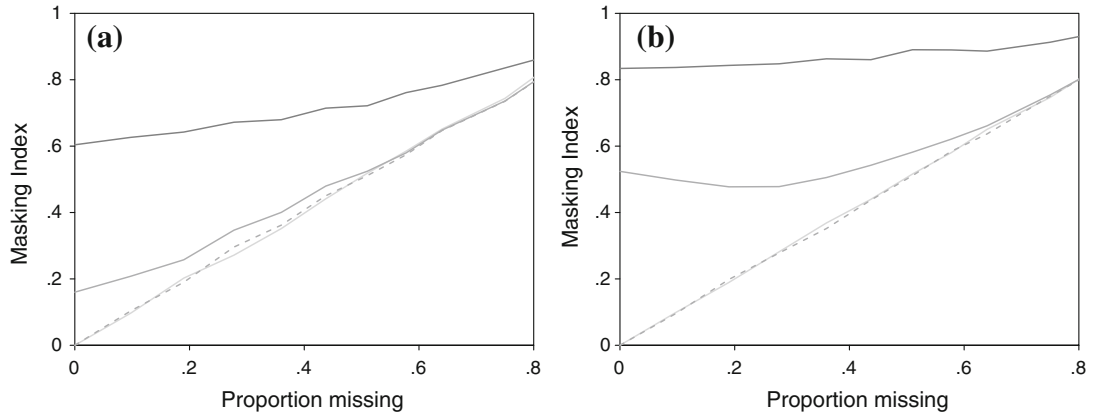


Fig. 6 Masking index for outlier detection in the presence of missing values $\mathcal{M}_{O/M}$—Case of Multivariate Outlier Detection Method. The X-axis denotes the proportion of missing values. The Y-axis is the masking index of Eq. 5, estimated empirically from the simulated datasets where the ground truth is known. The linear trend in the masking index is due to inner masking (*dashed line*). The color (*dark, medium, and light gray*) correspond to the outlier distributions used for corrupting the original data set. **a** Mahalanobis 5 % biv., **b** Mahalanobis 10 % biv.

the proportion of missing values (the dashed line), the amount of non-detection due to a lack of power (the gap between the solid curve and the dashed line) decreases.

5.2 Masking of outliers by duplicates

To study the masking index of outliers (type k) by duplicates (type k'), we used the same baseline dataset discussed in Sect. 5.1 and the datasets with 5 and 10 % injected outliers. We then introduced duplicates into the data.

Generating duplicates To inject duplicates,

- We first split the baseline datasets into non-outlying records and outlying records, creating D_N and D_O, respectively.
- We created three sets of building block duplicate records by drawing 100, 50, or 0 % from the outlying records D_O and the rest from the non-outlying records D_N.
- We replicated these building block duplicate sets 1, 2, and 5 times, yielding the final duplicate sets.
- The final duplicate sets were then added to the baseline data sets in turn.

– We also considered exact and approximate duplication. For approximate duplication there is an intermediate step applied to the duplicate sets before they are added to the baseline datasets: random values drawn from a uniform distribution with endpoints $\pm\epsilon$, $\pm 2\epsilon$, and $\pm 3\epsilon$ are added to the variable of interest (v). We chose ϵ to be $0.05\sigma_v$, where σ_v is the standard deviation of variable v.

We added exact and approximate duplicate records of size ranging from 250 to 2,500 to the original baseline datasets of size 5,000 while also varying the proportion of duplicates that themselves contain outliers.

Computing the masking index $\mathcal{M}_{O/D}$ In order to capture random variation, we replicated the process of each corruption and detection method 5 times, to average our results.

Figure 7 shows plots from the experiments. Panels (a), (b), and (c) correspond to duplicates generated only from non-outlying records (0 %), 50 % from outlying records and 100 % from outlying records. The shape of the dots corresponds to the respective outlying distributions used in Fig. 3. The solid lines and dots correspond to data with 5 % injected outliers, while the dashed lines connecting the circle symbols represent 10 % outliers. We report results only for the inner fences outlier detection method.

The X-axis shows the characteristics of the duplication used in the experiment. The left-most points of the curves show the masking index without any duplicates. This is followed by 3 groups of 4 points each, labeled "1X," "2X," and "5X," representing the cases where the records chosen for duplication are replicated 1, 2, or 5 times. Within each group of points, we have either exact duplicates, or duplicates shifted randomly by different amounts, labeled "0" (for exact), "ϵ," "2ϵ," and "3ϵ." Note that there is no strict order for the duplication characteristics. We join the dots so that the reader can more easily make out the differences between the dots.

We find that for outlying distributions that are very different from the population (blue), duplication of the non-outliers has minimal effect on outlier detection (panels (a) and (b) for 0 and 50 % non-outliers duplication). However, with 100 % of duplicates generated from outliers (panel (c), the masking index becomes high when the number of duplicates is large. With the other outlying distributions with means closer to that of the population, masking becomes very high even with moderate amounts of duplication. There is little difference between exact and approximate duplication, although this might be due to our use of a small value for ϵ.

5.3 Masking of duplicates by missing values

For this set of experiments, we used the same method of introducing missing values and duplicates as described in the previous sections. We used the following duplicate detection methods.

Cosine similarity The first method is a cosine similarity distance where the distance between the numeric vectors of any two records is the dot product of these two vectors. The two records are identified as being duplicates if this distance is smaller than a pre-specified threshold. We used thresholds corresponding to 1°, 5°, 10°, and 30°. Note that prior to applying the detection method, the individual variables are normalized to have zero mean and one standard deviation so that the different variables become comparable. Otherwise one variable might dominate the others in the distance computation.

Delta quantile similarity We propose a second duplicate detection method based on quantile similarity. For two records to be duplicates, all the numeric attributes have to be close together. Therefore, all the component-wise differences have to be small. We sorted each normalized attribute, computed the successive differences and ranked the differences into

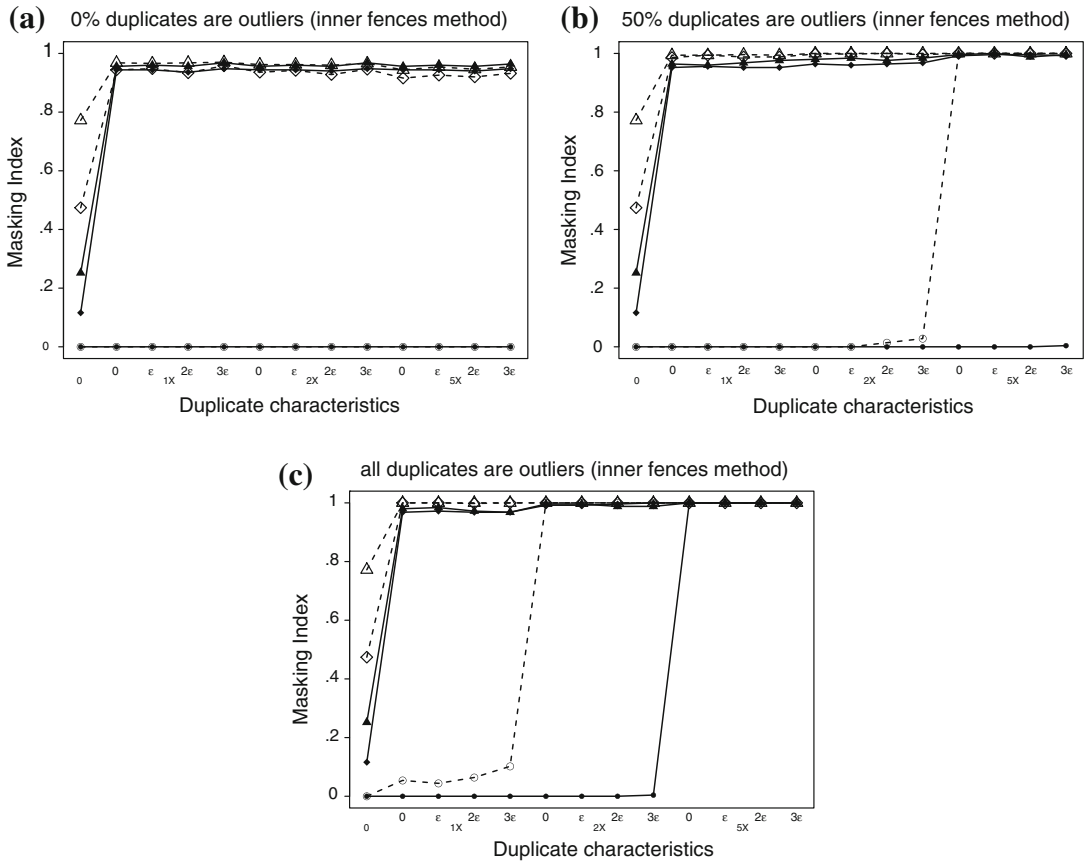


Fig. 7 Masking index for outlier detection in the presence of duplicates $\mathcal{M}_{O/D}$. The three panels correspond to where the duplicates come from, 0, 50, and 100 % from the outliers, respectively. The amount of outliers, 5 or 10 %, are indicated by *solid lines/dots* and *dashed lines/clear dots*, respectively. *Dots/circles* correspond to the *light gray* outlier distribution, *triangles* to the *dark gray* distribution and *diamonds* to the *medium gray* distributions shown in Fig. 3a. Within each panel, the leftmost point shows the masking index when there are no duplicates. The next set of 4 points are 1× replication (one exact, and 3 approximate) followed by 2× and 5× replications. **a** 0 % duplicates are outliers (I/O fences method), **b** 50 % duplicates are outliers (I/O fences method), **c** all duplicates are outliers (I/O fences method)

percentiles. We called this the *Delta Rank*. We then mapped each original attribute to its delta rank. For two records to be duplicates, it is necessary for the delta ranks to be small, that is the components should have the smallest possible deltas. To determine duplicates, we used the condition that the delta ranks of *all* the attributes have to be smaller than a given threshold. Varying the thresholds enables us to identify approximate duplicates.

Figure 8 shows the results of the simulation study that examines the masking effect of missing values on the detection of duplicates using the cosine and delta quantile similarity methods. The two plots in each row correspond to high and low thresholds used in the detection method. We also made plots with medium levels of detection thresholds (5° and 10°), but they are similar to the low threshold plot (Fig. 8b) and thus are not shown here. Within each plot, the solid and dashed lines reflect the masking index for detecting exact and approximate duplicates. For clarity, we show only the 3ϵ set of approximate duplicates—within each plot, the curves corresponding to 1ϵ and 2ϵ approximate duplicates lie between the curves for the exact and 3ϵ ones.

We find that detection of approximate duplicates is more difficult than detection of exact duplicates—the masking index is consistently higher throughout the range of proportions of

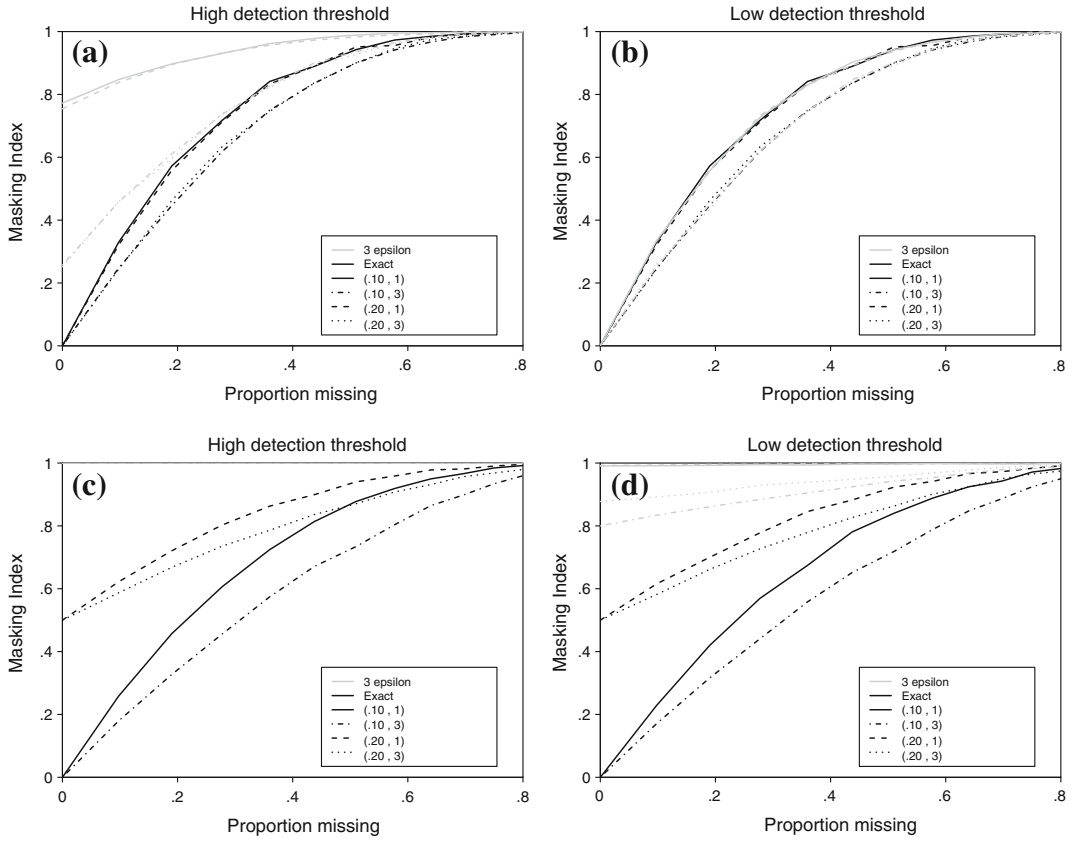


Fig. 8 Masking index for duplicate detection with missing values $\mathcal{M}_{D/M}$. The two panels correspond to different duplicate detection thresholds. The *solid* and *dashed lines* in each panel show the masking index for the cases of exact and approximate duplicates, respectively. The *different colors* represent different parameters used in the duplicate generation, namely the proportion of duplicate records added to the original dataset size and the number of replications from 1 to 3 times. **a** High threshold in cosine similarity of 1°, **b** low threshold in cosine similarity of 30°, **c** high threshold in delta quantile similarity, **d** low threshold in delta quantile similarity

missing values. This is expected since there is more uncertainty with approximate duplicates. The proportion of data that is (both exactly or approximately) duplicated did not seem to affect the masking index: The lines that have, respectively, 10 and 20 % of duplicated data have very similar curves. However, the number of replications (1–3) affects the masking index. Specifically, the masking index is lower when there is more replication [the (.10, 3) and (.20, 3) lines are lower than the (.10, 1) and (.20, 1) lines]. This is because when a duplicated record is replicated more times, there is a lower probability that all the duplicate records get masked by missing values.

Note that the masking index curves are not linear. Recall that the curves are expected to be linear if the masking effect is mostly due to inner masking. With outer masking, there is additional masking due to detection of the masking glitch in *other* records. In the current experimental scenario, the presence of missing values in other records can affect duplicate detection of any given record. Hence, there is scope for outer masking, and this is reflected in the plots in Fig. 8.

Finally, note that the two methods have very different characteristics. The delta quantile similarity method is sensitive to duplication, but more robust to missing values, whereas the cosine similarity method is not that sensitive to extent of duplication but more sensitive to missing values as demonstrated by the steepness of the curves.

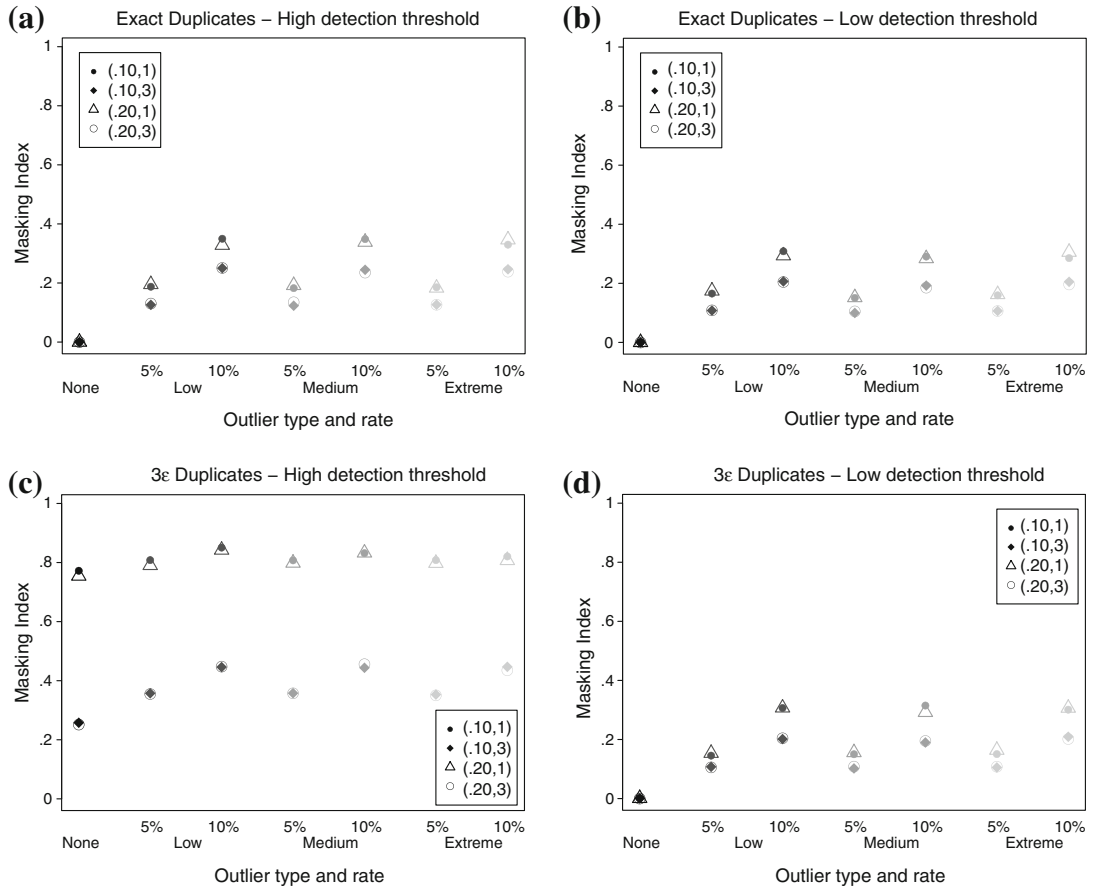


Fig. 9 Masking index for duplicate detection by the cosine similarity method in the presence of outliers $\mathcal{M}_{D/O}$. The *different symbols* correspond to different parameter settings for the duplicates (namely the proportion of duplicate records added to the original, 10 or 20 %, and the number of replications of 1 or 3 times), while the different *shades of gray* represent the different outlier distributions shown in Fig. 3. The outliers occur at two rates, 5 or 10 %. **a** High threshold in cosine similarity of 1°, **b** low threshold in cosine similarity of 30°, **c** high threshold in cosine similarity of 1°, **d** low threshold in cosine similarity of 30°

5.4 Masking of duplicates by outliers

In this set of experiments, we explore the masking of duplicate records by outliers. We introduce duplicate records and outlier values in the same way as described in Sects. 5.1 and 5.2, and perform duplicate detection using the cosine similarity and delta quantile similarity methods.

Figures 9 and 10 show the masking index obtained for the two duplicate detection methods. We find that, in general, the presence of outliers increases the masking index slightly, with a larger increase when there are more outliers present. The masking index appears to vary negligibly with the three outlier distributions we considered.

For both exact duplicate detection methods in Figs. 9a, b and 10a, b, the choice of threshold does not affect the masking index. With 3ϵ-approximate duplicates, however, using a higher detection threshold raises the masking index (and vice versa).

Moreover, for exact duplicate detection, cosine similarity method is less affected by the 20 % injected outliers compared with delta quantile similarity method where the masking index is approximately 4 times higher when 20 % outliers are injected.

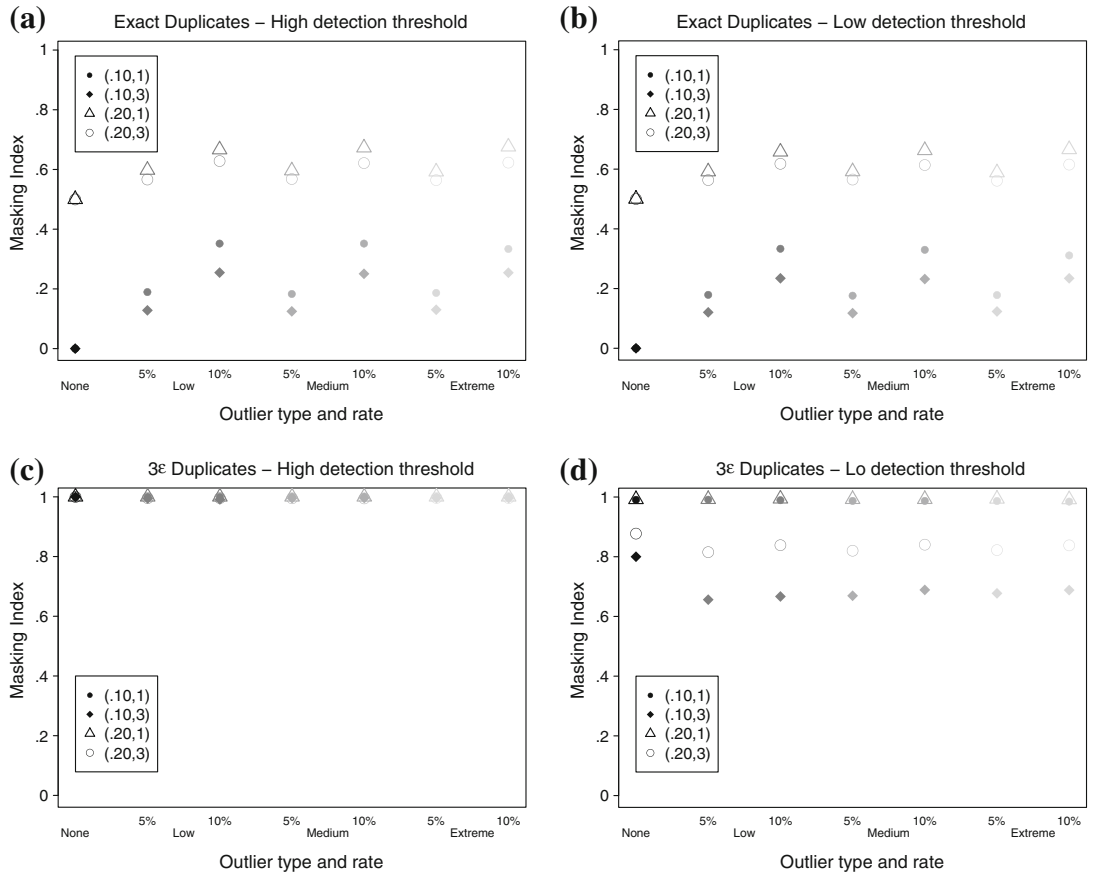


Fig. 10 Masking index for duplicate detection by the delta quantile similarity method in the presence of outliers $\mathcal{M}_{D/O}$. The *different symbols* correspond to different parameter settings for the duplicates (namely the proportion of duplicate records added to the original, 10 or 20 %, and the number of replications of 1 or 3 times), while the different *shades of gray* represent the different outlier distributions shown in Fig. 3. The outliers occur at two rates, 5 or 10 %. **a** High threshold in delta quantile similarity, **b** low threshold in delta quantile similarity, **c** high threshold in delta quantile similarity, **d** low threshold in delta quantile similarity

For approximate duplicate detection with cosine similarity method in Fig. 9c, the number of replications from 1 to 3 has significant impact in reducing the masking index for high threshold, whereas for the delta quantile similarity method in Fig. 10d, this reduction of the masking index is significant for low threshold.

Cosine similarity method has a much lower masking index for approximate duplicate detection than delta quantile similarity method for both thresholds in Figs. 9c, d and 10c, d. Relying on only one replication always degrades the approximate duplicate detection and more significantly for cosine similarity method as shown in Fig. 9c.

Finally, we can notice that the masking index drops in Fig. 10d when there are outliers versus no outlier. One explanation is that the outliers increased the bin size of the delta quantile similarity method, which effectively "lowers" the threshold for detection even more.

6 Masking index for real-world data

We use two publicly available real-world datasets, one on Internet advertisements and another much larger dataset on mean sea level differences. Datasets are described in the next subsections. Based on these, we:

– Demonstrate the estimation of the masking index when the ground truth is not known, and
– Use the masking index to choose a detection method that is least affected by masking.

In these two datasets, there are missing values and no duplicates, and we consider the masking effect of missing values (type k') on the detection of outliers (type k). The results are summarized in Fig. 11.

6.1 Estimating the masking index

Since the data already have missing values, outliers are already masked, i.e., outliers cannot be detected at the missing values (inner masking) and the power of the detection methods are already affected by the missing values (outer masking). Thus, to estimate the masking index, we use the following approach:

Step (1) We use the theoretical formulation given in Sect. 4.1 for the masking index, which requires estimation of the power π_O and the robustness $\rho_{O/M}$. The proportion of missing values p_M is also required, but this is, of course, easily obtained.

Step (2) Given multiple methods of detection, we select one method to detect outliers and treat it as the ground truth. With this specification of the ground truth, we estimate the power and robustness of the other methods, using steps (3) and (4) below. We rotate the role of the methods, so that each method gets to be treated as the ground truth.

Step (3) In order to estimate the power, we divide the data (D) into clean dataset (D_{clean}) with no missing values, and dirty (D_{dirty}). Due to inner masking, there are no outliers detected in D_{dirty}. The power of each method is obtained by comparing the number of outliers it detects in D_{clean} with the number of outliers of the ground truth.

Step (4) The robustness of a detection method is found by treating the clean D_{clean} as the complete data and injecting a proportion p_M of missing values to it, creating D_M. The choice of value of p_M is the actual amount of missing values in the original data D. By injecting missing values to the clean D_{clean}, we are re-creating the scenario of having proportion p_M of missing values in the original dataset D. By comparing the detections in D_{clean} with those in D_M, the change in power and hence the robustness at the observed level of missing values can be estimated.

Step (5) The injection of missing values is repeated multiple times to reduce sampling error and obtain more stable estimates of the robustness. An estimate of the masking index is then obtained using Eq. 5. Finally, we take the mean of all the masking indices obtained from the rotation of detection methods used for the ground truth, to obtain the masking index estimate of each method.

Identifying the ground truth can also be done using a voting mechanism. To estimate the masking index of one method, we use the other methods to determine the ground truth—if more methods agree, there is greater confidence that a glitch is real. The procedure described above can be considered a special case where a single method does the voting.

6.2 Internet advertisements data

The Internet advertisements data are described in [11] and are available at the UCI.[2] It contains 3,279 instances representing a set of possible advertisements on Internet Web pages. The features encode the geometry of the advertisement image (if available), specifically the *height, width*, and *aspect ratio*. Here, we focus on just the aspect ratio.

[2] http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements.

Data Set	Description	Missing	3σ	I/O Fences	Z-score .05	Z-score.01
Internet Ads	3,279 records 3 continuous variables	23 (0.7%)	25 (0.76%)	33 (1%)	30 (0,91%)	50 (1.52%)
		Masking $M_{O/M}$	**0.20**	**0.17**	**0.21**	**0.22**
Mean Sea Level Data	826,000 1 continuous variable	132,160 (16%)	14,520 (1.76%)	54,834 (6.64%)	32,717 (3.96%)	21,776 (2.64%)
		Masking $M_{O/M}$	**0.62**	**0.69**	**0.53**	**0.51**

Fig. 11 Masking index estimates for the real-world datasets

As shown in Fig. 11, the proportion of missing is small, about 0.7 %. This dataset has no duplicate record. We used four outlier detection methods: 3σ, inner/outer fences, z-score with $p = .05$, and z-score with $p = .01$. For these methods, the average estimated statistical power is 0.84, 0.84, 0.88, and 0.88, respectively. The corresponding estimated mean robustness values are 0.96, 1, 0.9, and 0.89, based on 100 independent replications of data created from injecting missing values to the clean portion of the data. Using Eq. 5, the corresponding estimated masking indices are 0.2, 0.17, 0.21, and 0.22, suggesting that for this dataset, the inner/outer fences method is "best" in terms of having the least amount of outliers masked by the missing values.

6.3 Mean sea level data

The mean sea level dataset is extracted from the Permanent Service for Mean Sea Level[3] and contains about 826,000 tuples describing the average change in sea level from tide gauges and bottom pressure recorders all around the world. The data goes back to the 1800s. This dataset has more missing values than the Internet advertisements dataset, with about 16 % missing values. It too has no duplicates. However, probably due to the extremely large data size, we find that the robustness of all the four detection methods are high, at around 0.998 with little difference between the methods. The ranking of the masking indices was thus determined by the relative power of the methods. The estimated power was 0.59, 0.57, 0.67, and 0.68, for the 3σ, inner/outer fences, z-score 0.05 and z-score 0.01 methods, yielding masking indices 0.62, 0.69, 0.53, and 0.51, so that the z-score 0.01 method has the lowest masking index for this data set.

In this paper, we assume that the glitch matrix can be computed off-line. Many other anomaly detection methods can be used to characterize the *dirtiness* of the dataset (and consequently, augment the size of the glitch matrix and the overall computing time). In the experiments on the two real-world datasets, we have demonstrated that the masking index is an effective method for determining a glitch detection method that is least affected by masking. Along with other data quality metrics, the masking index plays a critical role in the selection of data cleaning strategies, particularly in the case of iterative cleaning where it can be used to determine a stopping criterion for iteration. Future work will be devoted to the study of the efficiency, complexity, and scalability of our approach while expanding the set of anomaly detection methods.

[3] Permanent Service for Mean Sea level—PSMSL: http://www.psmsl.org/.

7 Related work

It is common to use data for constructing models that represent real-world behavior in compact and aggregated forms (e.g., formulas, charts, statistical classification or regression models, etc.). These summaries allow the decision maker to understand and analyze certain phenomena and behaviors. Model complexity and reliability may be significantly affected by data cleaning and preparation processes and their resulting data quality. The link between data quality and decision model correctness, which has been explored in a variety of studies (e.g., [5]), is still very complex and difficult to assess.

Removing anomalies and noisy data are an important goal of data cleaning because noise and errors hinder most types of data analysis. Most existing data cleaning methods from database research, data mining and statistics literature focus on removing noise as the result of low-level data errors from an imperfect data collection process [13], but the masking effect of a conjunction of anomalies can significantly bias data preparation and hinder analysis.

To the best of our knowledge, there is little work focused on iterative cleaning to efficiently detect and remove masked glitches. Except [4], most of the techniques currently detect or treat each data anomaly in isolation, and they do not exploit patterns of glitches for data cleaning. The detection is also clearly independent and disconnected from the cleaning process. In addition [6], defined the notion of statistical distortion as an essential metric for measuring the effectiveness of data cleaning strategies since a cleaning method may introduce new errors. Our approach addresses the challenges not addressed by prior work.

8 Conclusions

In this paper, we introduced the concept of *masking index*, a statistically rigorous way to quantify the effect of the presence of one type of data glitch on the detection of other types of glitches. We defined two different types of masking, inner and outer masking, to separate the different effects that one type of glitch has on the detection of another type. Using the fundamental relationship between statistical power and masking, we presented theoretical formulations of the masking index for pairs of different glitch types. In particular, we discussed in detail the interactions between outliers, duplicates, and missing values.

We illustrated the estimation of the masking index using synthetic and real-world data. With simulations, we can control the occurrence of glitches and establish the ground truth and therefore easily estimate the masking index. This allows us to study the behavior of the masking index of a particular method with respect to various characteristics of glitches. With the real-world datasets, where the ground truth is not known and masking is already present, we proposed a method for estimating the robustness and masking indices of multiple detection methods. This allowed us to identify detection methods that are less affected by masking.

An application of this work is in the area of anomaly detection for extremely large datasets where we want to limit the number of detection methods applied to the data. By extracting a smaller subset of the data and performing a comprehensive masking analysis on this subset, we can identify a small number of detection methods that are less affected by masking to be applied to the large data set.

Future work consists of extending this work in a number of fronts. Firstly, we will consider masking indices to quantify masking caused by multiple glitch types. This may be done directly or by combining the masking indices of individual glitch types. As an example, using the same probability considerations in this work, a possible masking index for duplicates

masked by outliers and missing values may be written in the form

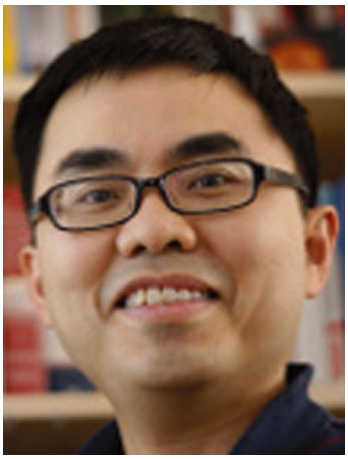
$$1 - \sum_{d} p_d (1 - p_{O \cup M})(1 - p_{O \cup M}^{d}) \pi_{Dd} \rho_{Dd/(M,O)},$$

where $p_{O \cup M}$ is the probability of being masked by an outlier, missing value or both, and $\rho_{Dd/(M,O)}$ is the new power of duplicate detection under the presence of outliers and missing values. We will examine this and other indices in more detail using simulation studies and applications to data. Secondly, we will consider using a generalized linear model based on multiple detection methods to estimate the ground truth when dealing with real data in applications. Lastly, we will address the important topic of iterative cleaning, and its effect on the masking index. This will be in the context of statistical distortion introduced by [6]. Essentially, the process of iterative cleaning has to trade-off the reduction in the masking index, along with other data quality criteria, and the statistical distortion caused by the cleaning.

References

1. Acuna E, Rodriguez CA (2004) Meta analysis study of outlier detection methods in classification, IPSI
2. Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, New York
3. Ben-Gal I (2005) Outlier detection. In: Maimon O, Rockach L (eds) Data mining and knowledge discovery handbook: a complete guide for practitioners and researchers. Kluwer, Dordrecht
4. Berti-Equille L, Dasu T, Srivastava D (2011) Discovery of complex glitch patterns: a novel approach to quantitative data cleaning, ICDE, pp 733–744
5. Blake R, Mangiameli P (2011) The effects and interactions of data quality and problem complexity on classification. J Data Inf Qual 2(2):8:1–8:28
6. Dasu T, Loh JM (2012) Statistical distortion: consequences of data cleaning. PVLDB 5(11):1674–1683
7. Davies L, Gather U (1993) The identification of multiple outliers. J Am Stat Assoc 88(423):782–792
8. Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
9. Hawkins D (1980) Identification of outliers. Chapman and Hall, London
10. Iglewics B, Martinez J (1982) Outlier detection using robust measures of scale. J Stat Comput Simul 15:285–293
11. Kushmerick N (1999) Learning to remove internet advertisements. In: Proceedings of the third annual conference on autonomous agents, AGENTS '99, pp 175–181
12. Rao CR (1973) Linear statistical inference and its applications. Wiley, New York
13. Xiong H, Pandey G, Steinbach M, Kumar V (2006) Enhancing data analysis with noise removal. IEEE Trans Knowl Data Eng 18(2):304–319

Laure Berti-Équille is a Senior Scientist at Qatar Computing Research Institute in the Data Analytics Group. Her research interests focus on developing novel data management and mining techniques for data integration, truth discovery, and anomaly detection. Prior joining QCRI, Laure Berti-Équille was a "Directeur de Recherche" at IRD, the French Institute of Research for Development (2011–2013), a visiting researcher at AT&T Labs-Research (NJ, USA) (2007–2009) and a tenured Associate Professor at University of Rennes 1 (France) (2000–2010). She is an associate editor of the ACM Journal of Data and Information Quality (JDIQ). She was the program co-chair of ICIQ 2012 and received a Marie Curie fellowship of the European Commission (Grant FP6-MOIF-CT-2006-041000).



Ji Meng Loh is Associate Professor in the Department of Mathematical Sciences at New Jersey Institute of Technology. His primary research interests are in the area of spatial statistics, especially the development of statistical methodology for the analysis of spatial point patterns. Ji Meng has made contributions to bootstrap methods for spatial data, anomaly detection of spatial point patterns, and understanding the effects of data quality issues on statistical inference. Prior to joining NJIT, Ji Meng served on the faculty at Columbia University and also as Principal Member of Technical Staff at AT&T Labs-Research.

Tamraparni Dasu is a Lead Member of Technical Staff in the Department of Statistics at AT&T Labs-Research. She joined AT&T Bell Laboratories immediately after receiving her Ph.D in Statistics from the University of Rochester in 1991. She has wide experience in mining massive telecommunication data, data streams, and network data. She is also an expert on data quality and has published extensively on this topic, including the book "Exploratory Data Mining and Data Cleaning," T. Dasu & T. Johnson, John Wiley, 2003. Her personal interests include literary translation and writing fiction.