Random Sampling-Arithmetic Mean: A Simple Method of Meteorological Data Quality Control Based on Random Observation Thought

SEN TIAN$^{1}$, JIN ZHANG$^{2}$, LINGYU CHEN$^{2}$, HONG LIU$^{2}$, AND YING WANG$^{3}$

$^{1}$School of Mathematics and Statistics, Hunan Normal University, Changsha 410081, China
$^{2}$College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
$^{3}$School of Humanities and Management, Hunan University of Chinese Medicine, Changsha 410208, China

Corresponding author: Jin Zhang (mail_zhangjin@163.com)

This work was supported in part by the research projects: the Hunan Provincial Department of Education Innovation Platform Open Fund Project under Grant 15K082; in part by the Hunan Provincial Department of Transportation Scientific and Technological Progress and Innovation Program Project under Grant 201927; in part by the Central Military Commission Department of Equipment Development Pre-research Project under Grant 31511010105; in part by the National Defense Science and Technology Bureau of Central Military Commission Department of Equipment Development Basic Research The planned project under Grant WDZC20200500119; in part by the Hunan Provincial Department of Education Project under Grant JG2018A012; in part by the Xiangjiangtong under Grant [2019] No. 291, Grant [2019] No. 248, Grant [2019] No. 370, Grant [2020] No. 9, Grant [2020] No.90, and Grant [2020] No.172; and in part by the Ministry of Education's Industry-University Cooperation and Collaborative Education Project under Grant 201901051021.

I. INTRODUCTION

Meteorological data [1] are widely used in military and civil fields and are the basis for meteorological operations, scientific research, and services. The quality of data directly affects the development of the meteorological career, which is very important for decision-making service and forecast verification. Meteorological data is the basic data of weather forecast and climate forecast, and its quality directly affects the accuracy of the weather forecast and climate forecast. For example, more than 2,500 ground-based meteorological observation stations were installed nationwide in 2011. Hence rapid and efficient quality control methods for detecting and marking errors and suspicious data are necessary.

A critical problem in the quality control of meteorological data is the quality of the claims. For example, the ground observation station fails the sensor and produces conflicting claims. The typical problem is that the automatic
meteorological observation stations in adjacent areas generate conflicting data due to the failure of sensors among the quality problems of claims. The types of data observed by ground meteorological observation stations vary, and the data structures are also different. The claims of the same object by meteorological observation stations at different places may be different, and the claims of the same object by different observation stations in the same place may also be different [2]. At the same time, due to errors, missing data, typos, out-of-date data, etc., data collected about the same object conflicts with each other. For example, Google’s search results for the highest temperature in Beijing include “29°C” from the China Meteorological Administration, “28°C” from the Beijing Meteorological Administration, and “30°C” from the US National Weather Service. So, which of these noisy messages is more trustworthy, and which represents the truth? In this and many similar problems, noise information from the same set of objects or events gathered from different sources must be summarized to get the facts. Truth discovery has lately received great attention for its resolving data conflicts, ensuring or improving the quality of collected data, and providing users with the most accurate information and data.

Truth discovery plays an important role in the information age. On the one hand, people need accurate information more than ever before, but on the other hand, inconsistent information is inevitable. As a result, truth discovery can be applied to many fields: health care [3], social perception [4], [5], information extraction [6], [7], recommendation system [8], [9], quality of service (QoS) [10]–[12], and so on. For example, to solve the problem of missing QoS data, Xin Luo et al [10] introduced effective second-order solvers into the latent factor model to accurately predict the missing QoS data. And [11] generate highly accurate predictions for missing QoS data by aggregating non-negative potential factor models.

In the existing truth discovery methods, since the reliability of the data source and truth are unknown, some solve the truth according to the principle of the iterative algorithm [13]–[17], some obtain the truth through the optimization formula [18]–[21], and others obtain the truth according to the probability graph model [22]–[24]. However, when calculating truth, most methods first consider the reliability of the data source, because a reliable data source provides accurate data, which adds complexity to the calculation of truth. And when the number of data sources increases, some truth discovery models, such as the model based on the optimization method, is too complex to solve the quality problem of claims, which requires a large amount of calculation and reduces the accuracy. Therefore, it is worthwhile devoting much effort to construct a model to reduce the complexity of the operation without reducing the accuracy of truth discovery.

It is a well-known fact that the mean method has been used as an estimate of truth in precision instruments for many years and verified theoretically and practically. And, we first considered using the mean method to resolve data conflicts to infer truth. However, the mean method did not achieve ideal results in the experiment and failed to reach the expected target. Meanwhile, few studies paid attention to the reasons why the results obtained by the mean method were always unsatisfactory in the research of truth discovery. Then, our goal is to extrapolate the truth, so the first step is to find a set of reliable observations. Therefore, it is desirable to exploit the reasons for the unsatisfactory mean method and to find the truth from conflicting meteorological data. In this paper, the idea of the random observation method is combined with the Arithmetic Mean to apply in the problem of meteorological data quality control, and a Random Sampling-Arithmetic Mean (RS-AM) method is proposed to solve the data conflict and deduce the truth.

Compared with the existing methods, the RS-AM method has the greatest advantage of simple calculation. This is because the RS-AM method does not need to estimate the reliability of data sources, but only needs to keep random extraction until a set of data observation values that meet the requirements are selected, and then the truth is calculated. In this work, based on the random sampling model, the method evaluates the goodness of fit between the expected distribution and the sampling distribution by repeatedly extracting the random observation vectors, to find the random observation vectors closest to the expected distribution. Then, the distance between the median and arithmetic mean of each group of claims is calculated by the distance formula, and the group of claims with the smallest distance is selected. Finally, random extraction is continued on the selected set of claims until the stop condition is met. The convergence of this method is proved by theoretical derivation. And, this paper proves the reason why the arithmetic mean is not ideal for truth estimation.

In summary, this paper mainly has the following contributions:

1. We theoretically prove the reasons for the mean method as truth estimation and the reasons for the unsatisfactory effect of arithmetic mean as truth estimation, which lays a foundation for the proposed model of this paper.

2. Based on the above reasons, we combine the idea of random observation with arithmetic mean and put forward a random sampling-arithmetic mean (RS-AM) model. Under the premise of not reducing the truth discovery accuracy, the complexity of the operation is reduced and the truth is deduced. Compared with the traditional method, this method does not calculate the weight of the data source.

3. We propose an algorithm that continuously screens and calculates the claims that meet the convergence conditions, and calculates the truth, and prove the convergence of the algorithm.

4. The RS-AM model and algorithm are verified on the real meteorological datasets. The results show that the method is effective in the quality control of meteorological data.
In the following section, we first illustrate some background knowledge. Then, in Section 3, we formulate the problem and detail the proposed method. In Section 4, experiments are conducted on real-world datasets, and we validate the effectiveness and efficiency of the proposed method. Related work is discussed in Section 5, and finally, we conclude the paper in Section 6.

II. RELATED WORKS
In the process of analyzing big data and social media information, data conflicts are ubiquitous in knowledge database construction [25], [26], Web [27], [28] crowdsourced data [29]–[31], and security and privacy protection [32]–[35], so truth discovery is regarded as a good way to solve data conflicts. Unlike the poll/average approach, which treats all data sources equally, the primary purpose of truth discovery is to infer the reliability of the data source and to infer reliable information from the reliability of that data source.

Since the reliability of the data source and the truth are unknown, the general principle of truth discovery is as follows: if a data source provides reliable information frequently, the data source is given high reliability. At the same time, if the information is provided by a reliable data source, there is a good chance that it will be selected as the truth.

In the general principles of truth discovery, truth calculation and data source reliability estimation are interdependent. Some proposed truth discovery methods [13]–[17] calculate truth and source weight step by step until convergence according to the principle of iterative algorithm. In many practical cases, since the truth and the weight of the data source are unknown, only the fixed truth can be used to calculate the weight of another data source or the fixed data source weight to find the truth. In the step of truth calculation, it is assumed that the weights of the source are fixed.

In literature [18]–[21], truth is obtained through an optimization formula, whose objective function is to calculate the weighted distance between the provided information and truth. The result will be closer to the information of high-weight data source by minimizing this function. At the same time, if a data source provides data that is far from the truth, it will be assigned a lower weight to minimize the total loss. These ideas fit perfectly with the general principles of truth discovery. In this optimization formula, two sets of variables are involved, the weight of the data source and the truth. The gradient descent method can be used to obtain a solution, in which one set of variables is fixed to solve the other set of variables. This causes the truth calculation step and the data source weight estimation step to iterate until they converge, similar to an iterative approach.

Some truth discovery methods [22]–[24] are based on the probability graph model (PGM). In this model, the claims are generated based on the corresponding truth and data source weights, and the functions connect them. In the case of a Gaussian distribution, for example, the truth can be set as the mean of the distribution, and the data source weight is the precision (reciprocal of the variance). Then, the claims are sampled from the particular distribution by the truth and weights of the parameters. If the data source weight is high, the variance is relatively small, so the claims “generated” will be close to the truth. In other words, the truth is close to the claim provided by the highly weighted data source. At the same time, the claim is close to the recognized truth, and to maximize the possibility, a known variance parameter (the high weight of the data source) will be estimated.

Comparison and Summarize: All three methods are widely used in truth discovery. Through the above comparison, the three methods have their advantages and disadvantages. In terms of interpretability, the iterative method is easier to understand and explain. And methods based on optimization and PGM can be used to improve the performance of truth discovery methods when some prior knowledge about data sources and objects is available.

III. PRELIMINARIES
A. QUALITY CONTROL OF METEOROLOGICAL CLAIMS
Meteorological claims need the meteorological instrument to observe, the meteorological equipment to collect data, the meteorological monitoring software to record, data transmission and analysis, and other important links. But in any link may be inaccurate, resulting in data errors.

Common types of ground claims are continuous data (temperature, humidity), random data (precipitation, visibility), and vector data (wind direction, wind speed).

Due to differences in the stability of meteorological observation stations, the observation environment interference as well as the data transmission and other uncertain factors, such as all kinds of error often made by the meteorological claims, mainly included: (1) lack of measurement data: due to natural disasters, such as wind and rain thundersnow abnormal crash, meteorological equipment worked abnormally, resulting in that meteorological claims cannot be collected and a certain period of time the data can’t be measured; (2) data error: in the process of data collection, the sensitivity of meteorological equipment components decreases due to environmental impact or long use time, and then the data collected will be different or even wrong; (3) data consistency error: meteorological data contains multiple elements, which meet certain constraints; (4) measurement system error: in the process of meteorological monitoring, there are differences between meteorological equipment so there will be different observation results for the same object.

The quality control of meteorological claims is to check and eliminate errors in claims by some methods.

The common methods used for the quality control of meteorological claims include: check the rationality of claims, check the limit value of climate science, check the extreme value of climate, check the internal consistency and check the time consistency, etc.

Each claim must be identified to represent the quality of the data. The common identification of data quality status includes correct data, suspicious data, wrong data, positive
TABLE 1. Meaning of quality control code of meteorological claims.

| Quality control code | Describe                                |
|----------------------|-----------------------------------------|
| 0                    | Correct data                            |
| 1                    | Suspicious data                         |
| 2                    | Wrong data                              |
| 3                    | The data has a positive value           |
| 8                    | Lack of measurement data                |
| 9                    | The data is not under quality control  |

FIGURE 1. Architecture for truth discovery.

Among them, “data is positive value” means that when the original data is doubtful or missing, a value can be calculated by certain statistical methods to replace the original data. “Data as modified value” means that when the original data is doubtful or missing, a correct value can be confirmed by consulting the observation station and replacing the original data.

As shown in Table 1, in the surface meteorological claims file (A), the meaning of quality control code is as follows:

B. TRUTH DISCOVERY

As shown in figure 1, a typical truth discovery architecture [13], [36] consists of three parts: a data source, a server, and a data consumer. The main task of the data source is to collect all kinds of data (such as weather, flight information, etc.), which has the characteristics of large quantity and low cost, but the collected data is prone to conflict. For example, the height of Mount Everest on Google is 29,035 feet, but the height of Mount Everest searched by Wikipedia is 29,029 feet. The main tasks of the server are data processing, data source management, and task assignment. The data consumer sends requests to the server, the server publishes tasks to the data source, the data source perceives tasks and completes tasks, uploads the perceived data to the server, and the server processes the data and returns the results to the data consumer. Compared with traditional measurement and calculation of truth, truth discovery has three obvious advantages. (1) Participants who complete the perception task published by the server do not need to have high professional knowledge and skills, which greatly reduces the task cost and maintenance cost. (2) With high time efficiency, it can help people choose the information with high reliability quickly. However, the traditional measurement and calculation method is time-consuming and labor-consuming. (3) Data consumers can timely feedback unreasonable results to data sources through the server.

However, there are some shortcomings in truth discovery. When the data source perceives the data, it is easy to be interfered with by objective factors (such as noise, environment, fault, etc.) and subjective factors (such as bad mood, temporary troubles, etc.), resulting in an inaccurate perception of data. When the server is processing the data, the inappropriateness of the model will reduce the efficiency of truth discovery. A disturbed data source has a serious impact on truth discovery. (1) The perception data uploaded by the data source is too small, and the server cannot accurately estimate the reliability of the data source [18]. (2) Inconsistent data types uploaded by data sources increase the complexity of truth discovery [19]. (3) The data source of the fault does not upload or upload the wrong perception data, so the truth cannot be found.

IV. RANDOM SAMPLING-ARITHMETIC MEAN METHOD

In this part, the proposed random sampling-arithmetic mean (RS-AM) method proposed will be described in detail in this section. First, this section introduces some of the concepts of truth discovery and presents the problems to be solved. Then, this section reveals the reasons why arithmetic averages do not work well as truth. In order to improve the quality of meteorological claims, the RS-AM method model is proposed in this section. This method evaluates the goodness of fit between the expected distribution and the sampling distribution by repeatedly extracting the random observation vector to find the random observation vector closest to the expected distribution. Meanwhile, the RS-AM algorithm is proposed. The advantage of this method is that the distance function can be used effectively to select data sources, and an efficient truth calculation method is obtained. In the rest of this section, the stopping condition of the algorithm is studied and the convergence of the model is analyzed.
A. PROBLEM FORMULATION

In this section, the terminology and symbols used in this paper are first introduced. Then, the problem to be solved is defined.

**Definition 1:** The person or thing of interest is an object.

**Example 1:** The maximum temperature of Beijing weather is an object of interest, so the maximum temperature of Beijing weather is an object.

**Definition 2:** Claims are information or data about an object provided by the source.

**Example 2:** A weather observation station in Beijing observed that the highest temperature in Beijing today is 10°C, so 10°C is the claim.

**Definition 3:** The most reliable information about an object is truth.

**Example 3:** There are three weather observation stations A, B, and C in Beijing. A, B, and C observed that the highest temperature in Beijing today is 9°C, 10°C, and 12°C respectively. The most reliable information is 10°C, that is, the truth of the highest temperature in Beijing today is 10°C.

Suppose there is a data source set $S$ that provides claims about an object set $N$. The claims of an object may only come from a subset of the data source set $S$. The truth set of object $N$ is represented as $X$ and it is a priori unknown. For an $n$-th object, $S_n$ represents a collection of data sources that provide claims about it. All claims of object set $N$ can be expressed as $C$. The set of claims for the $n$-th object is expressed as $C_n$.

For the $s$-th source, $\delta_s$ denotes the error between the claims and the truth. The smaller $\delta_s$ is, the closer the claim is to the truth, otherwise the further the claim is from the truth.

Source Discovery. Data source discovery is formally defined as follows: given the set of claims $C$ and the set of data sources $S$, the goal of data source discovery is to obtain a set of reliable data sources, so that the arithmetic mean and median of the claims of this set of data sources are infinitely close to or equal to each other.

Inferring truth. Through data source discovery, we obtain a set of data sources that provide reliable claims. The truth is then inferred from the claims provided by this set of data sources.

Table 2 summarizes the main symbols used in this chapter.

### Table 2. Main symbols used in this chapter.

| Symbol | Definition |
|--------|------------|
| $C$    | A collection of claims for all objects |
| $C_s$  | A collection of claims for an $n$-th object |
| $x_{n,s}$ | Claims provided by data source $S$ to $n$-th objects |
| $N$    | A collection of all objects |
| $n$    | The $n$-th object |
| $S$    | A collection of all data sources |
| $s$    | The $s$-th data source |
| $X$    | The truths for all objects |
| $\delta_n$ | The truths of an $n$-th object |
| $\delta_s$ | The error between claims and truths provided by data source $S$ |

B. REASONS FOR THE UNIDEAL EFFECT OF ARITHMETIC MEAN AS TRUTH ESTIMATION

1) THE MEAN METHOD HAS BEEN USED AS THE ESTIMATION OF TRUTH IN PRECISION INSTRUMENTS FOR MANY YEARS AND HAS BEEN VERIFIED THEORETICALLY AND PRACTICALLY.

When a series of equal precision measurements are made on an object, due to random errors, the measured values are different, so the arithmetic mean of all measured values should be taken as the final measurement result.

In a series of measurements, the algebraic sum of the $n$ measured values divided by $n$ is called the arithmetic mean.

We assume that the $x_1, x_2, \cdots, x_n$ is the value obtained from $n$ measurements, then the arithmetic mean $\mu$ can be expressed as follows:

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

The arithmetic mean is closest to the measured truth. From the law of large numbers in probability theory, it can be known that if the number of measurements increases indefinitely, the arithmetic mean value $\mu$ must approach the truth $\hat{x}$.

Measurement error refers to the measurement result minus the measured truth, referred to as the error. It is shown as follows:

$$\delta = x_i - \hat{x} \quad (2)$$

where the $\delta$ for error, $x_i$ is measured value, $\hat{x}$ for measurement of truth.

By the definition of error,

$$\delta_1 + \delta_2 + \cdots + \delta_n = (x_1 + x_2 + \cdots + x_n)$$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} x_i - n\hat{x}$$

$$\hat{x} = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} \delta_i}{n} \quad (3)$$

According to the offset of the random error, it can be known that:

- When $n \to \infty$, $\sum_{i=1}^{n} \delta_i / n \to 0$, so,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \to \hat{x} \quad (4)$$

Thus it can be seen that if an infinite number of measurements can be made on a certain quantity, the measured value will not be affected by the error, or the influence will be very
small and can be ignored. This is where the arithmetic mean is considered the closest approximation to the theory of truth as the number of measurements increases indefinitely.

2) WHY THE RESULTS OBTAINED BY THE MEAN METHOD ARE ALWAYS NOT IDEAL IN THE STUDY OF TRUTH DISCOVERY?

One premise of truth discovery research is that most claims provided by sources are truth or close to the truth.

In general, the claims provided by the source about the object obey a single peak symmetric distribution, for example, the Gaussian distribution, and the t-distribution are single peak symmetric distribution. When the claims conform to asymmetric unimodal distribution, the arithmetic mean, the median, and the truth of the claims are equal. However, there are always source-supplied claims that are far from most source-supplied claims, which is defined as outlier data in this paper. Therefore, the claims, because these outlier data obey a unimodal asymmetric distribution, known as a skewed distribution, are shown in figure 2.

The highest temperature claims in some meteorological data sets were randomly selected and the histogram of claims was drawn. As can be seen in figure 2, some claims deviated from most claims, which was called the long-tail effect or long-tail phenomenon [18]. The long tail effect has a great impact on data sampling and valuation. For example, it may lead to large deviations or inaccurate data valuation. The figure shows that the maximum observed temperature distribution is left long tail, that is, the data is left-skewed or negative skewed.

When the claims were skewed distribution, the arithmetic mean was easily affected by outlier data and could not well reflect the central trend of the data. Therefore, in the truth discovery study, the mean method always obtained the worst performance. The characteristic of medians, however, is that they are not affected by outlier data. Therefore, the median is a better indicator of the central trend of the data than the arithmetic mean.

In general, the arithmetic mean and median of the claims are equal when the claims are distributed symmetrically with a single peak, but outlier data makes the arithmetic mean and median not equal or even very far apart. Therefore, we’re going to keep narrowing the distance between the arithmetic mean and the median.

**Theorem 1**: When the claims have a unimodal symmetrical distribution, then there is a value \( x' \), which makes the sum of the errors from each point of the claim to \( x' \) is 0. Then this value \( x' \) is the estimated value of the truth, that is, \( x' = \hat{x} \).

**Proof**: Let \( x_1, x_2, \cdots, x_n \) be the values obtained from \( n \) measurements, then the arithmetic mean value \( \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \).

When the claims are distributed symmetrically with a single peak, let \( m \) be the median, then \( \mu = m \).

The error \( \delta_i = x_i - x' \), where \( x_i \) is the measured value, and \( x' \) is the sum of the errors from the claims to point 0.

Since \( \mu = m \), we know from formula (1) that when \( n \) is odd,

\[
\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{m-1} x_i + x_m + \sum_{i=m+1}^{n} x_i}{n}
\]

When \( \mu = m, m-1 = n - (m+1) + 1, \) so \( n = 2m - 1 \). Therefore, it can be obtained from equation (1),

\[
2 (m-1) \mu = \sum_{i=1}^{m-1} x_i + \sum_{i=m+1}^{n} x_i = \left[ \sum_{i=1}^{m-1} x_i - (m-1) \mu \right] + \left[ \sum_{i=m}^{n-1} x_i - (m-1) \mu \right] = 0
\]

\[
\sum_{i=1}^{m-1} \delta_i + \sum_{i=m+1}^{n} \delta_i = 0
\]

Therefore, there is a value \( x' \), and the sum of the errors from each point of the claim to \( x' \) is 0, and this value is \( \mu \). Therefore, \( x' = \mu = \hat{x} \).

Similarly, when \( n \) is an even number, there is a point \( x' \), and the sum of the errors from each point of the claim to \( x' \) is 0, and this value is \( \mu \). Therefore, \( x' = \mu = \hat{x} \).

So theorem 1 is proved.

According to theorem 1 above, for the arithmetic mean to be this point, the arithmetic mean and median have to be close enough to or equal to each other. Therefore, the following model is proposed in this paper. Several groups of claims are selected randomly from a set of claims and selecting a group of claims whose arithmetic mean and median absolute errors are small enough or equal to 0.

**C. RS-AM METHOD**

The main idea of the RS-AM method model is composed of two parts. (1) Firstly, claims that meet certain conditions are extracted through repeated random sampling, which is referred to as the RS process. (2) To calculate the truth of the group of claims extracted, this part is referred to as the AM process. In this part, this paper first explains the origin of the model through an example.

Because one premise of truth discovery research is that most claims provided by sources are truth or close to the truth. To address this assumption, this paper proposes the following example. In the example, the data source that provides claims
that are truth or close to the truth is represented by a white ball, otherwise, represented by black balls.

**Example 4:** There are 10 black and white balls in a box, of which 6 are white and 4 are black. We randomly draw 5 balls from it, as one set, note the color and put it back, then we draw the next set, draw 5 sets of balls in a row. So what is the probability that at least one of the five sets of balls contains at least 4 white balls?

As can be seen from the meaning of the question, 5 balls were randomly selected from the box that was put back. There are the following five situations:

(a) The probability that all the 5 balls are drawn are white is \( p(a) = \frac{3}{126} \).

(b) Among the 5 balls drawn, the probability of 4 white balls and 1 black ball is \( p(b) = \frac{30}{126} \).

(c) Among the 5 balls drawn, the probability of 3 white balls and 2 black balls is \( p(c) = \frac{33}{126} \).

(d) Among the 5 balls drawn, the probability of 2 white balls and 1 black ball is \( p(d) = \frac{30}{126} \).

(e) Among the 5 balls drawn, the probability of 1 white ball and 4 black balls is \( p(e) = \frac{3}{126} \).

Therefore, the probability that there are at least 4 white balls in the selected groups is \( p_1 = p(a) + p(b) = \frac{33}{126} \).

Then, the probability that at least one set of balls contains at least 4 white balls is \( p_2 = 1 - \left(1 - \frac{33}{126}\right)^5 \approx 0.753 \).

Below, look at example 5 based on example 4.

**Example 5:** There are 5 black and white balls in a box, of which 4 are white and 1 are black. We take out 3 balls randomly from them as a group, write down the colors and put them back, then we take out the next group, and take out 5 groups of balls consecutively. So what’s the probability that at least one of the five sets of balls is all white?

According to the meaning of the question, 3 balls were randomly selected from the box that was put back. There are the following two situations:

(a) The probability that all the 3 balls drawn are white is \( p(f) = 0.4 \).

(b) Among the 3 balls drawn, the probability of 2 white balls and 1 black ball is \( p(g) = 0.6 \).

Therefore, the probability that at least one set of balls are all white balls among the selected sets of balls is

\[
p_3 = 1 - 0.6^{3} \approx 0.922.
\]

As can be seen from the above example, after the continuous selection of white balls, the probability of at least one group of extracted balls being all white balls is increasing, that is, after the continuous screening of claims, the probability of at least one group of randomly selected observation values being all close to or equal to truth is increasing.

1) **RS-AM METHOD**

**Definition 4:** Suppose there are 1 claims in a set of claims, i.e., \( C_n = \{x_i\} \), \( 1 \leq i \leq l \). The arithmetic mean and median can be expressed as

\[
m = \arg \min_\varphi \sum_{i=1}^{l} |x_i - \varphi|
\]

In general, the arithmetic mean \( \mu \) and median \( m \) are different.

In probability theory, the Chekhov inequality leads to the theorem that for a continuous probability distribution with the expected and median, the distance between the expected and median is no greater than the standard deviation. However, this probability distribution theorem does not mean that the median and standard deviation also apply to the arithmetic mean values of finite data sets. In order to find a finite data set theorem similar to the continuous probability distribution, theorem 2 on the relationship between the arithmetic mean and median is proposed in this paper.

**Theorem 2:** For any finite data set, the distance between the arithmetic mean and the median of the sample is never greater than the standard deviation of the sample data, i.e.,

\[
|\mu - m| \leq \sigma
\]

**Proof:** That’s Jensen’s inequality,

\[
f\left(\frac{\sum y_i}{n}\right) \geq \frac{\sum f(y_i)}{n}
\]

where \( f(\cdot) \) is any convex function.

The square root \( \sqrt{\cdot} \) is a convex function. Let’s say \( y_i = (x_i - \mu)^2 \), then we get,

\[
\frac{\sqrt{(x_i - \mu)^2}}{n} \geq \frac{\sum |x_i - \mu|}{n}
\]

Therefore, it can be obtained from equations (9) and (10) that

\[
|\mu - m| = |E(x_i - m)|
\]

\[
\leq E(|x_i - m|)
\]

\[
\leq E(|x_i - \mu|)
\]

\[
\leq \sqrt{E((x_i - \mu)^2)} = \sigma
\]

Theorem 4.2 is proved.

It can be known from theorem 4.2 that \( |\mu - m| \leq \sigma \), from which we can get

\[
\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2} \geq \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x} + \hat{x} - \mu)^2}
\]

\[
= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\delta_i - \frac{n}{n} \sum_{i=1}^{n} \delta_i)^2}
\]

where \( \delta_i \) represents the deviation of the claims from the truth, that is, \( \delta_i = x_i - \hat{x} \).

From equation (12), we know that if the arithmetic mean of the claims is close to or equal to the median of the claims, that is, \( |m - \mu| \to 0 \). Therefore, when the equation (12) tends to 0, \( |m - \mu| \to 0 \). When the equation (12) tends to 0, there are two cases (a) and (b):

(a) For \( \forall \delta_i \to 0 \), that is, \( \exists \varepsilon > 0 \), \( \varepsilon \) is arbitrarily small such that \( |\delta_i - \varepsilon| > 0 \), which satisfies the situation.

(b) When \( \delta_1 = \delta_2 = \cdots = \delta_n \), the situation is satisfying.
And because one of the assumptions of truth discovery is that most sources are reliable, providing claims close to or equal to truth. Therefore, let $Pr(\delta_t)$ denote the probability of deviation of the claims from the truth,

Event $p$: $Pr(\delta_t < \varepsilon) > 0.5$.

Event $q$: $Pr(\delta_t > \varepsilon) < 0.5$.

Therefore, event $p$ corresponds to case (a) and event $q$ corresponds to case (b). The goal of the RS-AM method is to select a set of claims that are close to or equal to the truth from a set of claims provided by the data source, that is, the goal of the RS-AM method meets (a).

The advantages and limitations of arithmetic mean and median are mutually reinforcing. This paper aims to construct a truth discovery model based on simple arithmetic operations, which has the advantages of arithmetic mean and median. At the same time, the stop criterion of iteration is set, and the output of the model is close to the arithmetic mean of the median.

Next, this paper introduces the RS process in the proposed RS-AM method, as shown in figure 3.

In the model shown in figure 3, the first sampling is performed, and $I_{k}^{(1)}$ claims are randomly selected from the claim set of length $l$. $T$ groups are selected, whose arithmetic mean and median of each group is expressed respectively as $\mu_1, \mu_1^{(1)} \cdot \mu_T, \mu_T^{(1)}$. Here $l$ represents the number of claims in the set, and $I_{k}^{(1)} > 0.5l$, which ensures that most claims are drawn each time. Let $\theta$ denote the absolute value of the deviation between the arithmetic mean and the median, calculate the $\theta$ of each group separately, and select the group of claims with the smallest $\theta$ from the $t$ groups of claims, that is, $\theta_1 = \min |m_1^{(1)} - \mu_1^{(1)}|, i = 1, 2, \ldots, t$, then take it as the data set of the next sampling claim, the length of this group of claims is $I_{k}^{(1)}$.

Then, a second sampling is performed. Randomly extract $I_{2}^{(1)}$ claims from the claim set with the length of $I_{1}^{(1)}$, select $t$ groups and the arithmetic mean and median of each group are expressed respectively as $\mu_2, \mu_2^{(1)} \cdot \mu_T, \mu_T^{(2)}$. Here $I_{1}^{(1)}$ represents the number of claims in the set, $I_{2}^{(1)} > 0.5I_{1}^{(1)}$, which ensures that most of them are drawn of claims each time. Calculate the $\theta$ of each group separately, and select the group of claims with the smallest $\theta$ from the $t$ groups of claims, that is, $\theta_2 = \min |m_2^{(1)} - \mu_2^{(1)}|, i = 1, 2, \ldots, t$, and $\theta_2 < \theta_1$, take it as the next data collection of the claims, the length of this group of claims is $I_{2}^{(1)}$.

By analogy, when the $k$-th sampling is performed, $I_{k}^{(1)}$ claims are randomly selected from the claim set with a length of $I_{k-1}^{(1)}$, select $t$ groups and the arithmetic mean and median of each group are expressed respectively as $\mu_k, \mu_k^{(1)} \cdot \mu_T, \mu_T^{(k)}$. Where $I_{k-1}^{(1)}$ represents the number of claims in the set, $I_{k}^{(1)} > 0.5I_{k-1}^{(1)}$, which ensures that most of them are drawn of claims each time. Calculate the $\theta$ of each group separately, and select the group of claims with the smallest $\theta$ from the $t$ groups of claims, that is, $\theta_k = \min |m_k^{(1)} - \mu_k^{(1)}|, i = 1, 2, \ldots, t$, and $\theta_k < \theta_{k-1}$. Sampling stops when $\theta_k$ approaches or equals $\theta$, where the arithmetic mean of the set of claims is close to or equal to the median.

However, since we randomly extract the claim from the set of claims provided by a certain attribute from the data source, the extracted claims may be close to the truth or far from the truth, resulting in the following two situations:

Case 1: Randomly selected claims are all close to truth or truth. As shown in figure 4(b).

Case 2: Some of the randomly selected claims are near truth or truth, and some are far from the truth. As shown in figure 2.

The purpose of the RS process is to select a set of claims that meet the criteria, provided that the claims are close to the truth or are truth, and that the arithmetic mean and median deviation of the claims are at a minimum of, as shown in figure 4(b). But, because this paper uses $\theta_{ij} = |m_j^{(1)} - \mu_j^{(1)}|, j = 1, 2, \ldots, k, i = 1, 2, \ldots, t$, to determine the deviation of the arithmetic mean and median, so under normal circumstances, the deviation of figure 2 in case 2 is larger than the deviation of figure 4(a) and figure 4(b) in case 1, but this does not guarantee that the claims with a deviation of $\theta_{\text{min}}$ are all close to truth or truth. For example, in figure 4(a) of case 2, the extracted claims follow a single peak symmetric distribution, in which case, the deviation between the arithmetic mean and the truth is also very small. Therefore, for each group of randomly selected claims, this paper proposes the following calculation method for the distance between the arithmetic mean and median:

$$
\Theta_{ij} = |m_j^{(1)} - \mu_j^{(1)}| + \sum_{i=1}^{t} |x_i - \mu_j^{(1)}| \tag{13}
$$
where $\Theta_{ij}$ represents the distance of the arithmetic mean and median of the $i$-th group of claims in the $j$-th sampling, $j = 1, 2, \cdots, k$, $i = 1, 2, \cdots, t$, where $m_{ij}^{(i)}$ and $\mu_{ij}^{(i)}$ respectively represent the arithmetic mean and median of the $i$-th group of claims in the $j$-th sampling, $l_{ij}^{(i)}$ represents the number of claims in each group of the $j$-th sampling.

The equation (8) consists of two parts: (1) the first part: $|m_{ij}^{(i)} - \mu_{ij}^{(i)}|$ calculates the deviation $\Theta_{ij}$ of the arithmetic mean and median of the randomly selected claims each time, which ensures that the claims with a small deviation between the arithmetic mean and median is selected from the random claims of the $t$ groups. (2) the second part: $\sum_{i=1}^{t} |x_{ij}^{(i)} - \mu_{ij}^{(i)}|$ calculated the absolute error of each group of randomly selected observations. When this group of randomly selected claims is close to the truth or the truth, the absolute error will be small. However, when some of the randomly selected claims are close to the truth and some are far from the truth, the absolute error will be larger.

Based on the first part, by calculating the absolute error of each claim, it is guaranteed that the randomly selected claim is close to truth or truth. Therefore, by calculating the distance between the arithmetic mean and the median by equation (8), a set of claims that meet the criteria can be selected.

2) CALCULATION OF TRUTH
As can be seen from the above, the RS process of the RS-AM method selects a set of claims that are all close or equal to truth. Then, the AM process of the RS-AM method is carried out to calculate the arithmetic mean value of this set of claims.
Algorithm 1 Random Sampling-Arithmetic Mean

Input: $C_n$
Output: $\tilde{x}_n$

RS Process:
1: for $j = 1$ to $k$ do
2: for $l = 1$ to $7$ do
3: Randomly sample from the claim set $C_n$ of length $l$ and select $l^{(i)}$ claims as to the claims from group $i$, $l^{(i)} > 0.5l$, and calculate the arithmetic mean $\mu_j^{(i)}$ and median $m_j^{(i)}$ of each group of claims;
4: According to equation (8), calculate the distance $\Theta_{ij}$ between the arithmetic mean and median of each group of claims;
5: Select $\Theta_{ij_{\min}}$ this group of claims as to the next randomly selected claim set $C_n'$;
6: end for
7: end for
8: until $\Theta_{ij_{\min}} \rightarrow 0$ or meet the stopping criterion $S$; select this group of claims;

AM Process
9: Calculate the arithmetic mean of the set of claims $\mu$;
10: return $\hat{x}_n = \mu$;

Therefore, for the $n$-th object,

$$\hat{x}_n = \frac{\sum_{i=1}^{l} x_i}{l}$$

(14)

where $l$ represents the number of selected claims in this group, $x_i$ represents the selected claim in this group, and $\hat{x}_n$ represents the truth of the $n$-th object.

Example 6: At some point, five meteorological observation stations A, B, C, D, and E observed that the highest temperatures in Beijing were 25°C, 27°C, 28°C, 27°C, and 29°C respectively. Firstly, the RS process of RS-AM method is used to conduct random sampling to select the qualified claims, which are 27°C, 28°C, and 27°C observed by three meteorological observation stations B, C, and D respectively. Then, through the AM process, the truth is calculated, so at a certain moment, the highest temperature in Beijing is

$$\hat{x}_n = \frac{27 + 28 + 27}{3} \approx 28.3^\circ$$

(15)

3) OVERVIEW OF RS-AM ALGORITHM

So far, this paper describes the RS-AM method. Here the paper summarizes the two steps of the overall process of the RS-AM algorithm:

(a) Select the claims that meet the conditions. Through the RS process, calculate the deviation $\Theta_{ij}$ of the arithmetic mean and median of each group of claims according to equation (48), until the group of claims of $\Theta_{ij} \rightarrow 0$ is selected.

(b) Calculate the truth. In this step, a set of claims that meet the requirements can be obtained, and then, based on this set of claims, the truth can be calculated through the AM process.

The pseudo-code of the RS-AM algorithm is as follows:

where $j = k$ represents the number of random samples or iterations.

4) SOLVE THE STOPPING CRITERION $S$

As necessary research, two subsets are defined as,

$$C_u \overset{\text{def}}{=} \{x_i | x_i \in C_n, x_i > \mu\}$$

(16)

$$C_l \overset{\text{def}}{=} \{x_i | x_i \in C_n, x_i \leq \mu\}$$

(17)

Let $n_u$, $n_l$, $\mu_u$ and $\mu_l$ represent the number of claims and the arithmetic mean values of the two subsets respectively. Obviously, $C_u \cup C_l = C_n$, $n_u \cup n_l = l$,

$$n_u\mu_u + n_l\mu_l = n\mu = n_u\mu + n_l\mu$$

(18)

Define $\delta_u \overset{\text{def}}{=} \mu_u - \mu$, $\delta_l \overset{\text{def}}{=} \mu - \mu_l$, the equation (13) can be written as

$$n_u\delta_u = n_l\delta_l$$

(19)

Now, this paper examines the possible stopping criterion $S$. A possible stopping criterion $S_1$ is to ensure that the output result $\mu$ is close to the median, and the condition to ensure is

$$S_1 (\epsilon_1) : \Delta n = |n_u - n_l| \leq \epsilon_1$$

(20)

It is easy to see that the stopping criterion $S_1$ is satisfied when $\epsilon_1 = 1$. When the number of claims is even, the arithmetic mean is the median. When the number of claims is odd, the closest claim on both sides of the arithmetic mean is the median.

However, in some special cases, the stopping criterion $S_1$ requires a large number of iterations to meet, even an unlimited number of iterations. A simple method is to limit the number of iterations $k$ and set the maximum number of iterations $\epsilon_2$ in advance. The stopping criterion is

$$S_2 (\epsilon_2) : k \geq \epsilon_2$$

(21)

In general, $\epsilon_2$ depends on the number $n$ in the claim set $C_n$. Sometimes, it might not be a linear function of $n$.

A complex stop rule $S$ could be a combination of the above, for example,

$$S = S_1 (\epsilon_1) \lor S_2 (\epsilon_2)$$

(22)

It must be noted that the above possible stopping criteria are general. For all types of images and noise, it is difficult to find an optimal stopping criterion. For some specific data sets, there must be more effective stop criteria.

5) THE CONVERGENCE THEORY ANALYSIS OF THE ALGORITHM

The RS-AM method model proposed in this section is a process in which the arithmetic mean of the set of input or selected claims approaches the median of the set of input or selected claims. Here, this paper will analyze the convergence of the RS-AM algorithm.
Converge to the median. The output results of the RS-AM algorithm converge to the median \( m \), i.e. of the claims set, \[
\lim_{k \to \infty} \mu = m
\] (23)
where \( k \) is the number of iterations.

Proof: It can be obtained from the equation (14)
\[
\delta_u + \delta_l = \frac{1}{2} l \sigma (k) \left( \frac{1}{n_u} + \frac{1}{n_l} \right)
\]
\[
= \frac{l^2}{2n_u n_l} \sigma (k)
\] (24)
And because the arithmetic mean and median of the claims in the claim set are not equal, that is, \( m \neq \mu \), there can be \( \delta_u(k) + \delta_l(k) > 2\sigma(k - 1) \).

So,
\[
\sigma (k) = \frac{2n_u n_l}{l^2} (\delta_u + \delta_l) < \frac{4n_u n_l}{l^2} \sigma (k - 1)
\]
\[
= \frac{(\delta_u + \delta_l)^2 - (\delta_u - \delta_l)^2}{\sigma (k - 1)}
\]
\[
= \left[ 1 - \left( \frac{\Delta n}{l} \right)^2 \right] \sigma (k - 1)
\] (25)
And because \( m \neq \mu, 1 \leq (\Delta n)^2 \leq l^2 \), substitute it into equation (27) and it can be obtained that,
\[
\sigma (k) < \sigma (1) \left( 1 - \frac{1}{l^2} \right)^{k-1}
\] (26)
So, when \( m \neq \mu, \)
\[
\lim_{k \to \infty} \sigma (k) = 0
\] (27)
From theorem 2, there can be \( |m - \mu| \leq \sigma \), so \( \mu - \sigma \leq m \leq \mu + \sigma \).
And by the equation (21), while \( \lim_{k \to \infty} \sigma = 0, \lim_{k \to \infty} \mu = m \).
Proven.

V. EXPERIMENTS
In this part, this paper tests the proposed RS-AM method on the real meteorological data. Experimental results show that the RS-AM method is slightly better than the existing truth discovery methods in real data sets. This paper first discusses the experimental setup in Section 4.1 and then outlines the existing methods to be compared. Finally, in section 4.2, laboratory results and analysis of meteorological data sets are introduced.

A. EXPERIMENTS SETUP
In this section, the performance metrics are introduced and the baseline methodology is discussed.

1) PERFORMANCE INDEX
The problem is that there are multi-source input claims and reference ground truth. All data conflicts are conducted without supervision, so reference truths are only used for evaluation. In this experiment, two types of data are mainly concerned: classified data and continuous data. To evaluate the performance of various information conflicts or data conflict resolution methods, the following measures are adopted for meteorological claims.

(a) MAE: For continuous data, this paper uses the absolute mean error (MAE) metric, which measures the absolute mean distance from the truth discovery method output value to the ground truth. The calculation method is as follows:
\[
MAE = \frac{1}{n} \sum_{i=1}^{n} ||\hat{x}_n - x_n||
\] (28)
(b) RMSE: For continuous data, we can also use the root mean square error (RMSE) measurement. Compared with the average absolute error (MAE), this measurement penalizes larger distances more and less penalizes smaller distances. The calculation method is as follows:
\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} ||\hat{x}_n - x_n||}
\] (29)
(c) Error Rate: For classified data, the error rate is measured by calculating the percentage of mismatches between the output value of each method and the ground truth. For continuous data, the error rate can also be measured, where “mismatch” refers to when the distance from the reference truth is greater than a threshold (for example the threshold is 0.1% of the reference truth).
For all three measures, the lower the measure, the closer the estimation of the method is to the actual situation. For all measures, the lower the value, the better the method performance.

2) BASELINE METHOD
This paper compares the proposed RS-AM approach to the baseline approach below, which covers various approaches to resolving data conflicts. These methods can be divided into three categories:
The conflict resolution method applies only to continuous data. The following methods are suitable for continuous data because they ignore input from classification attributes.
- Mean: The Mean method simply takes the arithmetic mean of all claims on each property of each object as the final output.
- Median: The median method calculates the median of all claims for each attribute of each object as the final output.
- GTM [22]: Gaussian truth model (GTM) is a Bayesian probability method, which is especially suitable for solving conflicts on continuous data. Note that this method applies only to some data (continuous), while other truth discovery methods apply to all data (classified data and continuous data). Therefore, insufficient data can cause GTM performance to degrade when compared with other methods. However, because GTM is an important truth discovery method applied to continuous data, this paper still includes GTM in the comparison.
Among them, mean method and median method are traditional methods to solve data conflicts, and GTM is a truth discovery method considering source reliability estimation.

The conflict resolution method applies only to classified data. For the classification attribute, the majority voting method is adopted, which is the traditional method to solve the classification data conflict, and there is no source reliability estimation.

- **Voting**: the voting of the most frequent claims is the final output.

Conflict resolution based on truth discovery. Many existing data conflict resolution methods that consider source reliability (often referred to as “truth discovery” methods) aim to find the true “facts” for classification attributes. However, we can deal with conflicting data by continuously observing them as “facts”.

- **Investment** [37]: In this method, a source uniformly “invests” its reliability in the claims it provides, and the confidence of the claims grows according to a nonlinear function defined by the sum of the reliability of its provider’s “investments”. The sources then determine the reliability of the confidence of their claims.

- **Truth Finder** [14]: Truth Finder adopts the Bayesian analysis method. For each claim, its confidence is calculated as the product of its provider’s credibility. Adjusts a vote for a value by considering a similar function that takes into account the effect between facts.

- **AssuSim** [13]: AssuSim also uses Bayesian analysis and adopts the method of similarity function. At the same time, taking into account the supplementary votes adopted by 2-Estimates and 3-Estimates, other algorithms to solve the problem of source dependence in conflicts are proposed in [1]. As source dependence is not considered in this chapter, these algorithms are not compared in this chapter.

- **CRH** [19]: The conflict resolution based on heterogeneous data (CRH) is a framework for processing heterogeneous data. The truth is inferred from the claims provided by multiple conflicting sources. Each conflicting source involves multiple data types, and all data types are combined to estimate the reliability of the source.

- **CATD** [18]: Confidence-based truth discovery (CATD) is to infer the truth from conflicting data with the long tail phenomenon. The long tail phenomenon is a phenomenon in which many sources provide few claims. In this confidence-based truth discovery method, the confidence interval of the source reliability estimate can be obtained.

Comparing the proposed RS-AM method with these baseline methods, we can find that: 1) The RS-AM method is to randomly select a set of claims provided, so that the arithmetic mean and median of this set of claims are infinitely close to or equal, and then select a group of data sources that meet the requirements without estimating the reliability of the data sources; 2) After selecting the data sources that meet the requirements, calculate the arithmetic average of the claims provided by this group of data sources to infer the truth.

Based on the code provided by the authors of this paper, these baseline methods were implemented on Windows with 16 gigabytes of memory and an Intel I5 processor.

### B. EXPERIMENTS SETUP

In this section, the paper compares the RS-AM approach with the baseline approach, demonstrating the modeling capabilities under continuous and classified data, and demonstrating the effectiveness of the approach on real data sets.

1) **REAL DATA SET**

In this paper, the validity of the method is proved by using real meteorological claims.

Meteorological data set. The weather data integration task is a good test platform. Specifically, this paper integrates meteorological data collected from three platforms: Wunderground (http://www.wunderground.com.), HAM weather (http://www.hamweather.com.), and World weather Online (http://www.worldweatheronline.com.). For each object, we collect data of four properties: high temperature, low temperature, air pressure, and weather conditions, the first three of which are continuous data and the last one is classified data. Due to the small amount of pressure data collected, it is not used as the research object in this experiment. To obtain the ground truth and for evaluation, this paper collects real weather data of 20 American cities for 29 days, as shown in table 3.

As can be seen from table 3, there are 17,723 claims, 1,940 objects, and 1,700 ground truth. It should be noted that the number of objects in the table is greater than or equal to the number of reference ground truth because here there is only one ground truth for each object in the data set and some objects have not collected the true ground truth. Ground truth is not used as a data set. It only evaluates the performance of the method in the experiment.

| # Claims | 17723 |
|----------|-------|
| # Object | 1940  |
| # Ground truth | 1700 |

2) **ANALYSIS OF EXPERIMENTAL RESULTS**

Obtaining a set of high-quality claims is the key to truth discovery. Therefore, based on this point, this paper proposes the RS-AM method combining the advantages of existing methods.
The experimental performance of the truth discovery method on the meteorological data set is shown in Table 4, including MAE, EMSE, and error rate. As can be seen from Table 4, compared with CRH, CATD, GTM, and other baseline methods, the RS-AM method proposed by us achieves better performance on meteorological data sets, and the overall performance is slightly better than the existing truth discovery methods. Through experiments, the RS-AM method is the lowest among all methods in MSE and RMSE, with 3.9269% and 4.9626%, respectively. In terms of the error rate, although the RS-AM method is not the lowest. But the error rate of the two methods is basically the same as that of the CRH method with the lowest error rate. Therefore, MAE was reduced by 1.5% and RMSE by 2.8% compared with the best existing baseline method, CRH, while the error rates were basically flat.

At the same time, we can also see from Table 4 that Mean, Median, and Voting only aggregate multi-source information without considering the reliability of data sources among all the methods, that is, whether the data sources that do not provide data are reliable or provide real data, so their experimental performance is relatively poor. Investment, Truth Finder, and AccuSim input data in a classified way. Therefore, when processing continuous data, the experimental performance is not very ideal. However, considering the reliability of data sources, their experimental performance is slightly improved compared with Mean, Median, and Voting methods. Therefore, in this sense, methods such as GTM, CRH, and CATD are more suitable for dealing with continuous data types. Meanwhile, they take into account the reliability of data sources, so the experimental performance is improved compared with the six methods including Mean, Median, Investment, and AccuSim. Based on the advantages of these existing methods, the idea of RS-AM proposed in this paper is simpler, and the computational complexity is greatly reduced by determining and selecting the measured values that meet the requirements through random extraction. And, it gets better performance. Compared with the existing methods, the performance of the RS-AM method is slightly better than other truth discovery methods because the RS-AM method does not need to estimate the reliability of data sources, but only needs to select a set of data sources that meet the requirements and then conduct truth calculation.

**VI. CONCLUSION**

Due to errors, missing data, typing errors, outdated data, and other reasons, the meteorological data collected are likely to conflict with each other, which brings difficulties to the quality control of meteorological data. However, most of the existing truth discovery methods consider the reliability of data, which increases the difficulty of calculation. Therefore, we propose an RS-AM method to solve the conflict in meteorological data. In this work, the observation values that meet the convergence conditions are extracted by repeated random sampling, and then the truth values are calculated. The advantage of this method is that it does not need to calculate the weight of the data source, and at the same time, it reduces the complexity of the operation and deduces the truth without reducing the accuracy of the method. We experimented with this method in a real meteorological data set. The results show that this method is more effective and better than the existing truth discovery methods in the quality control of meteorological data.

**TABLE 4. Performance comparison of the meteorological data set.**

| METHODS   | MSE   | RMSE  | Error Rate |
|-----------|-------|-------|------------|
| Mean      | 4.9906| 6.4952| 0.4823     |
| Median    | 4.9903| 6.4987| 0.4526     |
| GTM       | 4.7448| 6.1356| 0.2736     |
| Voting    | 4.7953| 6.1865| 0.4899     |
| Investment| 4.8875| 6.8762| 0.2928     |
| Truth Finder | 4.7573| 6.6534| 0.2468     |
| AccuSim   | 4.5278| 6.2435| 0.2653     |
| CRH       | 3.9867| 5.1055| 0.1698     |
| CATD      | 4.0632| 5.2899| 0.1845     |
| RS-AM     | 3.9269| 4.9626| 0.1699     |
control of meteorological data. In the future, we plan to adapt the box model to more complex data quality issues, such as addressing semi-structured data quality issues.

REFERENCES

[1] M. Hlafa, Y. Zouhairi, M. Lemoudden, and E. Ziyati, “An approach for meteorological data integration and stream processing,” in Proc. 3rd Int. Conf. Syst. Collaboration (SysCo), Nov. 2016, pp. 1–5.

[2] S. A. Nishe, T. A. Tahrin, N. Kamal, M. D. S. Hoque, and K. T. Hasan, “Micro-level meteorological data sourcing for accurate weather prediction,” in Proc. IEEE Region Humanitarian Technol. Conf. (R-HTC), Dec. 2017, pp. 353–356.

[3] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, “People on drugs: Credibility of user statements in health communities,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (SIGKDD), Aug. 2014, pp. 65–74.

[4] C. Miao, W. Jiang, L. Su, Y. Li, S. Guo, Z. Qin, H. Xiao, J. Gao, and K. Ren, “Cloud-enabled privacy-preserving truth discovery in crowd sensing systems,” in Proc. 13th ACM Conf. Embedded Netw. Sensor Syst. (SenSys), 2015, pp. 183–196.

[5] D. Wang, L. Kaplan, and T. F. Abdelzaher, “Maximum likelihood analysis of conflicting observations in social sensing,” ACM Trans. Sensor Netw., vol. 10, no. 2, pp. 1–27, Jan. 2014.

[6] D. Yu, H. Huang, T. Cassidy, H. Ji, C. Wang, S. Zhi, J. Han, C. Voss, and M. Magdon-Ismail, “The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding,” in Proc. Int. Conf. Comput. Linguistics (COLING), Aug. 2014, pp. 1567–1678.

[7] F. Li, M. L. Lee, and W. Hsu, “Entity profiling with varying source reliabilities,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), Aug. 2014, pp. 1146–1155.

[8] X. Luo, W. Qin, A. Dong, K. Sedraoui, and M. Zhou, “Efficient and high-quality recommendations via momentum-incorporated parallel stochastic gradient descent-based learning,” IEEE/C 2A J. Autom. Sinica, early access, Sep. 24, 2020.

[9] X. Luo, M. Zhou, S. Li, and M. Shang, “An inherently nonnegative latent factor model for high-dimensional and sparse matrices from industrial applications,” IEEE Trans. Ind. Informat., vol. 14, no. 5, pp. 2011–2022, May 2018.

[10] X. Luo, M. Zhou, S. Li, Y. Xia, Z.-H. You, Q. Zhu, and H. Leung, “Incorporation of efficient second-order solvers into latent factor models for accurate prediction of missing QoS data,” IEEE Trans. Cybern., vol. 48, no. 4, pp. 1216–1228, Apr. 2018.

[11] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari, and A. Alabdulwahab, “Generating highly accurate predictions for missing QoS data via aggregating nonnegative latent factor models,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 3, pp. 524–537, Mar. 2016.

[12] X. Luo, H. Wu, H. Yuan, and M. Zhou, “Temporal pattern-aware QoS prediction via biased non-negative latent factorization of tensors,” IEEE Trans. Cybern., vol. 50, no. 5, pp. 1798–1809, May 2020.

[13] X. L. Dong, L. Berti-Equille, and D. Srivastava, “Integrating conflicting data: The role of source dependence,” VLDB Endowment, vol. 2, no. 1, pp. 550–561, Aug. 2009.

[14] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, “Corroborating information from disagreeing views,” in Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM), 2010, pp. 131–140.

[15] J. Pasternack and D. Roth, “Knowing what to believe (when you already know something),” in Proc. 23rd Int. Conf. Comput. Linguistics (COLING), Aug. 2010, pp. 877–885.

[16] X. Yin, J. Han, and P. S. Yu, “Truth discovery with multiple conflicting information providers on the Web,” in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2007, pp. 1048–1052.

[17] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, “Truth inference in crowdsourcing: Is the problem solved?” Proc. VLDB Endowment, vol. 10, no. 5, pp. 541–552, Jan. 2017.

[18] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, “A confidence-aware approach for truth discovery on long-tail data,” Proc. VLDB Endowment, vol. 8, no. 4, pp. 425–436, Dec. 2014.

[19] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, “Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation,” in Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD), Jun. 2014, pp. 1187–1198.
JIN ZHANG received the B.S. degree in communication engineering and the M.S. degree in computer application from Hunan University, Changsha, China, in 2002 and 2004, respectively, and the Ph.D. degree in biomedical engineering from Zhejiang University, Hangzhou, China, in 2007.

From 2009 to 2011, he held a postdoctoral position with Beijing Normal University. From 2012 to 2013, he held a postdoctoral position with The University of Chicago. Since 2014, he has been a Professor with Hunan Normal University. His research interests include intelligent software engineering and artificial intelligence.

LINGYU CHEN was born in Huanggang, China, in 1997. She received the B.S. degree in communications engineering from Hubei Normal University, Huangshi, China, in 2019. She is currently pursuing the master’s degree in software engineering with Hunan Normal University, China.

Her main research interest includes artificial intelligence and neural networks.

HONG LIU received the B.S. degree in computer software from the Huazhong University of Science and Technology, Wuhan, China, in 1983, and the M.S. degree in computer application from the National University of Defense Technology, Changsha, China, in 1988.

From 1983 to 1985, he worked as a Teacher with Hebei University. Since 1988, he has been a Teacher with Hunan Normal University. He is currently a Vice President of the College of Information Science and Engineering, Hunan Normal University and also in charge of Undergraduate Teaching. His research interests include software engineering and artificial intelligence.

YING WANG received the B.S. degree in software engineering and the M.S. degree in public administration from Hunan University, Changsha, China, in 2005 and 2008, respectively.

She has been an Associate Professor with Hunan University of Chinese Medicine, and the Director of the Department of Administration and Social Security. Her main research interests include administration, health economics, and policy.