METHODS FOR OUTLIER DETECTION IN METROLOGICAL STUDIES

Valeriy Aschepkov, PhD student,
National University of Radio Electronics,
Kharkiv, Ukraine, Email: ashhepkovvalera@gmail.com

https://doi.org/10.23939/istcmtm2024.03.025

Abstract. The article addresses the issue of outliers in metrological measurements, which can significantly distort research results and affect measurement accuracy. Outliers that substantially differ from other data points in a sample seriously threaten the reliability of metrological processes. In previous studies, the Isolation Forest model was applied to detect such outliers, demonstrating its effectiveness under certain conditions. For a deeper understanding and validation of the results, it is necessary to compare this approach with traditional robust methods, such as the Interquartile Range (IQR) and Median Absolute Deviation (MAD), already widely used in metrology.

This work compares the mentioned outlier detection methods with the Isolation Forest model. Special attention is given to the impact of outliers on data distribution and each method's ability to impact mitigation, enhancing reliability. The study encompasses an analysis of the characteristics of the method for the identification of strengths and weaknesses in the context of real metrological tasks.

Key words: Metrology, outliers, anomalies, uncertainty, error, Isolation Forest method, robust methods.

1. Introduction

In previous studies, we addressed the issue of outliers in metrological measurements [1]. The use of the Isolation Forest model for outlier detection demonstrated its effectiveness under certain conditions, showing the ability to identify anomalies that could distort measurement results and lead to an increase in Type A standard uncertainty. To gain a deeper understanding and validate the results, this method is compared with traditional robust methods.

Outliers pose a significant problem in the context of metrological research, as they can negatively impact the accuracy and reliability of measurements. Understanding their nature and impact on statistical indicators is crucial for developing effective methods for their detection and elimination [2-3]. While robust methods such as the Interquartile Range (IQR) and Median Absolute Deviation (MAD) are widely applied, new approaches like the Isolation Forest also require detailed analysis and evaluation

2. Goal

The comparison of the robust methods for outlier detection, such as the Interquartile Range (IQR) and Median Absolute Deviation (MAD), with the Isolation Forest model to improve the accuracy of measurements in the obtained data.

3. Outlier Detection Methods

Robust methods are based on the calculations/use of median absolute deviation (MAD) or interquartile range (IQR).

The **Interquartile Range** calculates the difference between the third quartile (Q3, upper quartile) and the first quartile (Q1, lower quartile). Quartiles are values that divide an ordered data set into four equal parts. Specifically:

- The first quartile (Q1) is the value below which 25% of the data lies;

The third quartile (Q3) is the value below which 75% of the data lies.

The formula for calculating the Interquartile Range is as follows:

$$IQR = Q_3 - Q_1 . \tag{1}$$

To identify outliers, the following rule is used: data points are considered outliers if they fall outside the following thresholds:

$$\textit{Lower threshold} = Q_1 - 1.5 \times IQR,$$
$$\textit{Upper threshold} = Q_3 + 1.5 \times IQR. \tag{2}$$

This method does not require assumptions about the data distribution and is robust to the influence of outliers, making it useful for analyzing small samples.

The **Median Absolute Deviation** (MAD) is the median of the absolute deviations of observations from the data median. It measures the dispersion of data around its median, providing additional robustness to the influence of outliers. The formula for calculating the Median Absolute Deviation (MAD) is:

$$MAD = M_e\left(\left|X - M_e(X)\right|\right), \tag{3}$$

where M_e is the median operator, $M_e(X)$ is the median of the random variable X. The median is the central value of a sample, which divides the data into two equal parts. Using the median provides a more stable estimate of the central tendency for samples with outliers. To identify outliers, values that significantly exceed the calculated MAD (typically by a multiple of 2 or 3 MAD) may be considered outliers [4].

The **Isolation Forest method** is based on the isolated anomalous data points by constructing random trees. The main idea is that anomalous points, or outliers, are easier to isolate from the majority. To achieve this, the method constructs a collection of trees (an isolation forest), where each tree is generated by randomly partitioning the data into sub-samples. At each node of tree, the data is split based on randomly selected features and their values until individual points are fully isolated.

After the isolation trees are built, the depth at which each data point is located is calculated. Points isolated at shallower depths in the tree are considered potential outliers. An anomaly score is calculated for each point, indicating how likely it is to be an outlier. According to the anomaly scores obtained, all data points are classified as outliers or normal points based on a predefined threshold. If the anomaly score exceeds the determined limit, the point is considered as an outlier.

4. Calculation Results

Several studies were conducted while preparing the "State primary standard of units of volume and mass flow of liquid, volume, and mass of liquids flowing through the pipeline" [5]. As part of one of these studies, the performance of three Coriolis flow meters of different diameters was measured at three different levels of liquid mass flow. The results of these measurements were processed according to the methodologies described in the final reports of the EUROMET international comparisons [6-11]. The study confirmed that the measurement results comply with international standards. This research highlighted the need to reduce Type A standard uncertainty by promptly identifying outliers in the measurement data and subsequently excluding them while processing the measurement results.

During the measurements of the Coriolis flow meters, the relative error was calculated using the following formula:

$$\varepsilon = \frac{(m_v - m_{ref})}{m_{ref}} \cdot 100\ \% \cdot \qquad (4)$$

where m_{ref} is the reference liquid mass value, m_v is the liquid mass value measured by the Coriolis flow meter.

The values of the relative measurement error at a given liquid flow rate represent a sample in which the presence of outliers needs to be identified. There were 9 samples, with sizes ranging from 16 to 33 repeated measurements. Due to the small sample size, calculating the data distribution may not be accurate, given the high risk of outliers affecting the distribution shape. Therefore, robust methods were specifically used to detect outliers.

Before conducting the calculations, robust methods require sorting the sample values in ascending order. The generally accepted thresholds for outlier detection methods are as follows:

- Interquartile range (IQR) method: 1.5;
- Median absolute deviation (MAD) method: 3.

However, based on the data from the studies and considering that the measurements were conducted on the state primary standard using Coriolis flow meters with an accuracy of 0.1%, the standard thresholds typically used for robust methods would not have identified any significant outliers. This is due to the high stability of the standard and the instruments, which reduces the likelihood of outliers occurring. Nonetheless, even minor deviations can be important for metrological studies and may impact Type A standard uncertainty. Lowering the thresholds provides the detection of potential deviations that might have gone unnoticed before.

For this purpose, the thresholds for the methods were set as follows:

– Interquartile Range (IQR) method: 0.4;

– Median Absolute Deviation (MAD) method: 1.5.

Fig. 1 shows the result of calculating one sample using different methods, where the presence of outliers can be visually assessed, and how these outliers are identified by applying methods.

The result of the outlier detection calculations is presented in Table 1.

Outlier detection reduces Type A standard measurement uncertainty, which calculation results are presented in Table 2.

The values from Table 2 are shown in Fig. 2.

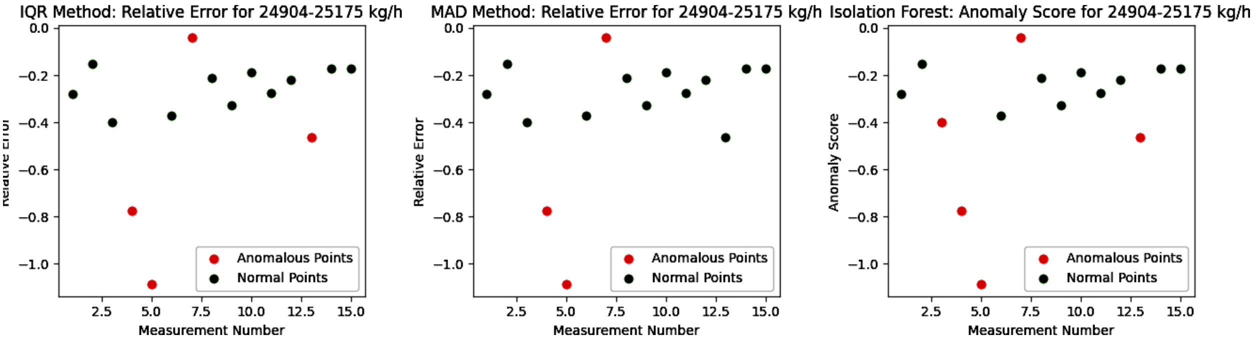

Fig 1. Detection of outliers using different methods at a mass flow rate of 25 t/h

Table 1. Results of outlier detection calculations

Mass flow rate point	Sample Size	Number of outliers detected with the IQR method	Number of outliers detected with the MAD method	Number of outliers detected with the Isolation Forest method
Flow meter №.1: 45 t/h	17	2 (11.26%)	5 (29.41%)	5 (29.41%)
Flow meter №.1: 25 t/h	15	3 (20%)	3 (20%)	4 (26.66%)
Flow meter №.1: 5 t/h	18	2 (11.11%)	2 (11.11%)	3 (16.66%)
Flow meter №.2: 5 t/h	17	3 (17.64%)	3 (17.64%)	5 (29.41%)
Flow meter №2: 2.5 t/h	16	2 (12.5%)	3 (18.75%)	5 (31.25%)
Flow meter №2: 1 t/h	33	9 (27.27%)	8 (24.24%)	10 (30.30%)
Flow meter №3: 1 t/h	17	5 (29.41%)	5 (29.41%)	6 (35.29%)
Flow meter №3: 0.5 t/h	33	2 (6.06%)	2 (6.06%)	8 (24.24%)
Flow meter №3: 0.1 t/h	16	4 (25%)	3 (18.45%)	5 (31.25%)

Table 2. Type A standard measurement uncertainty

Mass flow rate point	Type A standard measurement uncertainty,%			
	Before outlier removal	Removal by IQR method	Removal by MAD method	Removal by Isolation Forest method
Flow meter №.1: 45 t/h	3.08E-02	1.62E-02	2.55E-02	2.20E-02
Flow meter №.1: 25 t/h	6.94E-02	2.95E-02	2.95E-02	2.55E-02
Flow meter №.1: 5 t/h	1.74E-02	1.22E-02	1.22E-02	1.30E-02
Flow meter №.2: 5 t/h	2.01E-02	1.38E-02	1.38E-02	1.32E-02
Flow meter №2: 2.5 t/h	2.51E-02	2.07E-02	2.28E-02	1.88E-02
Flow meter №2: 1 t/h	1.04E-02	4.90E-03	4.80E-03	4.45E-03
Flow meter №3: 1 t/h	2.24E-02	5.65E-03	5.65E-03	5.65E-03
Flow meter №3: 0.5 t/h	1.67E-02	1.53E-02	1.53E-02	1.58E-02
Flow meter №3: 0.1 t/h	5.87E-02	3.87E-02	3.54E-02	3.31E-02

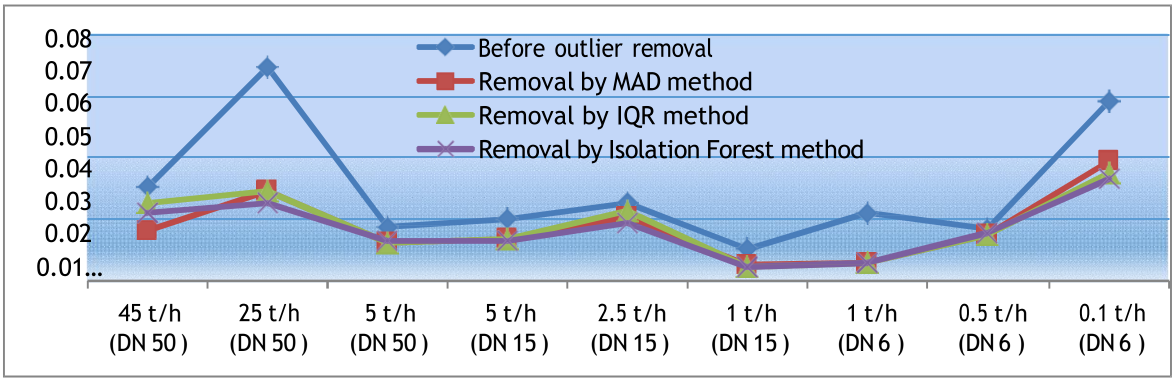

Fig 2. Calculation of Type A standard uncertainty for relative measurement error.

1. Comparison of Methods

Robust methods, like the Isolation Forest method, are insensitive to data distribution. However, the effectiveness of robust methods may decrease when working with data that exhibits a high degree of skewness. In such cases, the distribution can impact the accuracy of outlier detection when using robust methods.

One of the key advantages of robust methods is their simplicity and the absence of the need for complex tuning. These methods are easy to apply and do not require deep specialized knowledge, making them accessible in various scenarios. In contrast, the Isolation Forest method requires tuning several parameters, such as the number of trees in the ensemble and the proportion of outliers in the data, which may require certain expertise and experience to achieve optimal results.

Robust methods are based on defining a range of data values within which values are considered normal. Any values that fall outside this range are considered outliers. This approach makes robust methods particularly useful for detecting outliers in data where the main cluster of values is clearly defined, and any deviations from it are easily identified as anomalies (Fig. 3).

However, if the data cluster is split into two clusters, the Isolation Forest method may consider this as a normal condition. In Fig. 4, we can see how the model identifies two points as anomalies. To reduce the uncertainty based on the calculations of the mean values and the tendency of the data clustering, this aspect is considered as a disadvantage.

An important advantage of the Isolation Forest method is that the model transitions to other values based on the calculations, specifically to the anomaly score, which indicates how anomalous each point is relative to the most points in the sample. This allows for a visual assessment and observation of the anomaly level of each point, unlike robust methods. Based on this visual assessment, a more accurate threshold can be set here, and samples can be combined for a more comprehensive analysis and investigation of the causes of outliers. The possibilities of this approach are described in more detail in the article "Application of the Isolation Forest model for anomaly detection in measurement data" [1].

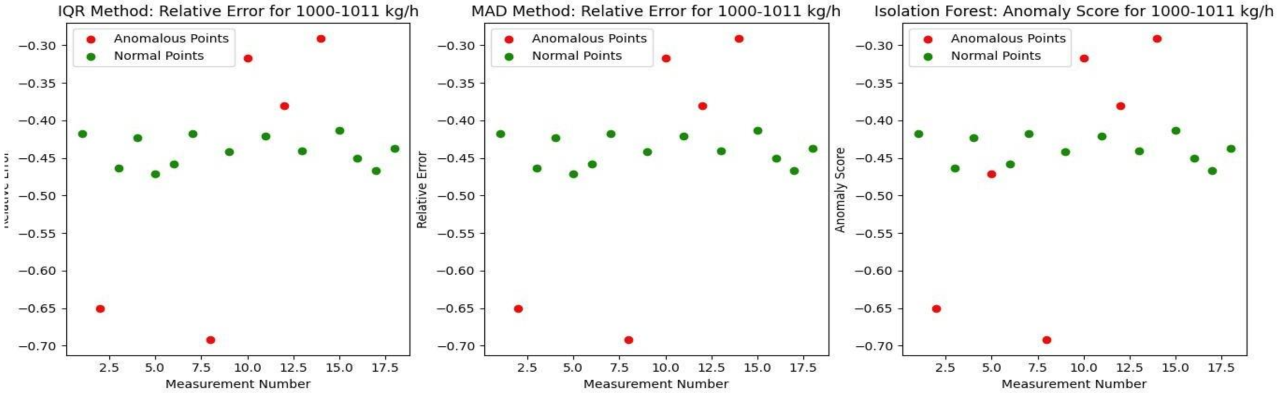

Fig. 3. Detection of outliers using different methods at a mass flow rate of 1 t/h.

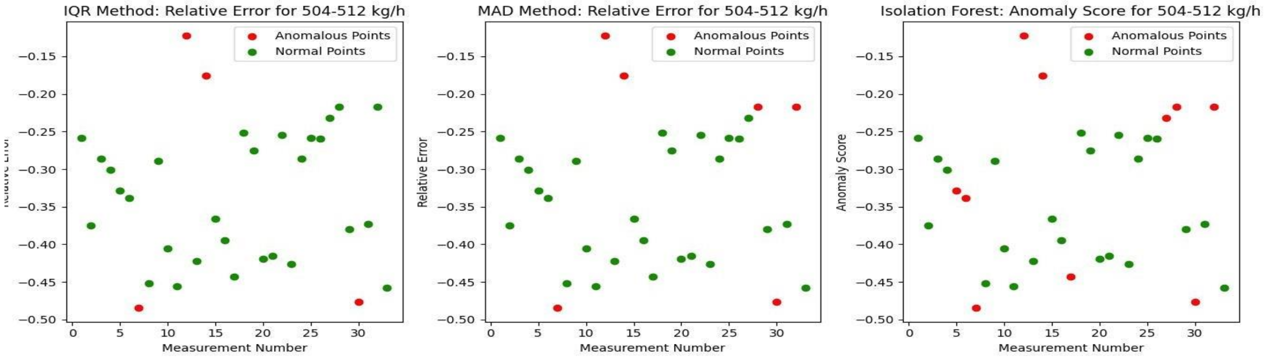

Fig. 4. Detection of outliers using different methods at a mass flow rate of 0.5 t/h.

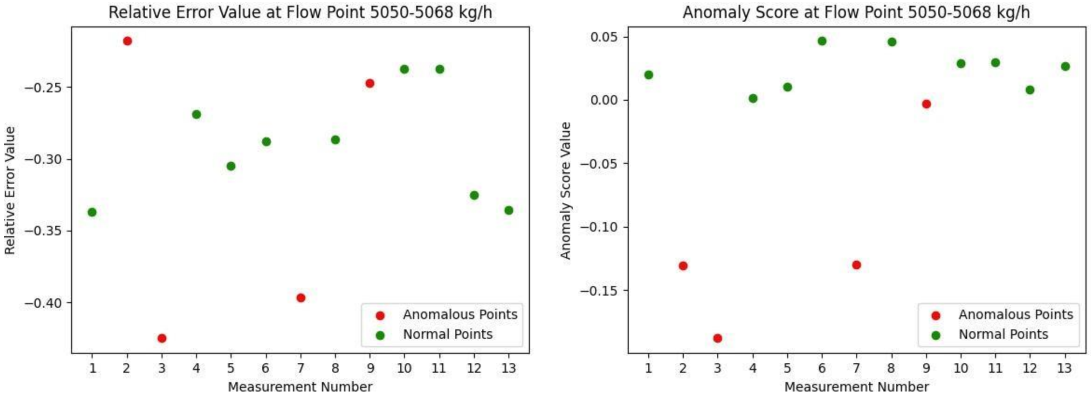

Fig. 5. Detection of outliers using different methods at a mass flow rate of 1 t/h.

2. Conclusions

As a result of the conducted research, three considered outlier detection methods — robust methods such as IQR, MAD, Isolation Forest method — successfully identified outliers in the metrological data. The detected outliers accounted for no more than 35.29% of the total sample.

These methods effectively reduced the Type A standard uncertainty, yielding similar results. However, there have been differences in the number of detected outliers and the corresponding uncertainty values. This may be caused by the dissimilar thresholds inherent in each considered method.

The Isolation Forest method allowed for a more sensitive threshold to outliers and reduced the uncertainty values. This indicates a direct relationship between the threshold value and the level of uncertainty, as a more sensitive threshold provides a more precise distinction between normal data and anomalies.

Further research is needed to explore the impact of threshold values on uncertainty levels and to develop a methodology that allows for optimal threshold setting, taking into account the specific characteristics of the data and ensuring more accurate outlier detection

3. Acknowledgment

I would like to express my sincere gratitude to scientific supervisor, Dr. Sc. V.V. Sklyarov, for his valuable advice and significant contribution to the research. I also thank the Department of Information and Measurement Technology at the Kharkiv National University of Radio Electronics. Your advice and guidance have made this research more successful and effective.

Furthermore, extend my profound thanks to the Scientific Center for Mechanical Measurements of the National Scientific Center "Institute of Metrology" for the useful discussions, constructive criticism, and comments.

References

[1] V.O. Ashchepkov "Application of the Isolation Forest Model for Anomaly Detection in Measurement Data" in *Innovative Technologies and Scientific Solutions for Industries* 2024, No. 1 (27), doi:10.30837/ITSSI.2024.27.236.

[2] T. V. Potanina, I. V. Mikhaylenko, "Investigation of Experimental Data Samples for Outliers: Comparison of Methods" in Integrated Technologies and Energy Saving, No. 3, 2023, doi: 10.20998/2078-5364.2023.3.07

[3] Wada K. “Outliers in official statistics” in *Japanese Journal of Statistics and Data Science,* 2020. No 3.pp. 669–691. doi:10.1007/s42081-020-00091-y

[4] M. Orellana and P. Cedillo, "Outlier Detection with Data Mining Techniques and Statistical Methods," 2019 International Conference on Information Systems and Computer Science (INCISCOS), Quito, Ecuador, 2019, pp. 51-56, doi: 10.1109/INCISCOS49368.2019.00017

[5] V.O. Aschepkov «Research of metrological characteristics of the state primary standard of unit of volume and mass flow rate of liquid during preparation for participation in international comparisons» in Ukrainian Metrological Journal. 2024. No.1 (77) doi: 10.24027/2306-7039.1.2024.300937.

[6] Batista E., Lau P. EURAMET regional key comparison EURAMET.M.FF-K4.b: Volume intercomparison at 20 L. Metrologia, 2009, vol. 46(1A):07013.doi: 10.1088/0026-1394/46/1A/07013

[7] Malengo A., Batista E., Arias R., Mićić L., Bošnjaković A., Mirjana M., Piluri E., Svendsen G., Huu M., Sarevska A. and others. Final report on EURAMET project 1395/EURAMET.M.FF-K4.1.2016: volume comparison at 20 L. Metrologia, 2020. vol. 57(1A):07021. doi: 10.1088/0026-1394/57/1A/07021

[8] Huovinen M., Frahm E. EURAMET.M.FF-S13final report. Metrologia, 2022, vol. 59(1A):07010. doi: 10.1088/0026-1394/59/1A/07010

[9] Geršl J., Lojek L. Final report on EURAMET project No. 1046: Intercomparison of water flow standards using electromagnetic flowmeters. Metrologia, 2013, vol. 50(1A):07002. doi: 10.1088/0026-1394/50/1A/07002

[10] Batista E. Final report on EUROMET key comparison EUROMET.M.FF-K4 for volume intercomparison of 100 ml Gay-Lussac pycnometer. Metrologia, 2006, vol. 43(1A):07009. doi: 10.1088/0026-1394/43/1A/07009

[11] Benkova M., Frahm E., Romieu K., Warnecke H., Büker O., Haack S., Akselli B., Mazur V., Berkmann C., Zygmantas G. Comparisons of standards for liquid flow rates under static load changes. Metrologia, 2024, vol. 61(1A):07003. doi: 10.1088/0026-1394/61/1A/07003