ISSN 1810-2328, Journal of Engineering Thermophysics, 2024, Vol. 33, No. 1, pp. 1–8. © Pleiades Publishing, Ltd., 2024.

On Applicability of IQR Method for Filtering of Experimental Data

B. B. Ilyushin*

Kutateladze Institute of Thermophysics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia

Received September 4, 2023; in final form, December 11, 2023; accepted February 2, 2024

Abstract—The results of testing the popular IQR (Interquartile Range) method for filtering experimental data are presented. It is shown that if the distributions of measured values differ greatly from the Gaussian distribution, this method gives a large error in the statistical characteristics, especially the higher moments. The earlier-developed statistical filtering method can take into account substantial skewness of distributions of measured values and can greatly reduce the filtering error.

DOI: 10.1134/S1810232824010016

INTRODUCTION

In many applications, processing and analysis of large volumes of measurements face appearance of anomalous data (outliers). In this case, it is important to determine whether a new anomalous observation belongs to the same distribution as the existing ones or it should be considered as a manifestation of new properties or phenomena of the object under study [1–3]. In many cases, such anomalies associated with emergence of new properties are the main purpose of analysis of obtained data, e.g., in cybersecurity [4, 5], medicine [6–8], biology [9], law machinery (financial fraud) [10, 11], etc. (see [2, 12–14]). Currently, neural-network and machine-learning methods for detecting such outliers are being actively developed and implemented [15, 16]. They enable automation of the process of search and analysis of anomalous outliers in a sequence of events and determination of their possible source and the degree of hazard they impose (see, for example, [17, 18]). Anomalous outliers may be due to measurement error or noise. When data are analyzed, such outliers can lead to an inflated error and significant distortions of parameters and statistical estimates (see, for example, [19–22]). Such "parasitic" outliers should be identified and discarded in further analysis. Statistical justification for dropping such outliers can be found in works dating back to the century before last [23]. Currently, filtering (suppression) of erroneous measurements has become an integral part of preprocessing of experimental data (see, for example, [24, 25]). It is often automated and is part of the software of a measuring equipment (see, for example, [26]).

There are a lot of works on methods for search and analysis of anomalous outliers (see, for example, reviews [12, 27–30]). The earliest ones are based on the basic assumption of identity and independent distribution of data, their mean and spread (dispersion/covariance) being the two most important statistical characteristics for their analysis (see [31] and references therein). If the law of data distribution is known in advance, methods based on the Pearson criterion [32] are often used (χ^2 criteria; see, for example, [7, 37]). The outlier identification method based on indicators such as the interquartile range [33, 8, 34, 5, 14, 35, 36] (IQR method) has become widely used in the practice of statistical processing of measurements because of its simplicity. In processing of experimental data by this method, one can filter out erroneous measurements in the tails of distributions. As noted above, erroneous measurements can introduce a significant error into calculation of the statistical characteristics. This paper will present a comparative analysis of the method based on the interquartile range and the method presented in [38], which was used in a number of works [39–42].

*E-mail: ilyushin@itp.nsc.ru

TESTING OF FILTERING METHODS

Despite the widespread use of the IQR method for filtering experimental data, it has a number of disadvantages. Calculation of quartiles necessitates sorting of each of implementations. This requires significant additional memory (if the time series of measurements must be stored). In addition, this method is applicable to implementations close to the normal (Gaussian) distribution. If the distribution has significant skewness $S_f = \langle \mathrm{f}^3 \rangle / \langle \mathrm{f}^2 \rangle^{3/2}$ and/or excess $E_f = \langle \mathrm{f}^4 \rangle / \langle \mathrm{f}^2 \rangle^2 - 3$, the use of the IQR method leaves a significant number of outliers, which leads to errors in determination of the statistical characteristics. Figure 1 shows as an example a histogram of PIV measurements of instantaneous velocity in the flow over a hydrofoil [39] after filtering of outliers by algorithms of the measuring system software (see, for example, [26, 41]). It can be seen that the distribution has significant skewness (it differs a lot from the normal distribution), and outliers are visible on both sides of the distribution core, on its tails. The lines show the cut-off boundaries for these outliers: on the right of $Q3$, the 3rd quartile plus the interquartile range $\Delta = Q3 - Q1$; on the left of $Q1$, the 1st quartile minus the interquartile range Δ. Some outliers remain after discarding of the outliers to the left of the left line and to the right of the right one. This is primarily due to the strong skewness of the distribution. Despite the insignificant number of the remaining outliers, the error in determination of the statistical characteristics with such filtering turns out to be large enough to hamper determination of the actual pattern of the flow (see, for example, [39, Fig. 4]).

In [38], a method was developed for filtering outliers in distributions with strong skewness. This method relies on construction of a model probability density function (PDF) of the measured distribution, which takes into account the strong skewness [43]. This function is a combination of two Gaussian distributions:

$$P(w) = \frac{a^+}{2\pi\sigma_+} \exp\left\{-\frac{(m^+ - w)^2}{2\sigma_+^2}\right\} + \frac{a^-}{2\pi\sigma_-} \exp\left\{-\frac{(m^- - w)^2}{2\sigma_-^2}\right\}, \tag{1}$$

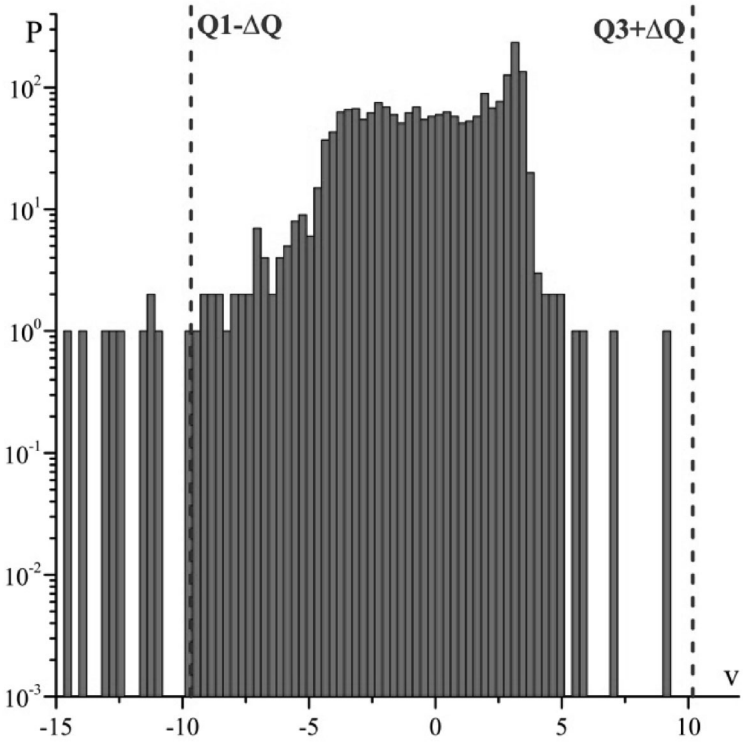

Fig. 1. Histogram of instantaneous velocity measurements in [39]; dotted lines are boundaries of outlier cut-off by IQR method.

where a^+, a^-, m^+, m^-, σ_+, and σ_- are parameters of the model PDF. They are calculated from the known variance σ and skewness coefficient S of the measured distributions [43]:

$$m^+ = \frac{\sigma}{4}\left[S + \sqrt{S^2+8}\right], \quad m^- = \frac{\sigma}{4}\left[S - \sqrt{S^2+8}\right],$$

$$a^+ = -\frac{S - \sqrt{S^2+8}}{2\sqrt{S^2+8}}, \quad a^- = \frac{S - \sqrt{S^2+8}}{2\sqrt{S^2+8}}, \tag{2}$$

$$(\sigma_+)^2 = \frac{\sigma}{16}\left[S + \sqrt{S^2+8}\right], \quad (\sigma_-)^2 = \frac{\sigma}{16}\left[S - \sqrt{S^2+8}\right].$$

The filtering involves rough primary analysis of the measured distribution using the Gaussian PDF. Next, in the case of strong skewness of the filtered distribution, the iterative process of additional filtering is started, based on model PDF (1) constructed from the calculated first three statistical moments of the current distribution ((1) and (2) are written for centered moments). For the purpose of identification and deletion of outliers, each bar of the data histogram is compared with the constructed model PDF. If the bar is higher than the PDF by more than α events (α is the filtering parameter), these events are regarded as outliers and are discarded. The α value depends on the database size: the larger is the database, the smaller α value can be taken (in practice, α ranges from 10 to 1000). The procedure is repeated with the updated model PDF (constructed from the histogram after deletion of outliers) until the filtered histogram exceeds the model PDF with a factor of α. Typically, no more than three such repetitions (iterations) are required for completion of the filtering.

For comparison of the method from [38] with the IQR filtering, with application of a pseudorandom number generator, a test histogram of 1000 implementations was constructed from model PDF (1) with strong skewness: $S = 1.157$ (see Fig. 2). For this purpose, the function $\mathrm{f}(p) = \left[\int_0^{\mathrm{f}} P(w)dw\right]^{-1}$ was calculated with the use of (2). It is the inverse function of the integral of

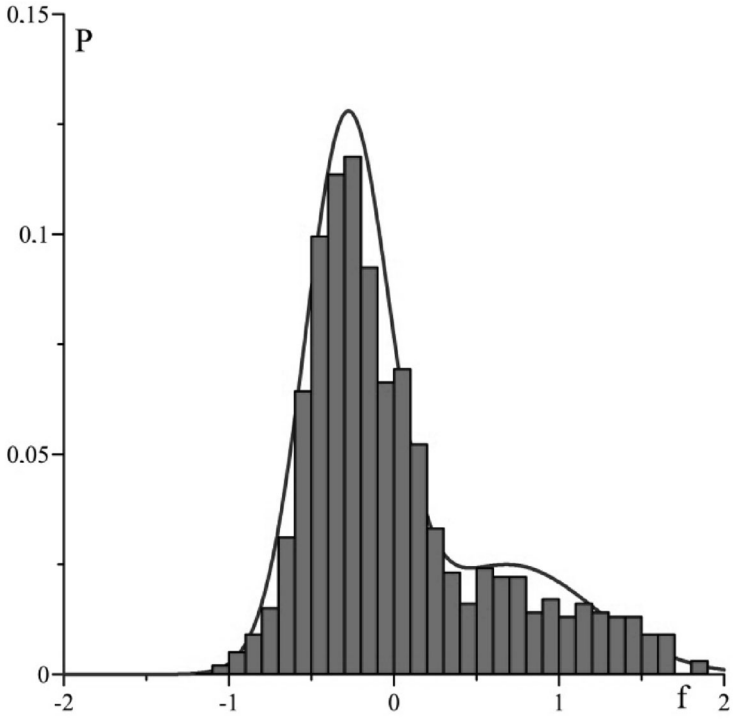

Fig. 2. Test histogram constructed based on model PDF (shown by line).

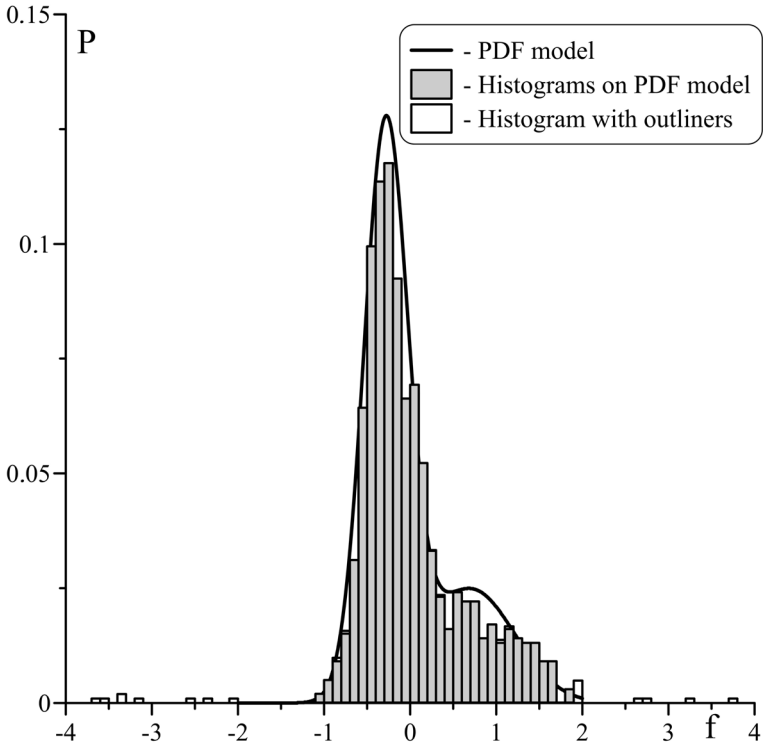

Fig. 3. Test histogram of Fig. 2 with added random outliers.

Table 1.

	Statistical characteristics of histograms		
	$\sigma = \sqrt{\langle \mathrm{f}^2 \rangle}$	$S = \frac{\langle \mathrm{f}^3 \rangle}{\langle \mathrm{f}^2 \rangle^{3/2}}$	$E = \frac{\langle \mathrm{f}^4 \rangle}{\langle \mathrm{f}^2 \rangle^2} - 3$
Original (model) histogram	0.5858	1.157	0.775
Test histogram with outliers	0.672 (**15%**)	0.421 (**63%**)	4.887 (**530%**)
Histogram filtered by method of [38]	0.5870 (**0.2%**)	1.144 (**1%**)	0.74 (**5%**)
Histogram filtered by IQR method	0.6270 (7%)	0.932 (**20%**)	2.507 (**223%**)

PDF (1) with the given parameters σ and S. The argument of this function was a pseudorandom number generator in the range [0; 1]. That resulted in a sequence with distribution close to $P(\mathrm{f})$ (1), which is shown in Fig. 2, along with $P(\mathrm{f})$ (line). Twenty random outliers in the range $[-4; +4]$ were added to this distribution (see the histogram in black in Fig. 3); some of them were found on the tails of the distributions. The test distribution (the histogram with the outliers) was filtered by the method of [38]. The result of the filtering is shown in Fig. 4. The lines in the figure show the boundaries of outlier cut-off by the IQR method. It can be seen that after the IQR filtering, some outliers remain. Despite their insignificant quantity (17 outliers out of 1017 measurements), their influence on the statistical characteristics of the distribution leads to a large error (see Table 1). Note that in Figs. 3 and 4, the original, test (with outliers), and filtered histograms do not coincide not only in the tails, but also in the core. This is due to the fact that the histograms are normalized and centered, and since addition or removal of even a few outlier realizations leads to minor changes in the normalization and the mean, there are seen marked small discrepancies in the core of the histograms.

Table 1 shows the statistical characteristics (standard deviation and skewness and excess coefficients) of the original histogram, histogram with outliers, and histograms filtered by the method

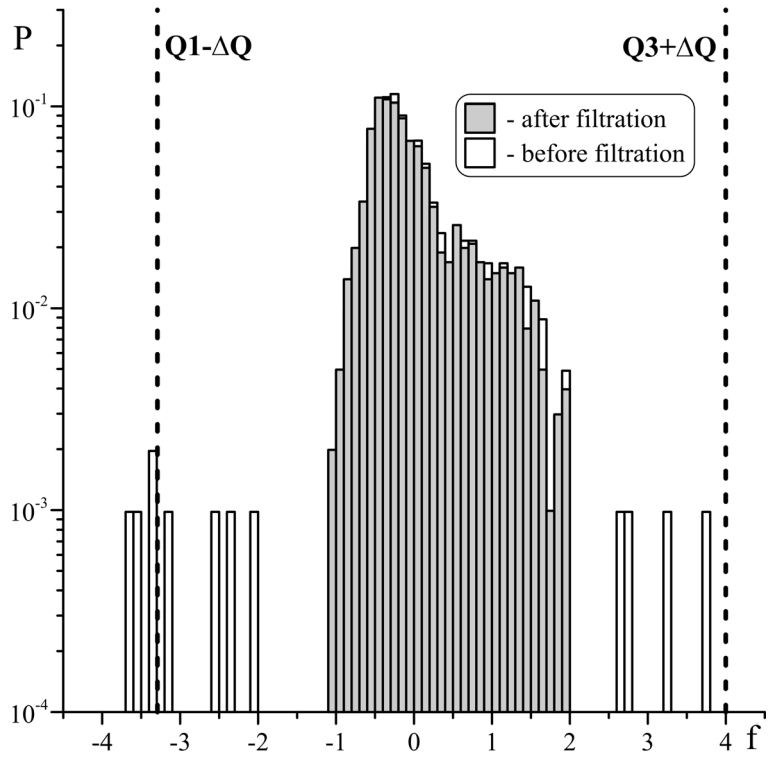

Fig. 4. Histograms in Fig. 3 filtered by IQR method and method from [38].

of [38] and the IQR method. The relative filtering error $\delta = \frac{|f_{ini}-f|}{f_{ini}}$ is shown in the parentheses. The table demonstrates that outliers in measurements can lead to a significant error. Despite their insignificant quantity (20 out of 1020), the relative error in the statistical moments (increasing with the moment order) is as high as 15% for the standard deviation σ to 530% for the excess coefficient E. It is also seen that although the IQR method can approximately halve the error in the calculated statistical characteristics, the error remains unacceptably large, especially for higher statistical moments. The filtering method of [38] makes it possible to filter out outliers with an error not exceeding 5% even for the excess coefficient.

CONCLUSIONS

By the example of a model numerical series, it is shown that random outliers can lead to a significant error in the statistical characteristics of a numerical series. For instance, adding 20 random outliers to a sequence of 1000 numbers leads to an unacceptably large error in the magnitude of the statistical moments. Therefore, such outliers must be identified and filtered out during processing of experimental data [19–22].

The presented results of testing the IQR filtering method on a model distribution show that this method gives a large error for distributions with significant skewness. The same conclusion can be drawn from analysis of the IQR filtering for real PIV measurements [39] of the instantaneous fluid velocity above a hydrofoil, the distribution of which is also highly asymmetric. The method in [38], on the contrary, enables qualitative filtering of experimental data: the error does not exceed 5% for the excess coefficient (the error increases with the moment order). This was shown by the testing results presented in this work, as well as the use of this method for processing measurement data in a number of experiments [38–40, 42].

Note also that when the IQR method is used in software for filtering large databases, it requires additional significant memory space for preliminary sorting of distributions. The method offered in [38] includes the iterative process for each of the distributions, as well as calculation of basic functions. However, the latter are insignificant because the parameters of these functions are obtained analytically and have a simple form, and the iterative process usually does not take more than two or three iterations [38].

FUNDING

The work was carried out with the financial support of the Russian Science Foundation (project no. 19-79-30075-P) and application of the Kutateladze Institute of Thermophysics, Siberian Branch, Russian Academy of Sciences infrastructure.

CONFLICT OF INTEREST

The author of this work declares that he has no conflicts of interest.

REFERENCES

1. Gupta, M., Gao, J., Aggarwal, C.C., and Han, J., Outlier Detection for Temporal Data: A Survey, *IEEE Trans. Knowl. Data Engin.*, 2014, vol. 26, no. 9, pp. 2250–2267; https://doi.org/10.1109/TKDE.2013.184
2. Aggarwal, C.C., An Introduction to Outlier Analysis, in *Outlier Analysis*, New York: Springer, 2013; https://doi.org/10.1007/978-1-4614-6396-2_1
3. Chandola, V., Banerjee, A., and Kumar, V., Anomaly Detection: A Survey, *ACM Comput. Surv.*, 2009, vol. 41, no. 3, pp. 1–58; https://doi.org/10.1145/1541880.1541882
4. Kumar, V., Parallel and Distributed Computing for Cybersecurity, *IEEE Distr. Syst. Online*, 2005, vol. 6, no. 10; https://doi.org/10.1109/MDSO.2005.53
5. Vinutha, H.P., Poornima, B., and Sagar, B.M., Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset, *Inform. Dec. Sci.*, 2018, vol. 701, pp. 511–518; http://dx.doi.org/10.1007/978-981-10-7563-6_53
6. Spence, C., Parra, L., and Sajda, P., Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model, in *Procs. of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, Washington, DC, USA: IEEE Computer Society, 2001; https://doi.org/10.1109/MMBIA.2001.991693
7. Ijaz, M.F., Attique, M., and Son, Y., Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods, *Sensors*, 2020, vol. 20, p. 2809; https://doi.org/10.3390/s20102809
8. Baharuddin, M.Y., Salleh, S.H., Zulkifly, A.H., et al., Design Process of Cementless Femoral Stem Using a Nonlinear Three Dimensional Finite Element Analysis, *BMC Musculoskelet Disord*, 2014, vol. 15, no. 30; https://doi.org/10.1186/1471-2474-15-30
9. Fay, D.S. and Gerow, K., A Biologist's Guide to Statistical Thinking and Analysis, *WormBook*, 2013; https://doi.org/10.1895/wormbook.1.159.1
10. Aleskerov, E., Freisleben, B., and Rao, B., Cardwatch: A Neural Network Based Database Mining System for Credit Card Fraud Detection, in *Procs. of IEEE Computational Intelligence for Financial Engineering*, 1997, pp. 220–226; https://doi.org/10.1109/CIFER.1997.618940
11. Hilal, W., Gadsden, S.A., and Yawney, J., Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances, *Expert Syst. Appl.*, 2022, vol. 193, p. 116429; https://doi.org/10.1016/j.eswa.2021.116429
12. Hodge, V.J. and Austin, J., A Survey of Outlier Detection Methodologies, *Artif. Intell. Rev.*, 2004, vol. 22, pp. 85–126; https://doi.org/10.1007/s10462-004-4304-y
13. Song, Y., Wang, Q., Zhang, X., et al., Interpretable Machine Learning for Maximum Corrosion Depth and Influence Factor Analysis, *npj Mater. Degrad.*, 2023, vol. 7, p. 9; https://doi.org/10.1038/s41529-023-00324-x
14. Jones, P.R., A Note on Detecting Statistical Outliers in Psychophysical Data, *Atten. Percept. Psychophys.*, 2019, vol. 81, no. 5, pp. 1189–1196; https://doi.org/10.3758/s13414-019-01726-3
15. Pimentel, M.A.F., Clifton, D.A., Clifton, L., and Tarassenko, L., A Review of Novelty Detection, *Signal Proc.*, 2014, vol. 99, pp. 215–249; https://doi.org/10.1016/j.sigpro.2013.12.026
16. Munir, M., Siddiqui, S.A., Dengel, A., and Ahmed, S., DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series, *IEEE Access*, 2019, vol. 7, pp. 1991–2005; doi: http://dx.doi.org/10.1109/ACCESS.2018.2886457
17. Domingues, R., Filippone, M., Michiardi, P., and Zouaoui, J., A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses, *Pattern Recogn.*, 2018, vol. 74, pp. 406–421; https://doi.org/10.1016/j.patcog.2017.09.037
18. Gupta, N., Eswaran, D., Shah, N., Akoglu, L., and Faloutsos, C., Beyond Outlier Detection: LookOut for Pictorial Explanation, in *Machine Learning and Knowledge Discovery in Databases*, Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G., Eds., Springer, 2018; https://doi.org/10.1007/978-3-030-10925-7_8
19. Zimmerman, D.W., A Note on the Influence of Outliers on Parametric and Nonparametric Tests, *J. Gen. Psychol.*, 1994, vol. 121, no. 4, pp. 391–401; https://psycnet.apa.org/doi/10.1080/00221309.1994.9921213

20. Zimmerman, D.W., Increasing the Power of Nonparametric Tests by Detecting and Downweighting Outliers, *J. Exp. Educat.*, 1995, vol. 64, no. 1, pp. 71–78; https://api.semanticscholar.org/CorpusID:120621848
21. Zimmerman, D.W., Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions, *J. Exp. Educat.*, 1998, vol. 67, no. 1, pp. 55–68; https://psycnet.apa.org/doi/10.1080/00220979809598344
22. Mowbray, F.I., Fox-Wasylyshyn, S.M., and El-Masri, M.M., Univariate Outliers: A Conceptual Overview for the Nurse Researcher, *Can. J. Nurs. Res.*, 2019, vol. 51, no. 1, pp. 31–37; https://doi.org/10.1177/0844562118786647
23. Peirce, B.O., Criterion for the Rejection of Doubtful Observations, *Astron. J.*, 1852, vol. 2, pp. 161–163; https://doi.org/10.1086/100259
24. Grubbs, F.E., Procedures for Detecting Outlying Observations in Samples, *Technometrics*, 1969, vol. 11, pp. 1–21; https://doi.org/10.2307/1266761
25. García, S., Luengo, J., and Herrera, F., *Data Preprocessing in Data Mining (Intelligent Systems Reference Library)*, 2015; http://dx.doi.org/10.1007/978-3-319-10247-4
26. Raffel, M., Willert, C.E., Wereley, S.T., and Kompenhans, J., *Particle Image Velocimetry: A Practical Guide*, 2nd ed., Berlin: Springer, 2007; https://doi.org/10.1007/978-3-540-72308-0
27. Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., and Walczak, B., Robust Statistics in Data Analysis—A Review. Basic Concepts. *Chemometrics Intelligent Lab. Syst.*, 2007, vol. 85, pp. 203–219; http://dx.doi.org/10.1016/j.chemolab.2006.06.016
28. Chandola, V., Banerjee, A., and Kumar, V., Anomaly Detection: A Survey, *ACM Comput. Surv.*, 2009, vol. 41, no. 3, pp. 1–58; https://doi.org/10.1145/1541880.1541882
29. Cousineau, D. and Sylvain C., Outliers Detection and Treatment: A Review, *Int. J. Psychol. Res.*, 2010, vol. 3, pp. 58–67; http://dx.doi.org/10.21500/20112084.844
30. Zimek, A. and Filzmoser, P., There and Back Again: Outlier Detection between Statistical Reasoning and Data Mining Algorithms, *Wiley Interdiscip. Rev.: Data Mining Knowledge Discovery*, 2018, vol. 8, no. 6; https://doi.org/10.1002/widm.1280
31. Rousseeuw, P.J. and Leroy, A.M., *Robust Regression and Outlier Detection*, New York: Wiley Interscience, 1987; http://dx.doi.org/10.1002/0471725382
32. Pearson, K., X. On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling, *London, Edinburgh, Dublin Philos. Mag.*, 1900, vol. 50, no. 302, pp. 157–175; https://doi.org/10.1080/14786440009463897
33. Beyer, H. and Tukey, J.W., Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass.—Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney, 1977, XVI, *Biometrical J.*, 1981, vol. 34, no. 4, pp. 413/414; https://doi.org/10.1002/bimj.4710230408
34. Chockalingam, S., Aluru, M., and Aluru, S., Microarray Data Processing Techniques for Genome-Scale Network Inference from Large Public Repositories, *Microarrays*, 2016, vol. 5, no. 3, p. 23; https://doi.org/10.3390/microarrays5030023
35. Rajendran, L.K, Bhattacharya, S., Bane, S.P.M., and Vlachos, P.P., Meta-Uncertainty for Particle Image Velocimetry, *Meas. Sci. Technol.*, 2021, vol. 32, p. 104002; http://dx.doi.org/10.1088/1361-6501/abf44f
36. Grossmann, F., Flueck, J.L., Roelands, B., Meeusen, R., Mason, B., and Perret, C., Characteristics of Official Wheelchair Basketball Games in Hot and Temperate Conditions, *Int. J. Environ. Res. Public Health*, 2022, vol. 19, no. 3, p. 1250; https://doi.org/10.3390/ijerph19031250
37. Pervunin, K.S., Timoshevskiy, M.V., and Ilyushin, B.B., Distribution of Probability of the Vapor Phase Occurrence in a Cavitating Flow Based on the Concentration of PIV Tracers in Liquid, *Exp. Fluids*, 2021, vol. 62, p. 247; https://doi.org/10.1007/s00348-021-03344-y
38. Heinz, O., Ilyushin, B., and Markovich, D., Application of a PDF Method for the Statistical Processing of Experimental Data, *Int. J. Heat Fluid Flow*, 2004, vol. 25, no. 5, pp. 864–874; https://doi.org/10.1016/j.ijheatfluidflow.2004.05.009
39. Ilyushin, B.B., Timoshevskiy, M.V., and Pervunin, K.S., Vapor Concentration and Bimodal Distributions of Turbulent Fluctuations in Cavitating Flow around a Hydrofoil, *Int. J. Heat Fluid Flow*, 2023, vol. 103, p. 109197; https://doi.org/10.1016/j.ijheatfluidflow.2023.109197
40. Alekseenko, S.V., Bilsky, A.V., Dulin, V.M., and Markovich, D.M., Experimental Study of an Impinging Jet with Different Swirl Rates, *Int. J. Heat Fluid Flow*, 2007, vol. 28, no. 6, pp. 1340–1359; https://doi.org/10.1016/j.ijheatfluidflow.2007.05.011

41. Tokarev, M.P., Markovich, D.M., and Bil'sky, A.V., Adaptive Algorithms for Processing Particle Images for Calculating Instantaneous Velocity Fields, *Vychisl. Technol.*, 2007, vol. 12, no. 3, pp. 109–131.
42. Severin, M.V., Timoshevskii, M.V., Ilyushin, B.B., and Pervunin, K.S., Turbulent Structure of a Free Bubble Jet: Analysis of the Higher Statistical Moments of Velocity Fluctuations, *PMTF*, 2023, no. 6, pp. 81–84; DOI: 10.15372/PMTF202315302
43. Ilyushin, B.B., Use of Higher Moments to Construct PDF's in Stratified Flows, in *Closure Strategies for Turbulent and Transitional Flows*, Launder, B.E. and Sandham, N., Eds., Cambridge University Press, 2001, pp. 683–699; https://doi.org/10.1017/CBO9780511755385