Extended Abstract for the
Symposium on Integrating AI and Data Science into School Education Across Disciplines 2025

DATA CLEANING IN MATHEMATICS EDUCATION: TEACHING STATISTICAL METHODS OF OUTLIER DETECTION

Jakim Eckert[1], Sarah Schönbrodt[2] & Martin Frank[3]
[1]Karlsruhe Institute of Technology, Germany, jakim.eckert@kit.edu
[2]Paris Lodron University Salzburg, Austria
[3]Karlsruhe Institute of Technology, Germany

Focus Topics: Data and Problems, Learning Materials

Why Data Science Education?

We live in a world increasingly shaped by data and algorithms. Everyday decisions are influenced by data science processes without one noticing it directly (Grzymek & Puntschuh, 2019). Consequently, it is essential to educate future generations in data science and to promote data literacy. Students should be taught to engage with data and data-driven algorithms in a reflective and informed manner from an early age (Schüller et al., 2021). This provides the motivation to develop teaching and learning material as part of a design research project with focusing on mathematical aspects of data processing. One of the first steps in data processes is data cleaning and data preprocessing. Data cleaning plays an important role in data science procedures and can have direct impacts on insights, predictions and decisions based on data science procedures.

Relevance of Data Cleaning

Anaconda's data science report (2022) states that data scientists spend almost 38% of their time on data preparation and data cleaning. Typical meta-models for data science processes such as PPDAC also emphasize data cleaning (or data preparation) as a key step (Wild & Pfannkuch, 1999). Main tasks of data cleaning include outlier detection, handling missing data, and data augmentation. Decisions made during data cleaning can have a crucial impact on the performance and results of subsequently applied algorithms or data analysis methods. The performance of many AI applications also heavily depends on the quality of data, as a large part of these applications is data-driven (Ntoutsi et al. 2020).

Within the Design Based Research (DBR) Project presented here we focus on outlier detection as one subtask of data cleaning. On the one hand, the existence of outliers in data sets can have a major impact on the results of models developed on the basis of these data sets. This is demonstrated in Fig. 1 using a linear regression model. On the other hand, data can also be incorrectly classified as outliers and removed from a data set even though it entails relevant information. A popular example is the discovery of the hole in the ozone layer. Data was labeled as outliers and removed from the data set due to an excessive restriction to so far typical ozone values. Due to incorrect data cleaning, the ozone hole was ‘discovered’ over 20 years later (Pötter, 2016).

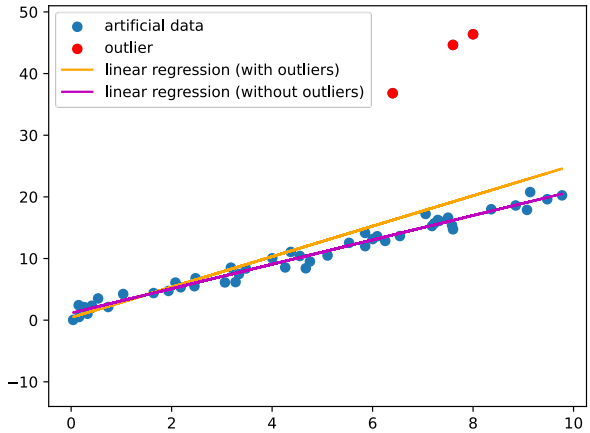

Figure 1: Regression with and without outliers

In S. Podworny & S. Schönbrodt (Eds.), *Towards Fostering AI and Data Science Literacy in Schools Across Disciplines. Proceedings of the 1st Symposium on Integrating AI and Data Science into School Education Across Disciplines (AIDEA1 2025)*, Salzburg, Austria.

A Design Research Project on Outlier Detection

The presented examples and explanations highlight the significance of outlier detection for the outcomes of data-based prediction models. At the same time, data cleaning methods are mathematically rich and therefore potentially of interest for mathematics education. In addition, data cleaning and outlier detection are good starting points for activities to promote data literacy, such as addressing ethical issues. Should outliers be removed automatically and what are the consequences? How can mathematical assumptions, such as the normal distribution of the sample, affect the identification of outliers? The design research project presented here investigates which and how the mathematical foundations of data cleaning and outlier detection methods can be made accessible to upper secondary students, thereby also motivating competencies in the area of data literacy.

Data science education and data literacy have gained increasing importance in educational research in recent years, especially in the fields of mathematics and computer science education (Fleischer et al., 2022; Markulin et al., 2022; Schönbrodt et al., 2022; Schüller et al., 2021). Earlier educational studies include aspects of data cleaning as a necessary but relatively minor part of data science projects (Fleischer et al., 2022; Markulin et al., 2022). In this study, we focus on data cleaning, especially on the subtopic of outlier detection, as a primary subject and the teaching and learning of its mathematical foundations.

Within the DBR project the following research interests (RI) will be addressed. These interests are aligned with the 4-level approach for specifying and structuring mathematical content from Hußmann & Prediger (2016):

RI 1: Which methods, concepts and key mathematical principles underlie outlier detection and which of these principals are appropriate to be specified, structured, and elementarized for secondary mathematics education? (Formal and semantic level)

RI 2: How can the influence of outliers and their removal on data science processes be taught? Which data, contexts and digital tools are suitable? Which educational approaches are suitable and enable a high level of personal activity? (Concrete level)

To investigate these interests, we conduct design research and develop teaching and learning materials for use with upper secondary students.

Our work on RI 1 includes specifying and structuring the mathematical foundations of data cleaning methods and highlighting possible elementarizations for upper secondary education. The elementarizations of relevant statistical methods draw on and partly go beyond high school statistical knowledge. The considered methods for outlier detection range from approaches based on extreme value considerations to distance and density-based methods, linear models to statistical tests (Aggarwal, 2017).

In the following, we look at extreme value analyses as an example to show where school mathematics can be used as a starting point and where there is potential to go beyond. School knowledge from analysis gives a first idea of detecting extreme values. At the same time, box plots can be used for a simple one-dimensional discrete approach which can classify whether extreme values can be labelled as outliers. Going beyond school knowledge, depth- or angle-based methods and the Mahalanobis distance, for example, can be used for analyzing multivariate discrete data (Aggarwal, 2017).

For the DBR project we decided to focus on boxplots, standard deviation, the Mahalanobis distance as well as statistical tests such as the Dixon or Grubbs test in order to start with methods closely related to school mathematics on the one hand and to show a certain richness of methods at the other hand. Another aspect is looking at contextual reasons to explain outliers after their identification. It should also become clear that data cleaning is not an end in itself and that it serves, for example, better predictions and increased accuracy. It should be considered what the prediction would look like with or without outliers to show the purpose of data cleaning.

Design decisions

Following the decision to prioritize the learning material about outlier detection based on boxplots, standard deviation, the Mahalanobis distance, and statistical tests we identified upper secondary students as the target audience. Looking at RI 2 and thus at educational approaches, data and digital tools to teach data cleaning we made some design decisions.

(D1) Using real data: First, we decided to work with real data sets to provide an authentic view on data cleaning. For handling real data sets, we choose to use Jupyter Notebooks and Python to create interactive digital worksheets.

(D2) Combine open modeling activities with direct teaching: In addition, we decided to combine open modeling activities with the direct teaching of mathematical foundations. More precisely, we draw on students' prior knowledge in an open start, to encourage them to be creative and come up with their own approaches to detect outliers in a given data set. The students explore the data set on their own. Once they have a basic understanding of the data set, they are asked to give a first rough definition of outliers. Their definitions of what typical candidates for outliers might be and their approaches how to detect these outliers are discussed.

(D3) Connecting students' ideas to detection methods: Afterwards their approaches are applied to real data sets and compared with various outlier detection methods. After categorizing and discussing their ideas, the students have the opportunity to choose between different methods that cover different mathematical levels.

(D4) Showing the influence of outliers on predictive models: As mentioned above, the influence of potential outliers is also emphasized with the help of regression predictions with and without outliers. Different identification methods can be compared with each other in terms of their predictions.

In empirical studies we will collect the students' answers and analyze them with the help of a qualitative content analysis. Consideration of prior knowledge in relation to outlier identification and categorization of students' mathematical ideas in comparison to scientifically used methods of outlier detection.

At the symposium, we will present the developed learning materials, educational concepts, and initial results from tests conducted with secondary students during the first design cycle. Outlier detection is just one example of data cleaning, which is intended to demonstrate its general relevance. Further research will focus on other subtopics of data cleaning such as the handling of missing data.

References

Aggarwal, C. C. (2017). Data Mining: The Textbook. https://doi.org/10.1007/978-3-319-14142-8

Anaconda (2022). State of Data Science 2022: Paving the Way for Innovation. https://www.anaconda.com/state-of- data-science-report-2022.

Bata (submitted). *Maschinelles Lernen lernen - Entwicklung und Erforschung einer Lehr-Lernumgebung in den Ingenieurswissenschaften.*

Fleischer, Y., Biehler, R. & Schulte, C. (2022). Teaching and Learning Data-Driven Machine Learning with Educationally Designed Jupyter Notebooks. *Statistics Education Journal, 21*(2), article 7.

Grzymek, V. & Puntschuh, M. (2019). *Was Europa über Algorithmen weiß und denkt – Ergebnisse einer repräsentativen Bevölkerungsumfrage.* Bertelsmann-Stiftung. https://doi.org/10.11586/2019006

Hußmann, S. & Prediger, S. (2016). Specifying and Structuring Mathematical Topics. *Journal für Mathematik-Didaktik, 37*(S1), 33–67. https://doi.org/10.1007/s13138-016-0102-8

Markulin, K., Bosch, M., Florensa, I. & Montañola, C. (2022). The evolution of a study and research path in Statistics. *epiDEMES, 1*. https://doi.org/10.46298/epidemes-7584

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., . . . Staab, S. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. *Wiley Interdisciplinary Reviews Data Mining And Knowledge Discovery*, *10*(3). https://doi.org/10.15488/10778

Pötter, B. (2016, 1. April). In letzter Minute: Rettung der Ozonschicht. ZEIT ONLINE. https://www.zeit.de/2007/37/A-Montreal-Protokoll (Original work published 2007).

Schönbrodt, S., Wohak, K. & Frank, M. (2022): Digital Tools to Enable Collaborative Mathematical Modeling Online. *Modelling in Science Education and Learning, 15*(1), 151–174, https://doi.org/10.4995/msel.2022.16269.

Schüller, K., Koch, H. & Rampelt, F. (2021). *Data-Literacy-Charta.* https://www.stifterverband.org/sites/default/files/data-literacy-charter.pdf

Wild, C. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review, 67*(3), 223-265.