Challenges and Prospects in Anomaly Detection of Sewer Monitoring Data: Annotating Synthetic Sewer Data with Known Sensor Failures

Jörg Rieckermann*, Andy Disch

Department of Urban Water Management, Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

Correspondence to: Jörg Rieckermann (joerg.rieckermann@eawag.ch)

Abstract

Managing sewer monitoring data poses challenges, often revealing quality issues. This study explores the feasibility of employing crowdsourcing methods, akin to artificial artificial intelligence, on sewer flow monitoring data. We devised a data annotation project, relying on the expertise of 12 sewer researchers to retrieve artificial anomalies introduced to the data. While recognizing limitations in annotation skills, we compiled time series from 7 locations and implemented visualizations for improved interpretation. Evaluation using the F1 criterion yielded mediocre scores (0.625±0.226), highlighting challenges in interpreting noisy raw data and varying analyst mental models. Despite the potential for data-driven modeling in urban drainage research, our results suggest challenges in obtaining annotations through crowdsourcing. Further work should focus on standards for data annotations, community efforts in data labelling, and identifying role model utilities providing open access to routine wastewater datasets.

1 Introduction

Advances in sensor technology and data transmission enable cost-effective and flexible process monitoring in drainage systems (Ruggraber et al., 2023; Kerkez et al., 2016; Blumensaat et al., 2017). Power supply and battery limitations in the past can be overcome with autonomous sensors utilizing energy harvesting technology (Mathis et al., 2022; Blumensaat et al., 2017), resulting in ubiquitous monitoring (Blumensaat et al., 2023). However, it is well known that quality control of monitoring stations is challenging (Brito et al., 2021) and data streams from a plethora of sensors requires automated approaches to efficiently validate real-time measurement data and identify anomalies or dubious data points (DWA-M 181, Becouze-Lareure et al., 2012; Mourad and Bertrand-Krajewski, 2002).

However, challenges arise when inspecting multiple signal types in real-time and employing computationally intensive analysis methods. Machine learning (ML) methods have been suggested as promising solutions (Aggarwal, 2017;

Branisavljević et al., 2011). Unfortunately, despite the buzz and hype around machine learning, they are not a universal solution. Two recent studies investigated advanced ML methods to effectively detect anomalies in sewer sensor data and concluded that i) much simpler methods performed equally or better than advanced methods and ii) that the degree of curating and preprocessing that data, including expert knowledge from engineers was important for a good performance (Trinnex, 2023). This supports the findings of (Russo et al., 2020) who concluded that, especially in data-driven modelling in the environmental field, access to expert-based data annotation is critical.

Unlike in domains with abundant everyday-type of data, such as images of cats, dogs and trees, etc., which can be annotated by non-experts, annotating wastewater monitoring data faces four specific challenges. First, open experimental datasets from sewers are largely lacking (Nedergaard Pedersen et al., 2021; Blumensaat et al., 2023) and, more importantly, there are only very few experts on sewer monitoring data which would be capable to annotate monitoring data. Third, wastewater professionals know that it is challenging to differentiate between process anomalies due to measurement malfunctions and real effects due to the stochastic nature of wastewater production, e.g. substantial industrial discharges that occur irregularly. Fourth, our flow meters and water level sensors are comparably imprecise and more often than not affected by factors like humidity, temperature changes, fouling and shifts in setpoints. This often leads to noisy signals which makes anomaly detection more challenging, especially for crowdsourcing methods.

Errors in annotated datasets for training and testing are difficult to avoid completely and the errors in crowdsourcing annotations can vary depending on several factors, such as the quality of the crowd workers, the clarity of the annotation task instructions, and the level of agreement among the crowd workers themselves. However, with proper quality control measures such as multiple annotators and consensus-based approaches, crowdsourcing annotations can achieve a reasonable level of accuracy comparable to or even surpassing that of expert annotations (Wang et al., 2020; Tang and Lease, 2011). The accuracy can also be improved, especially for popular datasets, through iterative refinement processes and the use of statistical techniques to identify and mitigate potential biases or errors in the crowd-contributed annotations.

Researchers examined popular open source datasets from the past two decades in computer vision, natural language processing, sentiment analysis, and audio processing (Northcutt et al., 2021). They found labelling errors of 3.4% on average, ranging from 0.15 to 10.12% for the individual datasets. They also conclude that even these small mislabellings make benchmark results from the test sets unstable and the popular “gorilla” incident drastically demonstrated how biases from the datasets can manifest in ML systems and have unintended consequences (Charte et al., 2019). For sewer datasets, given the problems mentioned above, we would probably expect even higher error rates. Unfortunately, studies on this topic are completely lacking and, to the best of our knowledge, errors we could expect in annotated sewer datasets are unknown.

In this paper we therefore investigate for the first time the applicability of crowdsourcing methods for annotating sewer water levels and flows. Specifically, we 12 experts manually identified and annotated synthetic anomalies in 7 time series of wastewater flows. We used a special annotation tool, which was capable of displaying multiple times series and visualizing upstream-downstream comparisons of multiple signals (Fig. 2). Although the study has a lot of limitations, we learned several key lessons, which will be very useful for future trials to annotate wastewater flow data.

2. Methods

2.1 Selection of experts on UWO wastewater flow data

We invited 12 experts, with different level of experience with flow monitoring data to incorporate domain expertise by consulting wastewater management professionals or subject matter experts. Their knowledge can help define specific anomaly scenarios and guide the creation of synthetic data that accurately represents real-world wastewater flow behavior.

The researchers involved in the project possessed significant knowledge and experience in sewer research, enabling them to accurately identify frozen water level sensors through flat time series and detect abnormal fluctuations in the data. The aim was to achieve high-quality annotations by leveraging their collective expertise.

2.2 Modelling anomalies in wastewater flow data

To generate synthetic flow data containing anomalies, we followed a five-step process. In the first step, we determined the characteristics and properties of the synthetic wastewater flow data that we intended to create. This involved identifying factors such as flow rates, temporal resolution, patterns, and specific anomalies to be simulated. The second step involved generating the base wastewater flow data. This was achieved by employing statistical models, mathematical equations, or simulations that replicated real-world wastewater flow patterns. Factors such as diurnal variations, seasonal variations, and other relevant patterns observed in real data were considered. A total of seven time series were compiled for the subsequent annotation task. In the third step, synthetic anomalies were introduced into the base wastewater flow data. We defined the types of anomalies we wanted to simulate within the synthetic data. These anomalies could include sudden spikes, drops, irregular patterns, or any other deviations from normal flow behavior. We applied appropriate algorithms or methods to introduce these anomalies into the base wastewater flow data.

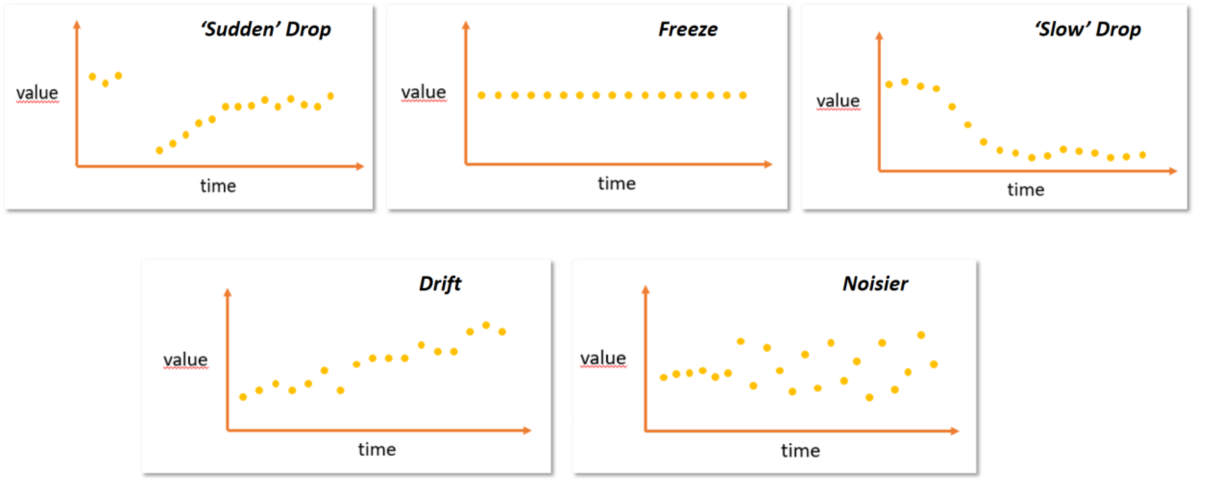

Figure 1: Overview of the types of virtual anomalies that were introduced on to the time series.

Table 1 presents an overview of the characteristics of each anomaly found in the synthetic flow data, including parameters such as magnitude, duration, and frequency. These parameters play a crucial role in determining the severity and occurrence of the synthetic anomalies within the generated data.

To ensure the usefulness and reliability of the created and labelled dataset, visible quality assurance tests were conducted. These tests aimed to assess the dataset's quality and appropriateness based on summary statistics. Various statistical measures and metrics were employed to evaluate the datasets properties, such as mean, median, standard deviation, range, and other relevant statistical summaries.

Table 1: Virtual anomalies and their characteristics and parameterization.

Type	Parameters	Value	Uncertainty/Distribution
Sudden Drop	drop	50%	10-190; U(10,190)
Freeze	duration	60h	30-80; N(60, 20)
	stop	0	
Slow Drop			
Drift	steepness	10%	
	duration		N(50, 100)
Noisier			N(0.1, 0.5)
	sign	+, -	

2.3 A custom annotation tool

After evaluating existing annotation tools, it was determined that none of them adequately met our requirements. As a result, we developed our own annotation tool based on the ADASen framework (Russo et al. 2020). This custom tool was specifically designed to accommodate the unique characteristics of our dataset, which involved multiple time series and the need to visualize sewer-specific information. The developed annotation tool incorporated features that allowed for efficient annotation and analysis of the data. It provided the capability to handle multiple time series simultaneously, enabling labellers to annotate anomalies across different variables or parameters. Additionally, the tool included specialized functionalities that facilitated the plotting and visualization of sewer-specific information, enhancing the labelling process by providing context and relevant domain-specific insights.

By creating a custom annotation tool tailored to our specific needs, we aimed to optimize the annotation workflow and improve the accuracy and efficiency of the labellers. This tool not only addressed the limitations of existing annotation tools but also provided enhanced functionality to effectively handle the complexities and nuances of our dataset, ensuring the quality and relevance of the resulting annotations.

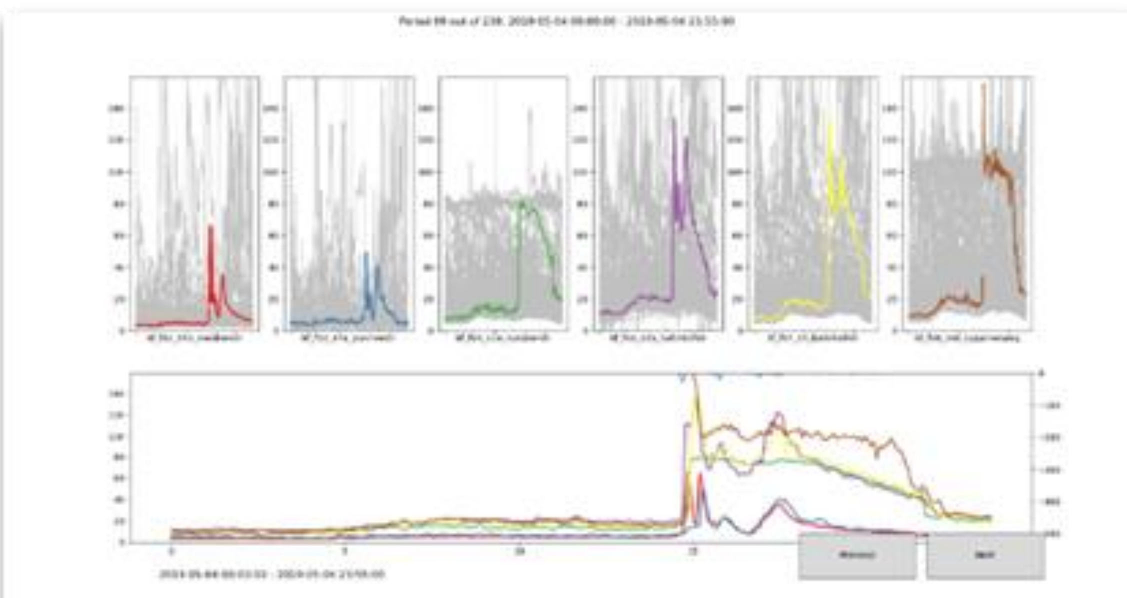

Figure 2 Screenshot of the annotation tool.

To facilitate the annotation process and enable seamless data sharing among team members, the scripts used in datValX pipline (Disch et al. 2023) were modified, and custom Python scripts were developed. These scripts allowed researchers to input their annotated data into a centralized database.

2.4 Annotation procedure

The selected group of labellers was briefed about the experimental procedure, including instructions on how to use the labelling tool effectively. It should be noted that the labels generated during the first session were not used for the subsequent analysis, which led to minor modifications in the tool. Labellers were given an opportunity to become familiar with the labelling tool in a practice session. Their labels from this session were not included in the final analysis but aimed to ensure their comfort with the tool and its functionalities. Labellers were provided with information about the dataset, including a map of the catchment area illustrating major flow paths and upstream-downstream relations. This briefing aimed to enhance their understanding of the dataset and the contextual factors that could influence anomaly identification. Following the briefing and tool familiarization, the labelling environment was set up for the labellers. Each labeller had access to the dataset and the labelling tool, enabling them to work independently.

Labellers individually reviewed the dataset and marked instances they deemed to be anomalies using the provided labelling tool. They relied on their own judgment and expertise to identify and annotate anomalies based on the information presented.

After the labelling process, a de-briefing meeting was conducted to analyze and interpret the collected data. The identified anomalies from each labeller were discussed, providing an opportunity for clarification, consensus building, and addressing any uncertainties or discrepancies.

2.5 Performance evaluation

To evaluate the quality of the annotations, the F1 criterion, commonly employed in previous studies, was utilized. The F1 criterion combines precision (hit rate) and recall (accuracy), providing a comprehensive assessment of the analyst's performance in terms of annotation quality (Disch et al., in prep.).

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

Precision is the ratio of correctly identified positive instances (true positives) to the total number of instances identified as positive (true positives + false positives). It quantifies the accuracy of positive predictions.

Recall, also known as sensitivity or true positive rate, is the ratio of correctly identified positive instances (true positives) to the total number of actual positive instances (true positives + false negatives). It measures the ability to capture all relevant positive instances.

The F1 score combines precision and recall into a single value, providing a harmonic mean that accounts for both measures. It ranges from 0 to 1, with 1 indicating perfect precision and recall, while 0 represents the worst performance.

2.6 Application to a model

The differently labelled datasets obtained from the annotation process were utilized to train a simple random forest model. This step aimed to investigate the impact of variations in the training dataset on the performance of the model. The labelled dataset, comprising instances with annotated anomalies, was used as the training dataset for the random forest model. The model was trained using this dataset to learn the patterns and characteristics of the labelled anomalies.

To evaluate the performance and generalization ability of the trained model, the synthetic dataset, which contained artificially generated flow data with known anomalies, was used as the test dataset. The purpose of using the synthetic dataset as the test dataset was to assess how well the model could re-detect the synthetic anomalies.

In this project, a Random Forest (RF) model was employed to analyze the training dataset with the goal of predicting labels for the given data. RF, an ensemble machine learning algorithm, was used to create a collection of decision trees during the training phase. Each decision tree was constructed using random subsets of both the data and features, and their individual predictions were aggregated to form a more accurate and robust prediction for the target labels. This approach was chosen for its versatility, ability to handle various types of data, and resistance to overfitting. Following the training process, the model's performance was likely evaluated using appropriate metrics on a separate test dataset to assess its predictive accuracy and generalization capabilities.

3 Material: Wastewater flow data from the Urban Water Observatory

This case study highlights data obtained from the Urban Water Observatory, an ongoing initiative aimed at monitoring in-sewer processes in the municipality of Fehraltorf, situated near Zurich in Switzerland. More information about the Urban Water Observatory can be found at the website www.eawag.ch/uwo. The study specifically focuses on analyzing in-sewer flow rates, which were recorded over a period of 2 years with a temporal resolution of 5 minutes.

Although the identification of anomalies in the data relies on visual inspection, the study found that these anomalies exhibit distinct characteristics and are easily distinguishable from normal data points, even for individuals without expertise in the field. This implies that the anomalies possess clear and distinguishable patterns within the observed domains. For researchers seeking access to the complete dataset, it is recommended to refer to the ERIC/open platform (ERIC-open, 2023). The full dataset, including all relevant information, is made available under the UWO project[1] and the reproducibility platform RENKUlab[2].

4 Results

Figure 5 showcases the labelling effort of an example labeller. The labels assigned by the labeller are depicted in red, while the synthetic anomalies are highlighted in green. The visualization provides a clear representation of the annotated anomalies and their locations within the dataset. This visualization aids the alignment and a visual assessment of the labeller accuracy in identifying and marking anomalies within the dataset.

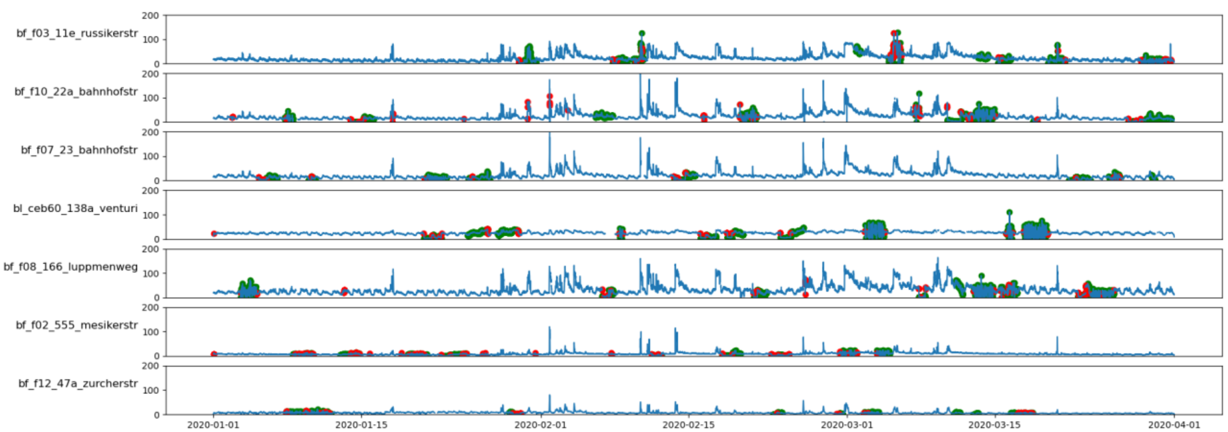

Figure 3 Example of data in blue, the synthetic labels in green and the labelling of an expert in red.

Table 6 presents the results of the conducted experiment. The median F1 score, a metric measuring the model's overall performance in terms of precision and recall, was found to be 0.625. The lowest F1 score observed was 0.07, indicating instances where the model's performance was relatively poor.

[1] https://doi.org/10.25678/000C5K
[2] https://gitlab.renkulab.io/andy.disch/datvalx-annotation-challenge

Table 2 Virtual anomalies and their characteristics and parameterization.

Expert	F1 score	RF score
1	0.08	0.07
2	0.74	0.1
3	0.67	0.1
4	0.66	0.08
5	0.65	0.09
6	0.62	0.09
7	0.56	0.09
8	0.48	0.08
9	0.63	0.09
10	0.38	0.09
11	0.63	0.08
12	0.07	0.07
Median	0.625±0.226	0.09±0.01

Furthermore, the results in the last column in Table 2 indicate that the application of the RF model did not significantly improve its performance, regardless of the quality of the labelling. This suggests that the model's performance was not significantly influenced by the specific labelling approach or the quality of the labelled dataset. These findings suggest that alternative approaches or modifications to the model may be necessary to achieve notable improvements in performance.

During the debriefing session, multiple differences in the application of the annotation tool were discussed among the labellers. These differences could include variations in annotation techniques, interpretation of anomalies, or discrepancies in labelling specific instances. One illustrative aspect highlighting these differences is presented in Figure 5.

However, it was noted during the discussion that the absence of a comprehensive catalogue or guideline hindered the objective assessment of these divergent issues. Without a standardized framework or reference, it becomes challenging to objectively evaluate and compare the variations in the application of the annotation tool. To address this limitation, there is a clear need for additional guidance and a well-defined catalogue that outlines specific annotation guidelines, best practices, and expected interpretations of anomalies. Providing such guidance would facilitate a more consistent and reliable application of the annotation tool among labellers, enabling better comparison and evaluation of the labelled data.

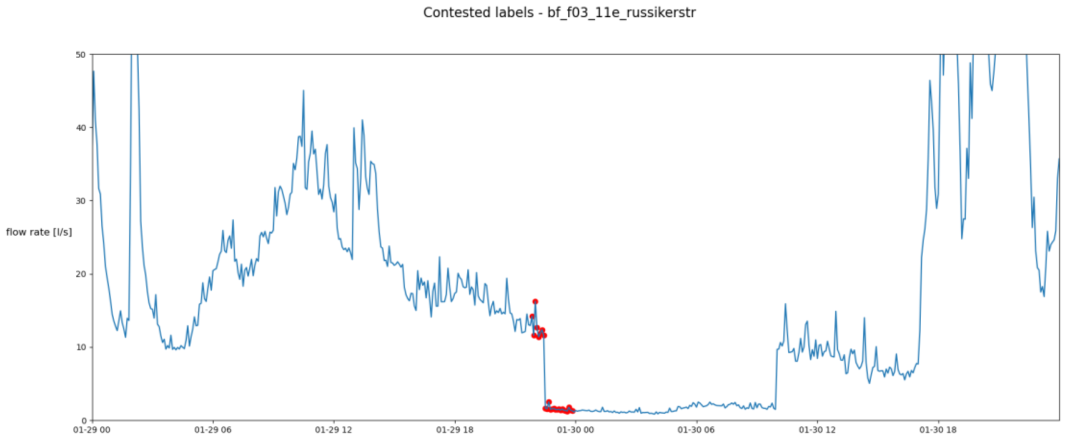

Figure 4 Example for imprecise start and end definition of anomalies.

5 Discussion

In this section, we discuss the various challenges we encountered when labelling a consistent dataset for urban drainage data. Firstly, the challenges related to tool handling were apparent. The choice and functionality of the labelling tool played a crucial role in ensuring consistent and accurate labelling. In some instances, the tools used lacked certain features or had limitations that hindered the labelling process. Addressing these issues by refining existing tools or developing new ones specifically tailored to the dataset's requirements would greatly enhance the efficiency and effectiveness of the labelling process.

Another significant difficulty arose from the precise definition of anomalies. Labellers faced ambiguity regarding when to start and label zones as anomalous. This uncertainty stemmed from the absence of clear guidelines on whether anomalies should be labelled a couple of points before the first abnormal point or precisely at that point. Additionally, labellers exhibited variations in their understanding of what constitutes an anomaly. To mitigate these challenges, providing labellers with more comprehensive and detailed briefings that emphasize the start and end points of anomalies would be beneficial in achieving more consistent labelling results.

Fatigue also emerged as a notable obstacle during the labelling process. The daytime and duration of labelling sessions influenced labellers' ability to maintain a high level of accuracy and consistency. As fatigue set in, labellers' concentration and attention to detail might have waned, leading to potential inconsistencies in labelling. To address this issue, implementing appropriate breaks, managing workload distribution, and ensuring labellers have sufficient rest periods would help mitigate the impact of fatigue and maintain the quality and consistency of the labelled dataset.

Lastly, a key challenge was the need for prior knowledge about the expected behaviour of the sensors in the system. Understanding how the sensors should behave under normal operating conditions is crucial for identifying anomalies accurately. Lack of comprehensive information and guidance regarding sensor behaviour can introduce biases and inconsistencies in the labelling process. Providing labellers with detailed information on the expected sensor behaviour, along with examples of normal and abnormal patterns, would improve their understanding and enhance the consistency of the labelling process.

6 Conclusions

In conclusion, the production of a useful labelled data set for urban drainage data presents significant challenges that need to be addressed. Firstly, the inherent noise within urban drainage data poses a major obstacle. The complex and dynamic nature of urban environments introduces various sources of interference and inaccuracies, making it difficult to obtain clean and reliable data. Efforts should be directed towards developing robust data collection and pre-processing techniques to mitigate the impact of noise and enhance the quality of the labelled data set.

Another crucial challenge stems from the differences in mental models among stakeholders involved in urban drainage systems. These differences encompass variations in system understanding and anomaly definition, which can lead to inconsistencies in labelling data. Overcoming this challenge requires fostering effective communication and collaboration between experts from different domains. Establishing a common understanding and consensus on key concepts and definitions will help improve the accuracy and consistency of the labelled data set.

Furthermore, the limitations of available tools for data labelling contribute to the challenges faced in producing a useful labelled data set. Existing tools may lack the necessary features and functionalities to handle the complexities specific to urban drainage data. It is essential to invest in the development of advanced tools tailored to the unique requirements of urban drainage data labelling. These tools should incorporate intelligent algorithms, data validation mechanisms, and user-friendly interfaces to streamline the labelling process and maximize the utility of the labelled data set.

Addressing these challenges is critical for advancing research and practical applications in the field of urban drainage. By improving the quality of labelled data sets, we can enhance the accuracy and reliability of analytical models, decision support systems, and predictive algorithms used in urban drainage management. Efforts should be made to foster interdisciplinary collaborations, refine data collection methodologies, and develop specialized tools to overcome the obstacles faced in producing a useful labelled data set for urban drainage data. Only by surmounting these challenges can we pave the way for automated data quality control and successful machine learning applications in the field of urban drainage.

7 Acknowledgements

We would like to express our sincere gratitude to all individuals who contributed to the completion of this paper. Their valuable support and assistance were instrumental in shaping our research and enhancing the quality of our work. Special gratitude to the participants of the expert labelling and the preparation of it, namely: Stephan Baumgartner, Viviane Furrer, Simon Bloem, Pravin Ganesanandamoorthy, Bruno Hadengue, Alejandro Figueroa, Christoph Ort, Christian Förster, Natalia Duque und Frank Blumensaat. We deeply appreciate the contributions of each individual mentioned above, as well as anyone else who has contributed to this paper indirectly. Without their assistance, this experiment would not have been possible.

8 References

Aggarwal, C. C.: Outlier Analysis, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-47578-3, 2017.

A Look Behind Using Machine Learning for Anomaly Detection: https://www.trinnex.io/insights/a-look-behind-using-machine-learning-for-anomaly-detection, last access: 8 September 2023.

DWA-M 181 - Durchflussmessung (9/2011) | Lieferprodukt | M 181-11: https://shop.dwa.de/DWA-M-181-Durchflussmessung-9-2011/M-181-11, last access: 8 September 2023.

(PDF) Using Embedded Sensor Networks to Monitor, Control, and Reduce CSO Events: A Pilot Study: https://www.researchgate.net/publication/228632159_Using_Embedded_Sensor_Networks_to_Monitor_Control_and_Reduce_CSO_Events_A_Pilot_Study, last access: 8 September 2023.

Welcome - ERIC-open: https://opendata.eawag.ch/, last access: 8 September 2023.

Becouze-Lareure, C., Bazin, C., Namour, P., Breil, P., and Perrodin, Y.: Multi-Level Approach of the Ecotoxicological Impact of a Combined Sewer Overflow on a Peri-Urban Stream, J. Water Resour. Prot., 4, 984–992, https://doi.org/10.4236/jwarp.2012.411114, 2012.

Blumensaat, F., Ebi, C., Dicht, S., Rieckermann, J., and Maurer, M.: Langzeitüberwachung der Raum-Zeit-Dynamik in Entwässerungssystemen mittels Niedrigenergiefunk, KA Korresp. Abwasser Abfall, 594–603, https://doi.org/10.3242/kae2017.07.001, 2017.

Blumensaat, F., Bloem, S., Ebi, C., Disch, A., Förster, C., Rodriguez, M., Maurer, M., and Rieckermann, J.: The UWO dataset: Long-term data from a real-life field laboratory to better understand urban hydrology at small spatiotemporal scales, https://doi.org/10.31224/3208, 30 August 2023.

Branisavljević, N., Kapelan, Z., and Prodanovic, D.: Improved real-time data anomaly detection using context classification, J. Hydroinformatics, 13, 307, https://doi.org/10.2166/hydro.2011.042, 2011.

Brito, R. S., Almeida, M. C., and Ribeiro, A.: Enhancing hydraulic data reliability in sewers, Water Pract. Technol., 17, 431–444, https://doi.org/10.2166/wpt.2021.093, 2021.

Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F.: Dealing with difficult minority labels in imbalanced mutilabel data sets, Neurocomputing, 326–327, 39–53, https://doi.org/10.1016/j.neucom.2016.08.158, 2019.

Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A., and Pak, C.: Smarter Stormwater Systems, Environ. Sci. Technol., 50, 7267–7273, https://doi.org/10.1021/acs.est.5b05870, 2016.

Mathis, S., Gruber, J.-M., Ebi, C., Bloem, S., Rieckermann, J., and Blumensaat, F.: Energy self-sufficient systems for monitoring sewer networks, 21. ITG/GMA-Fachtagung Sensoren und Messsysteme 2022, Nürnberg, Deutschland, 10.-11. Mai 2022, Accepted: 2022-11-11T09:37:09Z, 2022.

Mourad, M. and Bertrand-Krajewski, J.-L.: A method for automatic validation of long time series of data in urban hydrology, Water Sci. Technol., 45, 263–270, https://doi.org/10.2166/wst.2002.0601, 2002.

Nedergaard Pedersen, A., Wied Pedersen, J., Vigueras-Rodriguez, A., Brink-Kjær, A., Borup, M., and Steen Mikkelsen, P.: The Bellinge data set: open data and models for community-wide urban drainage systems research, Earth Syst. Sci. Data, 13, 4779–4798, https://doi.org/10.5194/essd-13-4779-2021, 2021.

Northcutt, C. G., Athalye, A., and Mueller, J.: Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks, https://doi.org/10.48550/arXiv.2103.14749, 7 November 2021.

Russo, S., Disch, A., Blumensaat, F., and Villez, K.: Anomaly Detection using Deep Autoencoders for in-situ Wastewater Systems Monitoring Data, https://doi.org/10.48550/arXiv.2002.03843, 6 March 2020.

Tang, W. and Lease, M.: Semi-Supervised Consensus Labeling for Crowdsourcing, 2011.

Wang, Q., Zhang, M., Tao, T., and Sheng, V. S.: Labelling Training Samples Using Crowdsourcing Annotation for Recommendation, Complexity, 2020, e1670483, https://doi.org/10.1155/2020/1670483, 2020.