SCOREwater

Smart City Observatories implement REsilient Water Management

DELIVERABLE D2.4

1ST VERSION OF DATA-DRIVEN MODELS REPORT FOR A WATER-SMART SOCIETY

Dissemination level	Public
Type	Report
Issued by	Eurecat
Contributing project partners	Hydrologic
Author(s)	Rubión, Edgar (ER); Ribalta, Marc (MR); van den Brink, Matthijs (MB); de Roover, Sam (SR)
Reviewed by	Corominas L., Lindblom E., Meijer H.
Keywords	AI, Machine Learning
Number of pages	136
Number of annexes	4
Date:	2020-07-28
Version:	V 1
Deliverable number	D2.4
Work Package number:	WP 2
Status:	Delivered
Approved by coordinator (IVL)	2020-07-28

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Copyright notices

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

REVISION HISTORY

Version	Reason for changes	Name	Date
1	Original release to EU	Edgar Rubión	2020-07-31

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

CONTENT

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

LIST OF FIGURES

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

LIST OF TABLES

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ABBREVIATIONS

Abbreviation	Definition
AI	Artificial Intelligence
Ada-boost	AdaBoost Classifier
CKAN	Compressive Knowledge Archive Network
CoA	City of Amersfoort
CRISP-DM	Cross Industry Standard for Data Mining
CSV	Comma Separated Values
DTC	Decision Tree Classifier
Dx.x	Deliverable x.x
EDA	Exploratory Data Analysis
EMA	Exponential Moving Average
EWS	Early Warning System
FN	False Negative
FP	False Positive
GBR	Gradient Boosting Regressor
IF	Isolation Forest
IoT	Internet of Things
KNN	K-Nearest Neighbours
KNR	K-Neighbours Regressor
LCA	Linear Discriminant Analysis
LOF	Local Outlier Factor
LR	Linear Regression
MA	Moving Average
NAB	Numenta Anomaly Benchmark
O-SVM	One class Support Vector Machines
P10	Percentile 10
P90	Percentile 90
PPV	Positive Prediction Value
Q1	Quartile 1
Q3	Quartile 3
QDA	Quadratic Discriminant Analysis
RFC	Random Forest Classifier
SDG	Sustainable Development Goals
SD	Standard Deviation (σ)
TN	True Negative
TP	True Positive
TPR	True Positive Rate

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Abbreviation	Definition
TSV	Tab Separated Values
UHI	Urban Heat Island
UoM	Unit of Measurement
URL	Uniform Resource Identifier
WaSCs	Water and Sewerage Companies
WPX	Work Package X

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

PROJECT ABSTRACT

SCOREwater focuses on enhancing the resilience of cities against climate change and urbanization by enabling a water smart society that fulfils SDGs 3, 6, 11, 12 and 13 and secures future ecosystem services. We introduce digital services to improve management of wastewater, stormwater and flooding events. These services are provided by an adaptive digital platform, developed and verified by relevant stakeholders (communities, municipalities, businesses, and civil society) in iterative collaboration with developers, thus tailoring to stakeholders' needs. Existing technical platforms and services (e.g. FIWARE, CKAN) are extended to the water domain by integrating relevant standards, ontologies and vocabularies, and provide an interoperable open-source platform for smart water management. Emerging digital technologies such as IoT, Artificial Intelligence, and Big Data is used to provide accurate real-time predictions and refined information.

We implement three large-scale, cross-cutting innovation demonstrators and enable transfer and upscale by providing harmonized data and services. We initiate a new domain "sewage sociology" mining biomarkers of community-wide lifestyle habits from sewage. We develop new water monitoring techniques and data-adaptive storm water treatment and apply to water resource protection and legal compliance for construction projects. We enhance resilience against flooding by sensing and hydrological modelling coupled to urban water engineering. We will identify best practices for developing and using the digital services, thus addressing water stakeholders beyond the project partners. The project will also develop technologies to increase public engagement in water management.

Moreover, SCOREwater will deliver an innovation ecosystem driven by the financial savings in both maintenance and operation of water systems that are offered using the SCOREwater digital services, providing new business opportunities for water and ICT SMEs.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

EXECUTIVE SUMMARY

The goal of D2.4 is to provide the first report describing designed and trained data-driven models. This description also includes the data pre-processing techniques used to split the information, detect and correct outliers, eliminate unrepresentative features and feature engineering. Additionally, information related to modelling and validation is also provided. It is important to note that a new version of this document will be presented on M36, enhancing current accuracy of data-driven models, and adding new ones.

D2.4 corresponds with the outcome of Task 2.2 "Exploratory data analysis, data cleansing and feature engineering", 2.3 "Design of advanced Machine Learning models" and 2.4 "Assessing models and algorithms".

The data-driven models are focused on 3 study cases:

- **Amersfoort case** focuses on providing smart models to improve the resilience in front of flood, heat and drought risk;
- **Barcelona case** focuses on advancing towards a resilient sewage system based on a prescriptive management; and
- **Gothenburg case** focuses on enhancing the urban resilience by smart monitoring construction pollution events.

The work done during this first year, which is described in this deliverable, includes all sorts of Machine Learning techniques such as sensor simulation, outlier detection, spatial predictions, data quality evaluation, drift detection and anomaly detection.

Amersfoort has available a dense network of temperature low-cost and hand-built sensors for measuring the impact of climate change on the city. The accuracy of the sensors is limited and hence, all the measured time series must be validated and corrected manually. This deliverable presents the first iteration of a data-driven model to automatically validate temperature data.

The Barcelona use case introduces the idea of predicting sediment level in all the sewer grid using spatial prediction, considering not only physical properties of the section but also properties of the nearby sewer sections and sediments to predict the sediment build in a specific section.

The Gothenburg use case presents a solution to early warning of pollution events on water of construction sites based on Novelty Detection, that is, detecting abnormal patterns in the water quality measurements.

Additionally, generic data-driven models are faced during this first year, which are used to detect abnormal flow patterns in the sewage system and to detect drifting behaviour of sensors.

The main outcomes are:

- A data-driven model based on algorithm Histogram-based *Gradient Boosting Classifier* to temperature validation. Low initial *Recall* Score and *Precision* Score was obtained, but it will be increased in the future enhancing the data model, adding spatial features, and improving the quality of the registers;
- A data-driven model based on *Gradient Boosting Regressor* algorithm to predict sediment accumulation in the sewer grid considering not only physical properties of the section but also properties of the nearby sewer sections and sediments. Results were good to predict low sediment accumulation, nevertheless they were worst with high sediment accumulation. More registers are expected during the rest of the SCOREwater project allowing improving the global results. Additionally, new strategies will be faced, like the prediction of future sediment level in a section using the trend of the past values;
- A data-driven model to predict anomalies in the water from construction sites based on algorithm *Isolation Forest*. Due to the collected data quality, only one month of normality was used to train the model and two anomalies to evaluate the model. Despite this data issue, the model results were good enough to plan future iterations following the same approach, *Novelty Detection*. More quality data will be collected during the project to improve the data-driven model;

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

- A data-driven model to detect anomalies on water quality sensors, obtaining a good scoring despite of a counterpart of detecting some false positives. *One-class Support Vector Machine* was the most reliable algorithm tested. In the future, more data will be added, and the team will experiment with deep learning algorithms such as *Deep Belief Networks* or hybrid solutions between auto-encoders and *O-SVM*;
- A data-driven model to drift detection on ammonium and turbidity sensors. The study compares a batch of classification algorithms, from linear predictions to ensembles and neural networks, to evaluate real-time anomaly detection models. The empirical results highlight the feedforward *Artificial Neural Network* as the best model, obtaining high *NAB* and *Precision* scoring; and
- A data-driven model to validate data quality of flow patterns. The data-driven model built by using the *Quadratic Discriminant Analysis* algorithm demonstrates encouraging results, especially if it is trained with a sufficiently large and representative dataset. The optimization of hyperparameters could improve current results, and hence should be considered during the second year of the SCOREwater project

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

1. INTRODUCTION

SCOREwater focuses on enhancing the resilience of cities against climate change and urbanization by enabling a water smart society that fulfils SDGs 3, 6, 11, 12 and 13 and secures future ecosystem services.

WP2 is aimed at providing a set of data-driven models to build smart water infrastructures supporting urban resilience, contributing to fulfil KPI 2, 4, 5, 6 & 17. Then, the WP exploits heterogeneous data and apply novel data analytics and machine learning techniques to create and validate smart water services. Data are enhanced with sharp capabilities such as signal conditioning for missing data and outliers.

The goal of D2.4 "1st version of data-driven models report for a water smart society", which corresponds with the outcome of Task 2.2 "Exploratory data analysis, data cleansing and feature engineering", 2.3 "Design of advanced Machine Learning models" and 2.4 "Assessing models and algorithms", is to provide the first report describing designed and trained data-driven models. This description also includes the data pre-processing techniques used to split the information, detect and correct outliers, eliminate unrepresentative features and feature engineering.

It is important to remark the interdependencies and relationships between this deliverable and the rest of work packages and deliverables. D2.4 takes advantages of the datasets gathered on D2.1 "Testbed data and sensor validation" to advance in the development of the first version of data-driven models. D2.4 also takes advantage of the guidelines and scripts provided by Task D2.6 "1st version of streamlined model evaluation environment" to ensure the quality of data-driven models. Finally, the collected results of D2.4 will be integrated in the SCOREwater platform through Task 3.2 "Integration of sensors, algorithms and models".

1.1. SCOPE

As part of the SCOREwater project, three case studies are faced in WP2.

Amersfoort case focuses on the potential impacts of climate change in an urban environment and on the effectiveness of adaptation measures. Three potential impacts are monitored: (a) **flood risk** due to intensified precipitation urban public space is increasingly vulnerable to flooding, caused by several mechanisms (precipitation unable to enter the drainage system or sewerage system, water flowing out of the sewerage system through spill-ways or manholes and flooding from surface water or ground water); (b) **heat risk** due to increasingly longer periods of high temperatures the impact of the urban heat island effect increases as well; and (c) **drought risk** due to increasingly longer periods without precipitation, urban public vegetation suffers from water shortages

The objective of the Amersfoort case is to assess the impacts of these risks and to investigate the effectiveness of measures taken to reduce the effects. To this end, for several sites in Amersfoort monitoring networks have been designed that cover the dominant variables related to these risks. Where available, existing sensors and sensor networks have been used. The main variables observed are temperature, air humidity, soil moisture and ground water levels. Additionally, surface water and sewerage data – water levels, discharge, pumping hours – are made available.

Currently, the city council of Barcelona has maintenance and cleaning routines based on sediment levels to reduce the risk of blockages and odours. The **Barcelona case** focuses on assessing the potential impacts of human behaviour (for example, not-allowed discharges) and natural factors (for example, infrastructures degradation, rain) on sediment accumulation. For that, sedimentation will be monitored by applying novelty techniques based on AI. The prediction of sedimentation accumulation will allow to minimize the need of physical inspections of the sewage system, with the consequent improvement of the quality of life of workers. Additionally, Barcelona will advance towards a more resilient sewage system managed in a prescriptive way.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Construction industry is one of the major sources of pollution, responsible for around 4% of particulate emissions, more water pollution incidents than any other industry, and thousands of noise complaints every year (Gray, 2020). Gothenburg minimises such silt and pH pollution by installation of portable and monitored treatment stations on building sites. Nevertheless, the application of pre-emptive techniques to anticipate the problems, that is, prepare for the unexpected can be key to face pollution events. In this section, the design of data-driven models to provide an early warning system for water pollution events on construction is addressed. Then, the **Gothenburg case** focuses on studying novel techniques based on AI to early warning of pollution events on water of construction sites.

Finally, Quality Assurance (QA) plays an essential part in any analytical project to ensure the validity and reliability of data. Effective QA ensures that decisions are made with an appropriate understanding of evidence and risks, and helps analysts ensure the integrity of the analytical output. Until now, two different data-driven models related to QA have been provided on SCOREwater. One for detecting anomaly flow patterns on sewage system and another to detect drift on water quality sensors.

1.2. DOCUMENT OUTLINE

In particular, this document provides key information about the data-driven models, including:

a) a brief introduction to the CRISP-DM methodology, which is followed throughout the project to design and validate the data-driven models (see Section 2);

b) one section for each study case, where is depicted deeply each data-driven model, how they were designed and their results (see Section 3, 4 and 5);

c) a section for describing the general data-driven models related to data quality assurance (see Section 6);

d) main conclusions of the analysis (see Section 7);

e) external references cited throughout the document (see Section 8); and

f) annexes including useful information to understand the data analysis such as Data Science and Machine Learning concepts (see Annex 1) and Scoring Metrics (see Annex 2), among others.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

2. HOW THE DATA-DRIVEN AND AI MODELS ARE CREATED

The design of the data-driven and AI models is based on a robust and well-proven methodology for data mining, *CRISP-DM*. The *CRISP-DM* methodology (Wirth & Hipp, 2000) provides a structured approach, based on an idealized sequence of events, to planning a data mining project. It is flexible, and in practice, many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. Below, Figure 2-1 presents the steps of the *CRISP-DM* methodology:

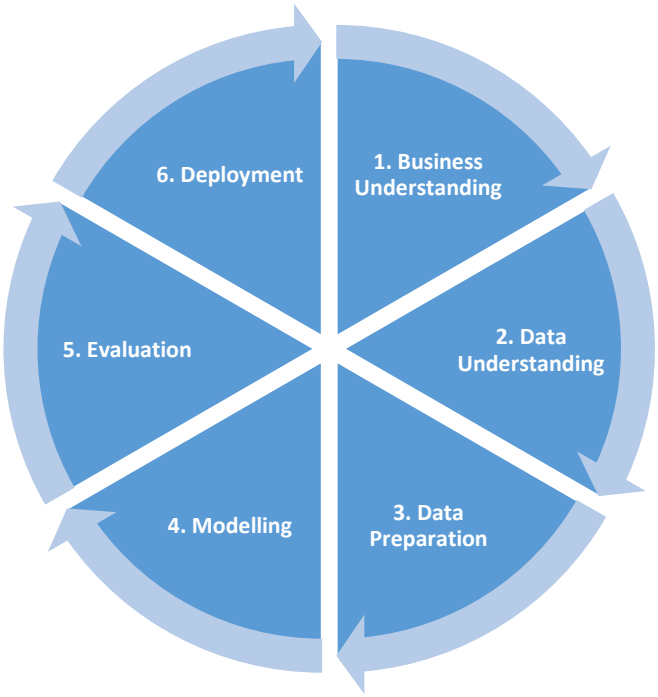

Figure 2-1. Structured approach of CRISP-DM methodology

2.1. FIRST STAGE: BUSINESS UNDERSTANDING

The first stage (1. Business Understanding) goal of the *CRISP-DM* methodology is to uncover important factors that could influence the outcome of the project. For that, it is important to understand what you want to accomplish from a business perspective, including the objectives and available resources.

Basically, this stage follows four steps:

- define objectives from a business perspective;
- describe the current situation of resources;
- define objectives from an AI point of view; and
- list data mining success criteria.

To set business objectives, the primary objective should be described from a business perspective, including other related questions that you would like to address. For example, the primary goal for the Barcelona case might be to minimize the sewage blockage by predicting if it will be blocked in the near future. Related business questions might be "Do rainy days impact sewer blockage?" or "Will the socio-economic level of sewerage users affect the blocking frequency?".

The second step is to assess the current resource situation, where the business provides detailed information about all the resources to be considered. To achieve a consistent result, an inventory of resources should be provided, listing the available resources for the project. It should include:

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

- data (fixed extracts, access to live, warehoused, or operational data);
- computing resources (hardware platforms); and
- software (data mining tools, other relevant software)

Concerning *determine AI objectives* step, a business goal states objectives in business terminology, while a data mining goal sets project objectives in technical terms. For example, the business goal might be to "minimize the sewage incidences". A data mining goal might be "Predict the probability of sewage section being blocked in a few weeks, given their maintenance actions over the past three years, infrastructure information (length, section, material….), and demographic and weather information".

Finally, data mining success criteria should be defined. They are the criteria for a successful outcome to the project in technical terms, for example, a certain level of predictive accuracy.

2.2. SECOND STAGE: DATA UNDERSTANDING

The second stage of the *CRISP-DM* process requires to acquire the data listed in the project resources to explore and analyse them and extract the understanding. The stage has four steps, which are:

(i) collect initial data;

(ii) describe data;

(iii) explore data; and

(iv) verify data quality

The initial collection includes data loading, if this is necessary for data understanding. For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. If you acquire multiple data sources, then you need to consider how and when they are integrated.

A table with an initial data collection report is provided, it lists the data sources acquired together with their locations, the methods used to acquire them and any problems encountered and their resolutions achieved.

Table 1. Template for details about data source acquisition

Datasource	Location	Method used to acquire	Problems
Datasource 1	n/a	Send via email	n/a
Datasource 2	URL XXX	Query to REST API with python script	No problems identified

Additionally, a description of the data is essential to understand them, therefore they should be examined the "gross" or "surface" properties of the acquired data should be reported. For that, a table template is provided, which includes for each data source a brief description, its format, its number of records and fields.

Table 2. Template for general details about data source

Data Source	Description	Format	# Registers	# Fields
Data source 1		CSV		
Data source 2		TXT		

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 3 enhances the collected information about the data sources, describing the features part of the data sources. The description includes the identifier of the feature, a brief description, the type of information (numerical, date, alphanumerical, categorical…), the unit of measurement (UoM) and the data source of which it is part.

Table 3. Template for general details about features

Feature	Description	Type	UoM	Data Source
Length	Length of sewage pipe section	Numerical	Meters	Data Source 1

Another important step is to explore the features of collected data by using data mining querying, data visualization, and reporting techniques. This Exploratory Data Analysis (EDA) is comprised by a number of steps listed below:

(i) identification of variables and data types;

analysis of the basic metrics such as the mean, standard deviation (σ or SD) of each variable;

(ii) graphical univariate analysis using box plots, histograms, pie charts, etc…;

(iii) multivariate analysis using scatter plots, area plots and 3d plots; and

(iv) correlation analysis using correlation matrix, and in case of using time series data, autocorrelation and cross correlation.

These analyses may directly address the data mining goals. They may also contribute to or refine the data knowledge, and feed into the transformation and other data preparation phases needed for further analysis. The results of the data exploration are described through tables and plots, including first findings or initial hypotheses. Moreover, data quality is verified, checking if data features are correct or contain errors or missing values. In case of finding errors or missing data, determine when happen and how common these events are.

2.3. THIRD STAGE: DATA PREPARATION

This is the stage of the project where the dataset is produced and described to be used during the modelling stage. This stage contains five steps:

(i) select data;

(ii) clean data;

(iii) construct data;

(iv) integrate data; and

(v) format data.

The features and quantity of them to be used for analysis is decided and reasoned, applying criteria based on the relevance of the data, data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types. Once the features have been selected, they are cleaned by applying different techniques with the aim of raising the data quality required. The decisions and actions taken to clean the data are documented. Also, the process of creation of new features and records, or transformation of themselves is detailed during this stage. Multiple tables or data sources are usually available to build the models, so they are merged and combined in order to create a single dataset. This process, which also includes the aggregation (e.g. accumulated rainwater during the last week), will be documented.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

2.4. FOURTH STAGE: MODELLING

The fourth stage is aimed at designing accurate models to face the *AI objectives* defined on first stage (business understanding). For that, there are four steps:

(i) select the data-driven algorithm;

(ii) generate the train-test environment;

(iii) build model; and

(iv) assess model.

As the first step in modelling, the specific machine learning algorithm or algorithms are selected (e.g. Support Vector Machine, AdaBoost) taking advantage of the conclusions extracted during the second and third stage, data understanding and data preparation respectively. The intended plan for training, testing, and evaluating the models are described. It is important to note that D2.6 "First Version of streamlined model evaluation environment" (M12) describes how to assess the data-driven models, including the list of key reference indicators. To run the modelling tool on the prepared data set, the tuning of the hyper parameters (parameters of the Machine Learning algorithms) is essential. It is depicted including the used reasoning to adjust the parameters. Once the model is built, the results are described including the interpretation of the models according to the domain knowledge and the data mining success criteria, which is defined in first stage (business understanding). Therefore, the results are only judged by the analytics point of view. Later, the outcomes are validated taking into account domain expert knowledge on stage five. Finally, if several models are created, they are ranked according to the evaluation criteria providing a list of generated models qualities.

2.5. FIFTH STAGE: EVALUATION

The fifth stage, *evaluation*, is addressed to assess the efficiency and generalization of the model designed throughout the previous stage, *modelling*, from the business point of view. To sum up, there are three steps:

(i) evaluate the results;

(ii) review the process; and

(iii) determine next steps.

During this stage, the degree to which the designed model fits with the business and AI objectives will be assessed jointly with domain experts. Moreover, if the model obtained is deficient, it will be seeking to determine if there is some business reason. Depending on the results of the assessment and the process review, how to proceed will be decided through a list of possible actions and decisions.

2.6. SIXTH STAGE: DEPLOYMENT

The main aim of the deployment stage is to integrate the designed models in an architecture or module to be executed on real time or batch. This aim is out of the scope of the WP2, where only data-driven models are created and persisted. WP3 includes a task, Task 3.2 "Integration of sensors, algorithms, and models", whose objective includes the deployment of the models. Therefore, it will be addressed by the deliverable D3.3 "Integration and connection of sensors and algorithms to the SCOREwater Platform, including processing, storage and transformation to Open API".

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

3. AMERTSFOORT CASE

3.1. FLOOD EARLY WARNING

The City of Amersfoort (COA) is increasingly vulnerable to flooding due to intensified precipitation. On July 28, 2014, a precipitation event caused flooding throughout the city. The water blocked several tunnels and the entrance to a railway station, causing congestion and interruptions of the train service. An analysis of the event showed that the statistical return period for the event drops from 100 years in the climate of 2014 to 20 years in the climate of 2050. This event, and similar, less extreme situations in the past years, raised the awareness of the potential impact of relatively frequent flooding and an increased interest in preventive measures, like an early warning system.

3.1.1. ITERATION 1

3.1.1.1. BUSINESS UNDERSTANDING

The business objective of the flood early warning system is to create a window of opportunity for the City of Amersfoort (COA) to take preventive measures (e.g. warn citizens or the fire department, set up road blocks/detours) aimed at reducing the negative impacts of a precipitation event.

COA has identified two locations, where the potential impact of flooding is highest:

1. the tunnel at the Schothorst railway station. Flooding of the tunnel blocks the entrance to and the exit from the railway station; and
2. the 'Stadsring' tunnel. The Stadsring is a ring road around the city centre and is a critical part of the approach routes for police, ambulance, and fire department.

In the first iteration, the flood early warning system will be focused on these two locations.

The window of opportunity that could be created by the early warning system is a trade-off between the accuracy of the system, the forecast horizon, and the quality of the input data. The uncertainty of the precipitation forecast increases with the forecast horizon, i.e. the precipitation depth can be forecasted more accurately 2 hours ahead than 6 hours ahead. The precision metric is the ratio between the number of correctly detected anomalies and the number of the predicted anomalies, and will be used to evaluate the quality of the algorithm (more detailed information on Annex 1). In the first iteration, a forecast horizon of 2 hours will be used and a time step of 15 minutes: every 15 minutes a new binary forecast (flooding/no flooding) is produced for the next two hours.

As historical flooding events are rare, there are only few observations to train a data driven model. Therefore, an artificial data set has to be produced using a hydrodynamic model of the sewerage system and street levels of the city of Amersfoort. By feeding this model many different precipitations events and recording different aspects of the floods that are simulated (occurrence, location, duration, level), a dataset is obtained with which a data driven model can be trained and validated. This hydrodynamic model (D4.17) is currently in preparation, which implies the necessary data sets for the flood early warning system are not yet available. The first iteration will start as soon as the data is available (expected in the fall of 2020).

3.1.1.2. DATA UNDERSTANDING

To create an early warning system described above, it is necessary to find a relationship between the precipitation intensity - i.e. a combination of depth and duration - and the occurrence of flooding at the specified locations. There is no historical data set available of flooding events, except for the visual observation of flooding of the Schothorst tunnel on July 28, 2014.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Therefore, it is necessary to create a dataset that can be used to train a classifier algorithm. This dataset can be created with the help of a hydrodynamic model of the sewerage system of the city of Amersfoort, by feeding a set of varying precipitation events to this hydrodynamic model and recording occurrence, duration and severity (i.e. maximum water depth) per event.

The required hydrodynamic model is currently under construction as part of deliverable D4.18 of this project. As a consequence, the required data for the Flood Early Warning system is not yet available. This use case will be described in deliverable D2.5.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

3.2. HEAT RISK

Climate change causes longer periods of hot weather, and a higher rate of occurrence of extreme temperatures. COA wants to investigate heat stress (observed and experienced), and how COA can (re)develop its city in such a way that it is able to deal with rising temperatures.

3.2.1. ITERATION 1

3.2.1.1. BUSINESS UNDERSTANDING

To measure heat in Amersfoort, a relatively dense network of temperature sensors was installed and has been operated by a citizen initiative 'MeetJeStad'(MJS) since 2017. As these sensors have been hand built at costs as low as possible, the accuracy of the sensors is limited. In order to use the data, the time series have to be validated and corrected. As validation by hand is tedious and subject to inconsistencies or errors, the MJS platform and COA would like to implement an automated data driven validation service. As a first iteration, the AI objective of the data driven classification model is to detect and flag outliers and other non-NaN (Not a Number) value anomalies, with recall (or accuracy) and precision score of 80%. The recall (or accuracy) metric is the ratio between the number of correctly detected anomalies and the number of the observed anomalies, whereas the precision metric is the ratio between the number of correctly detected anomalies and the number of the predicted anomalies.

Besides the MJS data source, COA has installed a second network of temperature sensors, at fewer locations but with sensors that are expected to have a higher accuracy level than the MJS sensors. To quantify the difference in quality of both data sets, the data driven model should be operated on both data sets.

3.2.1.2. DATA UNDERSTANDING

Three different data sources are found that describe temperatures in the city at several locations (Table 4 and Table 5). The first and second data source are respectively the measurement databases and validation of the temperature measurements in the MeetJeStad (MJS) project. In this citizen-initiated, citizen-science project, temperature and humidity are measured with stationary sensors at multiple locations spread across Amersfoort. 148 sensors are registered by the MJS organisation, but not all provide (meta) data and are thus not used in this project. In the first iteration of the Heat-risk case, 35 sensors in the Schothorst neighbourhood are selected. The third data source is similar to the previous two, except that measurements are taken by the city of Amersfoort (COA) with 5 sensors. The MJS project is aimed at understanding the urban climate, whereas the COA sensors are mainly placed for measuring air quality in the city and thus measure 6 more variables besides temperature and humidity.

Table 4. Data source acquisition – Heat Risk

Data source	Location	Method used to acquire	Problems
RAW MeetJeStad	http://meetjestad.net/data?type=sensors&start=YYYY-MM-DD,HH:mm&end=YYYY-MM-DD,HH:mm&ids=#[,#[,#-#]]&format=csv Or https://meetjestad.net/data/	Query to API via Python Or Query to MeetJeStad API	CSV is in truth tab delimited
Validated MeetJeStad		Send via e-mail	No problems identified

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Data source	Location	Method used to acquire	Problems
Air quality COA		Send via e-mail	No problems identified

Table 5. General details about data source – Heat risk

Data Source	Description	Format	# Registers	# Fields
RAW MeetJeStad	MeetJeStad raw data	CSV	148, 35 used in first iteration	20, but 5 usable
Validated MeetJeStad	MeetJeStad validated temperature data	CSV	148, 35 used in first iteration	2
Air quality COA	City of Amersfoort raw data	CSV	5	11

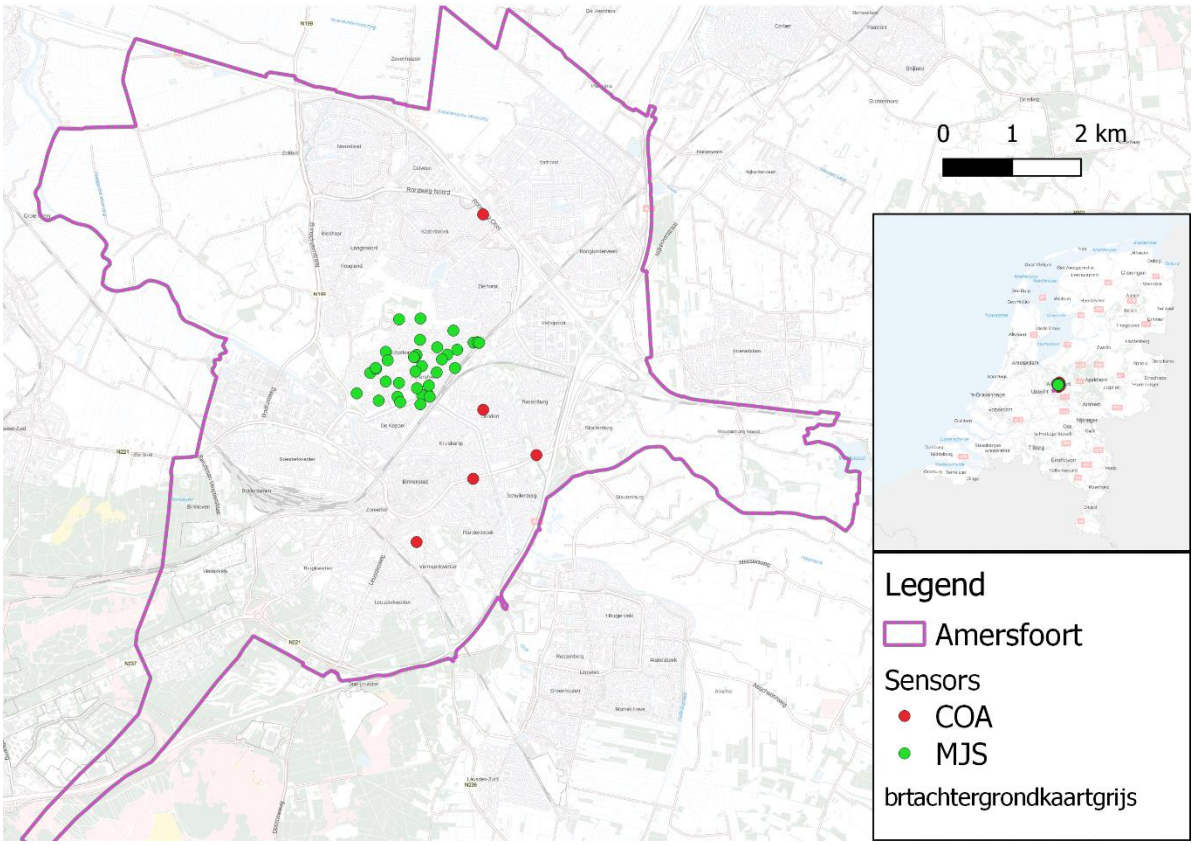

Figure 3-1. Spread of the MeetJeStad (MJS) and city of Amersfoort (COA) sensors

Figure 3-1 presents the MJS sensors providing (meta) data and COA sensors. All variables provided by the data source files, together with integrated metadata, are shown in Table 6.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 6. General details about features – Heat risk

Feature	Description	Type	UoM	Data Source
Id	Sensor identification code	Numerical	N/A	RAW MeetJeStad
Timestamp	Time at measurement	Date	YYYY-MM-DD HH:mm:ss	RAW MeetJeStad
Coordinates	Longitude and latitude of the sensor	Coordinates	Degrees with WGS84	RAW MeetJeStad
Temperature	Temperature measured at sensor	Numerical	Degrees Celsius	RAW MeetJeStad
Humidity	Humidity of the air measured by sensor	Percentage	N/A	RAW MeetJeStad
Timestamp	Time at measurement	Date	YYYMMDDHHmm / YYYY-MM-DD HH:mm:ss	Validated MeetJeStad
Temperature	Temperature measured at sensor[1]	Numerical	Degrees Celsius	Validated MeetJeStad
Device id	Sensor identification code	Numerical	N/A	Air quality COA
Timestamp	Time at measurement (meetmoment)	Date	YYYY-MM-DD HH:mm:ss	Air quality COA
Record id	Record identification code (rij)	Numerical	N/A	Air quality COA
Air pressure	Air pressure at sensor (s_barometer)	Numerical	hectopascal	Air quality COA
CO_2 concentration	Concentration of carbon dioxide measured by sensor (s_co2)	Numerical	Parts per million	Air quality COA
Humidity	Humidity of the air measured by sensor	Percentage	N/A	Air quality COA
NO_2 concentration	Concentration of nitrogen dioxide measured by sensor (s_no2)	Numerical	Parts per million	Air quality COA
PM10 concentration	Concentration of micro particles smaller than 10 micrometer measured by sensor	Numerical	Parts per million	Air quality COA
PM2.5 concentration	Concentration of micro particles smaller than 2.5 micrometer measured by sensor	Numerical	Parts per million	Air quality COA
Temperature	Temperature measured at sensor	Numerical	Degrees Celsius	Air quality COA
Sound pressure level	Sound pressure level measured by sensor (v_audio_total)	Numerical	decibel	Air quality COA

[1] Per row temperatures are given for all sensors.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

In the following, the different data sources are described in more detail. Besides, results of simple analyses and visualisations of the data are given.

Data sources 1 & 2 – MeetJeStad

The raw data (data source 1) can be acquired via a connector written in Python or via the website https://meetjestad.net/data/. Data is collected by citizens, part of a citizen science project to give insights in the urban climate. Due to this, measurement errors may be present in data (wrong placement of sensor stations, etc.). However, the organisation also provided a file with validated temperature data (data source 2).

In the first iteration of the heat-risk case, data was gathered for 35 sensors from the MJS database between January 1st, 2018 and December 31st, 2019. Below, a summary is given of the copied data (data source 1):

- Start date: Varying per sensor
- End date: Varying per sensor
- Interval: Irregular
- Sensors: 35, spread irregularly across Schothorst neighbourhood in Amersfoort
- Quality: dubious, but: clean timeseries available (data source 2)

The validated dataset (data source 2) contains temperature data between January 1st, 2018 and October 19th, 2018 for the same 35 sensors as in the raw dataset. The records are given at a regular interval of 15 minutes.

The following procedure was used to create the validated temperature dataset:

1. raw data intervals are regularised to 15-minute intervals, by averaging the data in each 15-minute bin
2. citywide temperature quartile 1 (Q1), quartile (Q2) and quartile 3 (Q3) values are determined for each timestamp
3. at each timestamp possible temperature anomalies are identified by being smaller than Q1 or larger than Q3
4. a sub selection is made of the values that are 6 °C larger or smaller than the citywide Q2/median
5. the selected values are compared with local values at the corresponding sensor and are deleted if they also do not fit the trend of the sensor data (opinion of the data scientist)

Anomalies that remain in the dataset after validation are due to:

- sensors taken indoors (to i.e. prevent being vandalised)
- relocation of the sensor
- placement of the sensor

The remaining anomalies may be removed by a more rigorous validation procedure, but a large probability exists that ‘real’ values are then deleted rather than anomalies.

Figure 3-2 shows temperature and humidity measurements for one of the sensors, together with the validated temperature data. The yearly seasonality of data is clearly visible, and so is the daily seasonality if a closeup is made.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

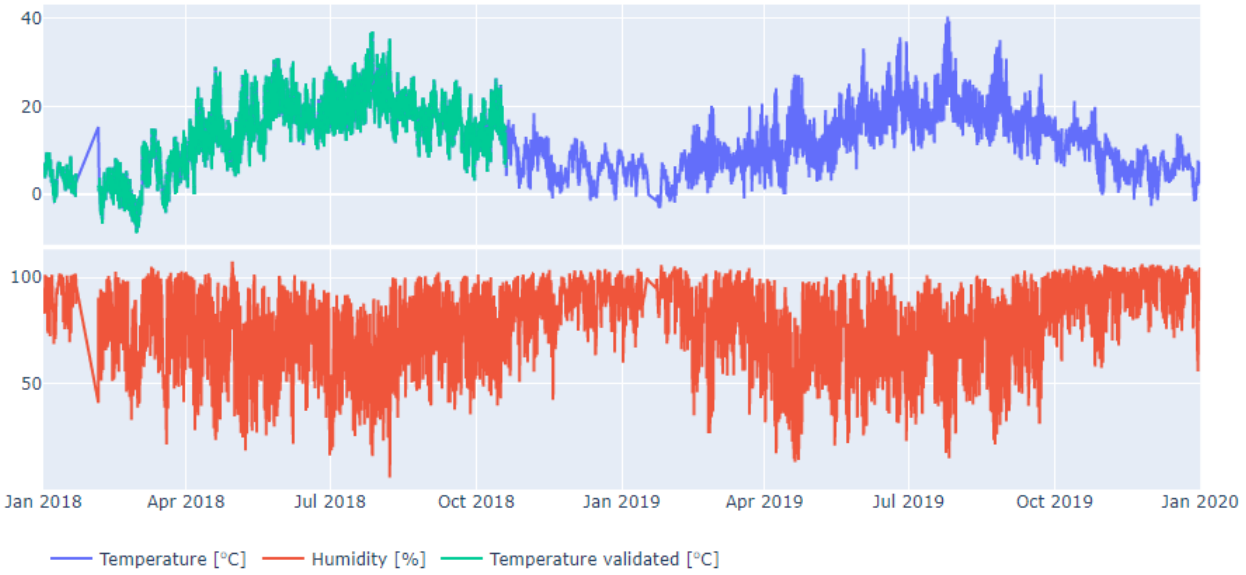

Figure 3-2. Temperature and humidity data as acquired for sensor 25

Only the temperature data was interesting for determining heat risk. A global statistical analysis of the temperature data was done for all sensors. Table 7 is the result. The statistics for the nearby KNMI measurement station 'De Bilt' are added for comparison. The following conclusions could be drawn from the global statistical analysis:

- Sensor 43 has the least amount of data gaps when compared to the recording period (729 days of observations with only 1 day missing)
- The average and median values aligned, although the average is in most cases slightly higher than the median, sometimes up to 1 ºC
- A likely no data value for temperature is -26 ºC, since the minimum measured temperature should be around -8 ºC
- Maximum temperatures of 40-43 ºC should probably not be considered as anomalies, since these extremes did occur during July 2019. However, the maximum temperature at sensor 148 (46.6 ºC) seems to be too high and at sensor 272 (115.8 ºC) seems to be a measurement error
- Statistical values for sensors that measured for most of the time between January, 2018 and December, 2019 are similar to the values of the KNMI measuring point, although the average, median, minimum, and maximum temperatures are slightly higher, possibly due to the Urban Heat Island-effect

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 7. MeetJeStad sensor statistics with statistics of temperatures measured at the KNMI measurement station De Bilt

Variable	Temperature							General						
Sensor	Average	Median	Max	Min	Standard deviation	Variance	Skewness	Number Of Observations	Start Date	End Date	Total Days	Days Without Observations	Longitude	Latitude
13	6	5.6	28.9	-7.9	5.5	30.2	0.6	9150	01-01-2018 10:02	11-06-2019 09:41	525	421	5.391367	52.17203
25	12	11.4	40.4	-8.6	7.6	57.8	0.3	57426	01-01-2018 10:04	31-12-2019 23:54	729	19	5.381648	52.17167
33	13.1	13.8	37.8	-8.2	8.6	74.6	0	24915	01-01-2018 10:16	18-10-2018 19:21	290	0	5.398068	52.17666
43	11.9	11.3	42.6	-26	7.5	56.7	0.3	63652	01-01-2018 02:42	31-12-2019 23:59	729	1	5.395629	52.17291
49	10.9	9.4	37.7	-7.6	7.6	58.3	0.4	35926	01-01-2018 00:01	19-06-2019 07:55	534	123	5.380412	52.17114
59	13.1	13	42.6	-8.4	7.7	59.5	0.3	36344	01-01-2018 11:08	31-12-2019 23:49	729	191	5.392614	52.16843
63	13.6	13.4	38.3	-1.9	6.7	45.2	0.3	39658	07-07-2018 19:06	31-12-2019 23:45	542	59	5.390015	52.17133
71	23.6	23.1	37.4	13.2	5.1	25.9	0.4	1855	07-07-2018 19:24	05-08-2018 19:29	29	7	5.393013	52.16805
76	12.7	12.3	41.1	-8.4	7.7	59	0.3	53464	01-01-2018 00:16	31-12-2019 23:52	729	83	5.396763	52.17353
88	11.1	10.1	36.4	-6.5	6.9	47.9	0.5	27786	07-07-2018 20:15	12-06-2019 11:05	339	7	5.386767	52.16733
109	12.3	11.7	42.7	-8.8	7.8	61	0.4	55394	01-01-2018 00:07	31-12-2019 23:47	729	49	5.386544	52.17813
110	12.3	11.9	39.8	-8.3	7.6	57.6	0.2	50342	01-01-2018 00:42	29-08-2019 09:29	605	0	5.40234	52.1751
122	12.1	11.4	40.5	-8.7	7.8	61.3	0.3	49872	07-02-2018 14:28	31-12-2019 23:56	692	88	5.394515	52.17122
127	12.2	11.7	42	-8.4	7.6	57.8	0.3	53121	01-01-2018 00:13	31-12-2019 23:58	729	50	5.394623	52.17448
139	13.2	12.9	40.9	-6.5	7.4	55.1	0.3	38259	30-06-2018 16:54	02-10-2019 02:11	458	0	5.389784	52.17318
145	11.8	10.8	38.6	-6.8	7.4	55	0.5	30843	07-02-2018 16:52	08-07-2019 06:28	515	156	5.386188	52.16803
148	13.3	12.8	46.6	-6.9	7.8	60.4	0.4	35248	30-06-2018 16:50	09-09-2019 21:08	436	0	5.390242	52.17348

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Sensor	Average	Median	Max	Min	Standard deviation	Variance	Skewness	Number Of Observations	Start Date	End Date	Total Days	Days Without Observations	Longitude	Latitude
173	12.2	11.1	36.4	-4.6	7.1	49.8	0.5	29309	30-06-2018 17:14	12-06-2019 13:46	346	0	5.389696	52.17321
174	18.4	18.1	38.1	2.6	6.2	38.3	0.3	8347	04-02-2018 20:21	18-10-2018 20:25	256	158	5.390985	52.17546
177	13.1	12.7	40.6	-6.8	7.4	55.2	0.3	36771	04-02-2018 19:18	20-09-2019 04:35	592	150	5.391434	52.16823
182	10.4	9.4	37.1	-9.3	7.8	60.3	0.3	35166	01-01-2018 00:00	11-03-2019 02:05	434	1	5.377534	52.16848
205	14.6	14.9	41.7	-1.9	7.2	51.2	0.2	18233	09-05-2018 13:32	31-12-2019 23:59	601	369	5.403155	52.17515
250	12.1	11.1	37.1	-6.8	7.3	53.4	0.3	31846	07-07-2018 20:10	13-08-2019 00:40	401	26	5.391069	52.17823
269	21.1	20.7	40.9	12.7	4.6	21.1	0.3	1013	17-06-2018 13:42	20-07-2018 17:05	33	21	5.403531	52.17507
270	12.8	12.2	41.2	-7.2	7.7	59.1	0.4	34359	09-07-2018 14:01	06-09-2019 04:45	423	4	5.386526	52.16984
272	13.7	13.7	115.8	-26	8	64.7	-0.3	35798	17-06-2018 14:01	31-12-2019 23:29	562	125	5.398406	52.17178
273	11.9	10.9	39.4	-7	7.5	55.7	0.5	29959	01-07-2018 14:18	27-06-2019 22:20	361	0	5.382196	52.16753
274	12.3	11.5	39.3	-7.8	7.7	59.6	0.4	31568	01-07-2018 13:36	03-08-2019 17:54	398	6	5.398862	52.17416
276	12.9	12.6	41.7	-8.1	7.8	61.5	0.3	34132	01-07-2018 13:07	31-12-2019 23:49	548	123	5.391066	52.16706
279	12.3	11.4	41.6	-26	7.4	55.1	0.4	40002	09-07-2018 13:00	31-12-2019 23:44	540	53	5.381558	52.17179
284	13	12.6	40.1	-6.6	7.4	54.9	0.3	36640	07-07-2018 19:02	09-09-2019 04:18	428	0	5.39033	52.16915
288	12.1	11.5	39.7	-7.7	7.2	51.9	0.4	42319	07-07-2018 20:03	31-12-2019 23:56	542	28	5.392847	52.16951
289	12.8	12.3	38.4	-6.1	7.3	53.4	0.3	35282	09-07-2018 13:29	22-08-2019 20:10	409	0	5.3837	52.17002
290	12.7	12.1	42.1	-26	7.7	59.5	0.4	38377	09-07-2018 07:45	31-12-2019 23:44	540	87	5.383709	52.1739
300	13.2	12.9	42.7	-7.2	7.5	57	0.4	38133	09-07-2018 13:24	02-10-2019 13:06	449	4	5.384125	52.17279
KNMI-De Bilt	11.3	10.8	37.2	-8.4	7.1	50.8	0.2		01-01-2018 00:00	31-12-2019 23:59				

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

In the further global analysis of the data, a scripting environment for Python, named Jupyter notebook, is used. The analysis consists of the inspection of:

- histograms
- auto correlation
- cross correlation
- plots of temperature vs. time

Conclusions are reported in this paragraph together with only some visualisations. The data and notebook can be shared upon request.

Histograms for most sensors exhibit a behaviour similar to the histogram for sensor 25 (Figure 3-3):

- two 'peaks', at 5-7 ^{o}C and at 15-17 ^{o}C. These peaks seem to correspond to frequent temperatures during the respectively winter and summer period.
- 'shallow' valley in between the two peaks
- slightly right skewed//left tail shorter than right tail (corresponds to mostly positive skewness in Table 7)

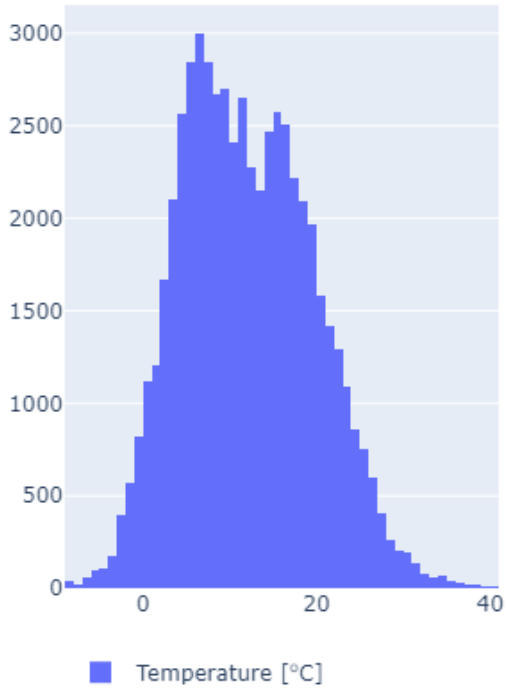

Figure 3-3. Histogram for raw temperature data at sensor 25

The correctness of a temperature measurement could be derived from prior measurements; a temperature of 20 ^{O}C seems odd if the temperature 15 minutes ago at the same location was measured to be 10 ^{O}C. To train an algorithm that flags outliers and other non-NaN anomalies, information is needed on which previous measurements could be used to determine if the current measurement is correct. This information is explained with autocorrelation (see Annex 1).

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

With autocorrelation, it is determined how data points in a time series correlate to data points with the same (delayed) index in the same time series. A timeseries is thus correlated with its copy. If no changes are made to the copy, the correlation value (usually the Pearson correlation) is equal to 1. If the indices of the copy are shifted one index position to the left or right, or by one *lag*, the correlation value describes how well data points are correlated with the previous or following data point. Furthermore, recurring signals could be observed with autocorrelation in data that is linked to recurring events. An example is temperature data. As the data is linked to a day-night cycle, high autocorrelation values are expected if the used lag corresponds to a 24 hour-shift of data points.

A requirement for autocorrelation is that data points are equally distributed over time. Therefore, the raw datasets were formalised to datasets with values per 15 minutes. Each value is the average of raw values within this 15-minute window.

The correctness of a temperature measurement can also be evaluated with averages taken over larger windows. Besides autocorrelation on the formalised 15-minute temperature averages, autocorrelation was also carried out on datasets resampled to:

- hourly-averaged values
- daily-averaged values
- weekly-averaged values
- monthly-averaged values

Autocorrelation coefficients for a multitude of lags were calculated for each resampled dataset of temperature data for each sensor. The autocorrelation plots (Figure 3-4) have on the y-axis the correlation value and the lag on the x-axis. As said before, each lag represents a shift in indices of the data points, so if data points are gathered each 15 minutes, a lag of 1 signifies that a data point is compared to a data point of 15 minutes ago, and likewise a lag of 10 signifies that a data point is compared to a data point of 150 minutes ago.

The following observations are made:

- a daily seasonality in Figure 3-4a) and b), as local maxima occur at lags representing a multitude of 24 hours (for example, 96 lags for 15-minute-averaged values and 24 lags for hourly)
- a yearly seasonality in Figure 3-4d) and e), as local maxima occur at lags representing one year (52 lags for weekly-averaged values and 12 lags for monthly)
- rapidly changing correlation coefficients for sensors with data gaps or data for a limited period
- a bandwidth within which correlation coefficient values are described for most sensors
- sensors of which correlation coefficients are not contained within the bandwidth, often have limited data (sensors 13 (orange), 174 (yellow), 269 (light blue), and 71 (green))

The last three observations inform us that dropping temperature data measured by some sensors from the training data set, might improve the quality of the trained algorithm if results should disappoint.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

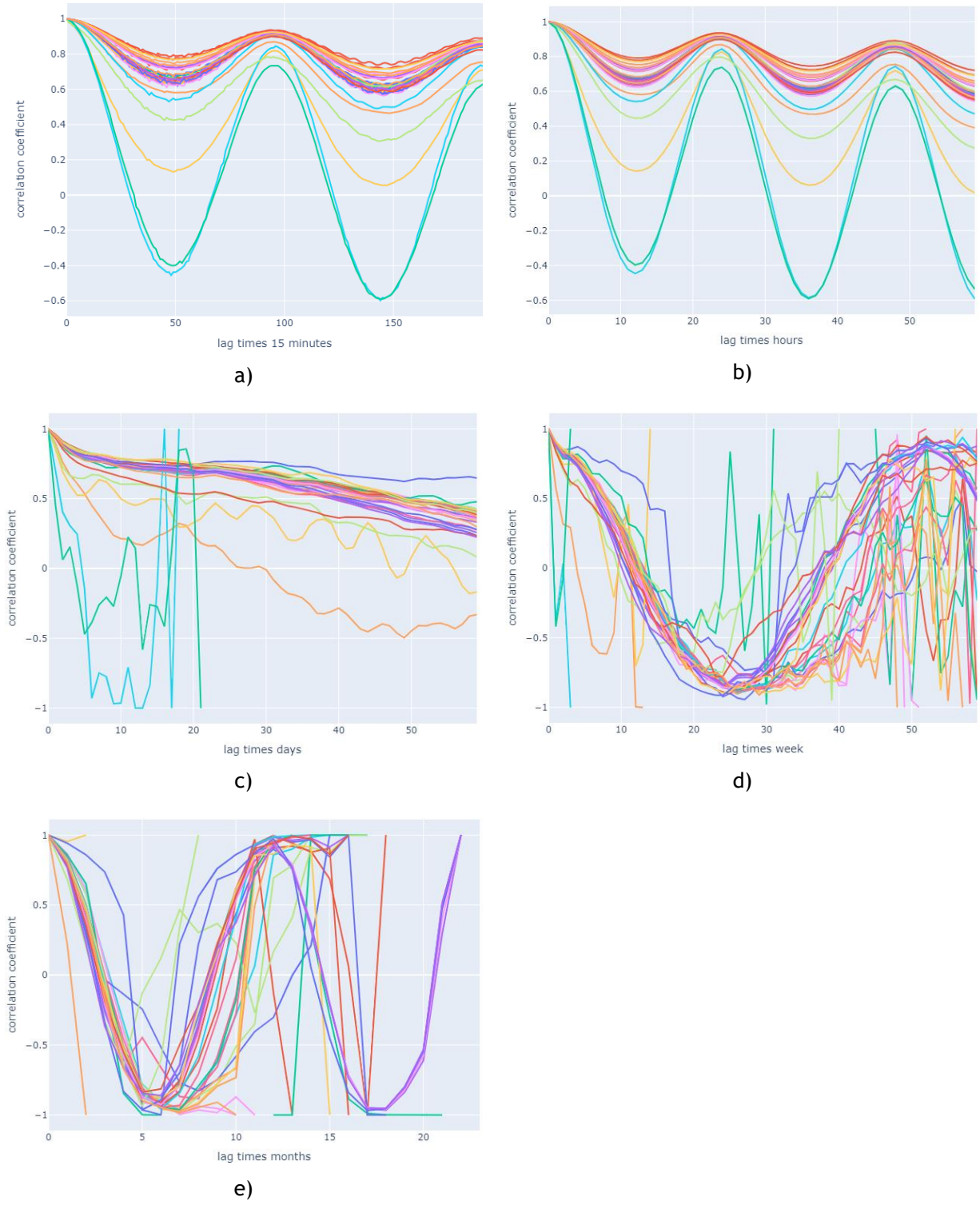

Figure 3-4. Autocorrelation coefficients for all sensors for temperature averaged over different intervals, namely a) 15 minutes, b) hours, c) days, d) weeks, and e) months. Lags taken correspond to the aggregation period, i.e. for daily data is lag=1 equal to a lag of 1 day, lag-2 equal to 2 days, etc.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The correctness of a temperature measurement could not only be evaluated with previous observations made by the same sensor, but also with temperature measurements made at the same time by other sensors. Cross-correlation of temperature data measured by different sensors provides information on which other sensors could be used for the evaluation of correctness of measurements of a certain sensor.

With cross-correlation (see Annex 1), it is a measure of similarity of two series as a function of the displacement of one relative to the other. The correlation value (usually the Pearson correlation) is equal to 1 if one timeseries has the same data point pattern as the time series with which it is correlated. N.B.: data points do not need to have the same values at the same indices, only the overall pattern has to be the same. If data of two sensors are well-correlated and a relatively high temperature is measured at a certain time by a sensor, the temperature measurement at the other sensor is also expected to be relatively high. If this is not the case, it might be a reason to doubt the correctness of the measurement done by either sensor.

For the cross-correlation analysis, the 15-minute averaged dataset was used, that was originally made for the auto correlation analysis. The reason is that data timestamps are originally not set at a regular intervals and cross correlation cannot be carried out if corresponding data points do not share the same timestamp.

Figure 3-5 the cross-correlation values pairwise between sensors in a cross-correlation matrix. The cross-correlation value between one sensor and the other is the same as the cross-correlation value between the latter and the former. This allows the matrix to be simplified to a triangular matrix. If one needs to know the cross-correlation value between temperatures measured by sensor 182 and another sensor, one either looks in the row or the column headed by ‘sensor182’, depending on where a pair is made between sensor 182 and the sensor of interest. For example, the cross-correlation between sensors 182 and 139 is 0.99 and is found at the intersection of row ‘sensor182’ and column ‘sensor 139’. The cross-correlation value between a sensor and its copy is equal 1 and is thus left out of the matrix.

The following observations are made:

- Almost all sensors have a high cross-correlation coefficient with all sensors. Measurements made by a certain sensor could thus be checked and, if needed, corrected with the use of temperature data measured by other sensors.
- Sensor 13 is highly correlated with only some sensors (182, 25, 33, 43, 49, 69 and 76)
- Some sensors have a slightly less strong correlation with most sensors ($r<0.98$), being sensors:
    - 174
    - 269
    - 272
    - 274
    - 290
    - 33
    - 63
    - 71

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

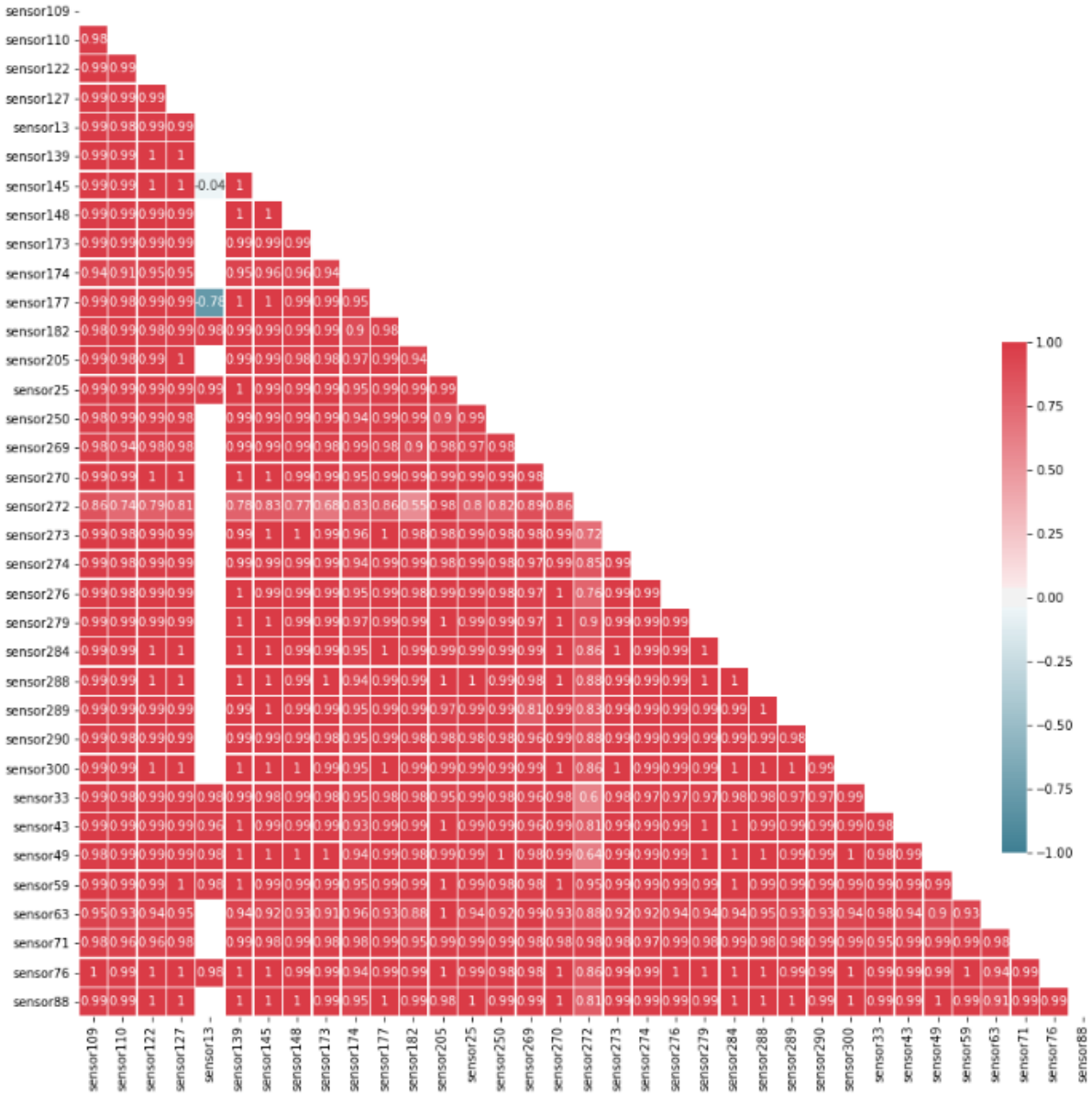

Figure 3-5. Cross-correlation of 15-minute averaged temperature data

By plotting temperature against time only, and not datetime, values at the same time of day can be compared. This provides the option to create bandwidths for certain periods of time with which new temperature values can be tested for being an anomaly. Figure 3-6 presents the raw and validated temperature data at sensor 43, plotted against time of day. From this kind of graphs, temperature bandwidths and temperature trends during the day could be quantified. A strength is that outliers, such as depicted in the figure, are detected immediately. Also, similar daily temperature trends could be distinguished; at most sensors, temperatures rise and fall in a sinusoid fashion during the day, but for some sensors (110, 139, 145, 148, 205, 272, 276, and 288) a slow rise and a steep fall are observed. One should keep in mind that in Figure 3-6 no distinction is made in seasonal variation and that the bandwidth is thus very broad. Nonetheless, significant outliers can still be detected as seen in the raw data figure around 22:00. Bandwidths per day of the year could be plotted in the future as timeseries for multiple years are then available

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

A weakness is, however, that the upper side of the bandwidth is likely to increase further in coming years due to climate change and the more frequent occurrence of extremely high temperatures. An algorithm that focuses on temperature bandwidths should thus be able to keep learning from new (validated) data.

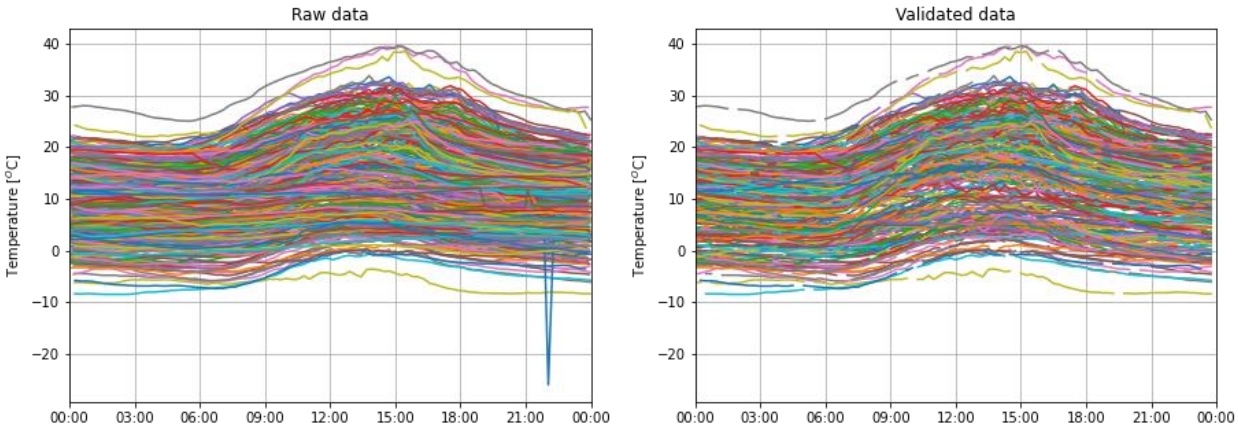

Figure 3-6. Temperature data measured by sensor 43 plotted against time

Data source 3 - City of Amersfoort

The data is acquired from a contact person of the city of Amersfoort (COA). The data was collected during a pilot in which air quality was measured. Professional sensors were used and that these have been installed by professionals, and thus it is assumed that the data contains a relatively small number of anomalies caused by human or measuring errors. The assumption on data quality cannot be tested as at the time of writing no validated dataset is available.

Data was gathered for 5 sensors from the COA database between November 1st, 2018 and November 30th, 2019 (full length of the project). Below, a summary is given of the data:

- Start date: 01-12-2018 02:00
- End date: 30-11-2019 23:00 (except for one sensor, which gathered data till 21-11-2019 15:00)
- Interval: Hourly
- Sensors: 5, with:
    - 3 located near main roads (sensors 14542, 14544 and 14577, with sensor 14544 having more surrounding vegetation))
    - 2 located in the middle of neighbourhoods (sensors 14522 and 14531)
    - all sensors were placed on streetlights, which always lid by the sun during the day
- Quality: good (assumed)

Figure 3-7 shows temperature and humidity measurements of one of the sensors, sensor 14531. In July 2019, a possible anomaly is observed, with unexpected continuous relatively low summer temperatures. Upon closer observation, it is concluded that this ‘anomaly’ is caused by data gaps; the measurements were only taken at certain hours of the day (instead of all hours), which were mostly night-time hours. This explains why only low temperatures are present for this period in the dataset. Similar ‘anomalies’ for other periods are seen at sensors 14542 and 14544.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

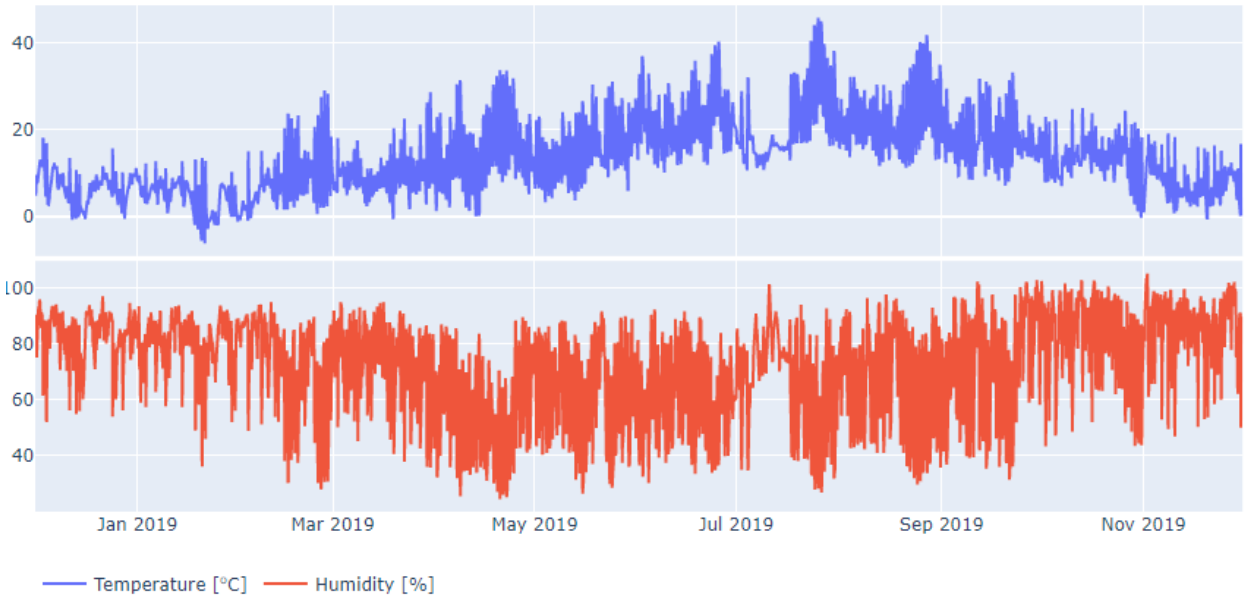

Figure 3-7. Temperature and humidity data as acquired for sensor 14531

Only the temperature data was interesting for determining heat risk. A global statistical analysis of the temperature data was done for all sensors. Table 8 is the result. The following conclusions could be drawn from the global statistical analysis:

- sensors 14577 has the least amount of data gaps when compared to the recording period (8719 observations with 0 days missing).
- median and average temperatures are not as well aligned as was the case with the MJS sensors; average temperatures at all COA sensors are at least 1 °C higher than the median temperatures.
- with all maximum temperatures being above 44 °C (and the maxima of the MJS sensors), it could be assumed that the quality of measured extreme temperatures is dubious. An explanation for the high temperatures could be that the air temperature measured inside the measurement stations was hotter than the actual air temperature, because all sensors are in full sunlight during the whole day. Sensor 14544 has relatively more surrounding vegetation when compared to the surroundings of the other sensors, which might explain the measured maximum and the minimum temperatures where respectively relatively lower and higher than at other sensors that are more in the open
- when compared to the statistical values for temperatures measured by the KNMI at station De Bilt between December, 2018 and November, 2019, COA sensors seem to have consistently measured higher temperatures (higher average, median, and minimum temperatures), but also higher and more frequent high temperatures (higher maximum temperatures and higher standard deviation, variance, and skewness values). The same is true to a less extent when temperature statistics of COA sensors are compared to temperature statistics of MJS sensors.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 8. COA sensor statistics with statistics of temperatures measured at the KNMI measurement station De Bilt

Temperature							
Sensor	Average	Median	Max	Min	Standard deviation	Variance	Skewness
14522	13.5	12.5	47.2	-5.3	8.0	63.6	0.7
14531	13.3	12.1	45.8	-6.2	8.4	70.7	0.7
14542	13.3	12.3	47.9	-6.3	8.2	67.3	0.7
14544	13.6	12.4	44.7	-5.7	8.3	68.8	0.6
14577	14.3	13.0	51.5	-5.2	8.6	74.3	0.8
KNMI-De Bilt	11.2	10.7	37.2	-7.8	6.6	44.1	0.3

General							
Sensor	Number Of Observations	Start Date	End Date	Total Days	Days Without Observations	Longitude	Latitude
14522	8494	01-12-2018 02:00	21-11-2019 15:00	355	0	5.391	52.15
14531	8457	01-12-2018 02:00	30-11-2019 23:00	364	0	5.405	52.167
14542	8292	01-12-2018 02:00	30-11-2019 23:00	364	6	5.405	52.193
14544	8326	01-12-2018 02:00	30-11-2019 23:00	364	3	5.416	52.161
14577	8719	01-12-2018 02:00	30-11-2019 23:00	364	0	5.403	52.158

In the further global analysis of the data, a jupyter notebook is used. The analysis consists of the inspection of:

- histograms
- auto correlation
- cross correlation
- plots of temperature vs. time

Conclusions are reported in this paragraph together with only some visualisations. The data and notebook can be shared upon request.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Histograms for most sensors exhibit a behaviour similar to the histogram for sensor 14577 (Figure 3-8):

- a peak at 5-7 ^{O}C
- an edge right of the peak at 15-17 ^{O}C
- right skewed//left tail shorter than right tail (corresponds to mostly positive skewness in Table 7)
- noteworthy peak at the 17-18 ^{O}C bin for sensor 14542

The histograms for COA data are more normal-distributed than they are for MJS data, with respect to the presence of a single peak. Because of this and the right-hand tail reaching higher positive values than in the histograms for the MJS sensor data, the skewness is more pronounced in the COA histograms.

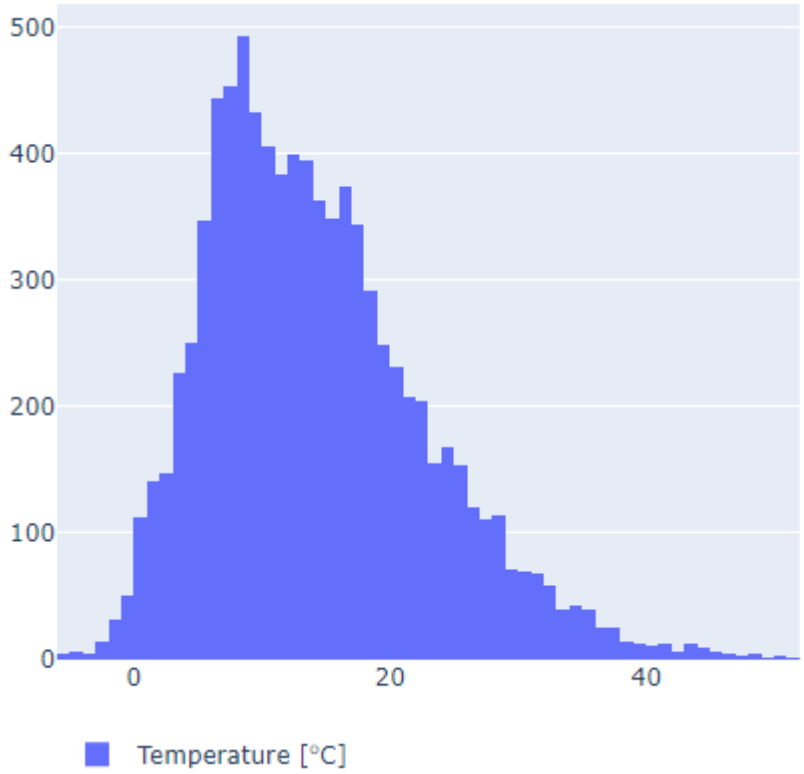

Figure 3-8. Histogram for raw temperature data at sensor 14577

Autocorrelation was carried out for the COA data too. The how and why of autocorrelation is explained in section of MJS data exploration. Also, the correctness of a COA temperature measurement can be evaluated with averages of larger windows. Besides autocorrelation on raw dataset with hourly observations, autocorrelation was also carried out on datasets resampled to:

- daily-averaged values
- weekly-averaged values
- monthly-averaged values

Autocorrelation coefficients were calculated for each dataset. Figure 3-9 presents the results. The following observations are made:

- a daily seasonality in Figure 3-9a) and b), as local maxima occur at lags representing a multitude of 24 hours (for example 96 lags for 15-minute average values and 24 lags for hourly)
- a yearly seasonality in Figure 3-9 c) and d), as local maxima occur at lags representing 24 hours (52 lags for weekly average values and 12 lags for monthly)

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

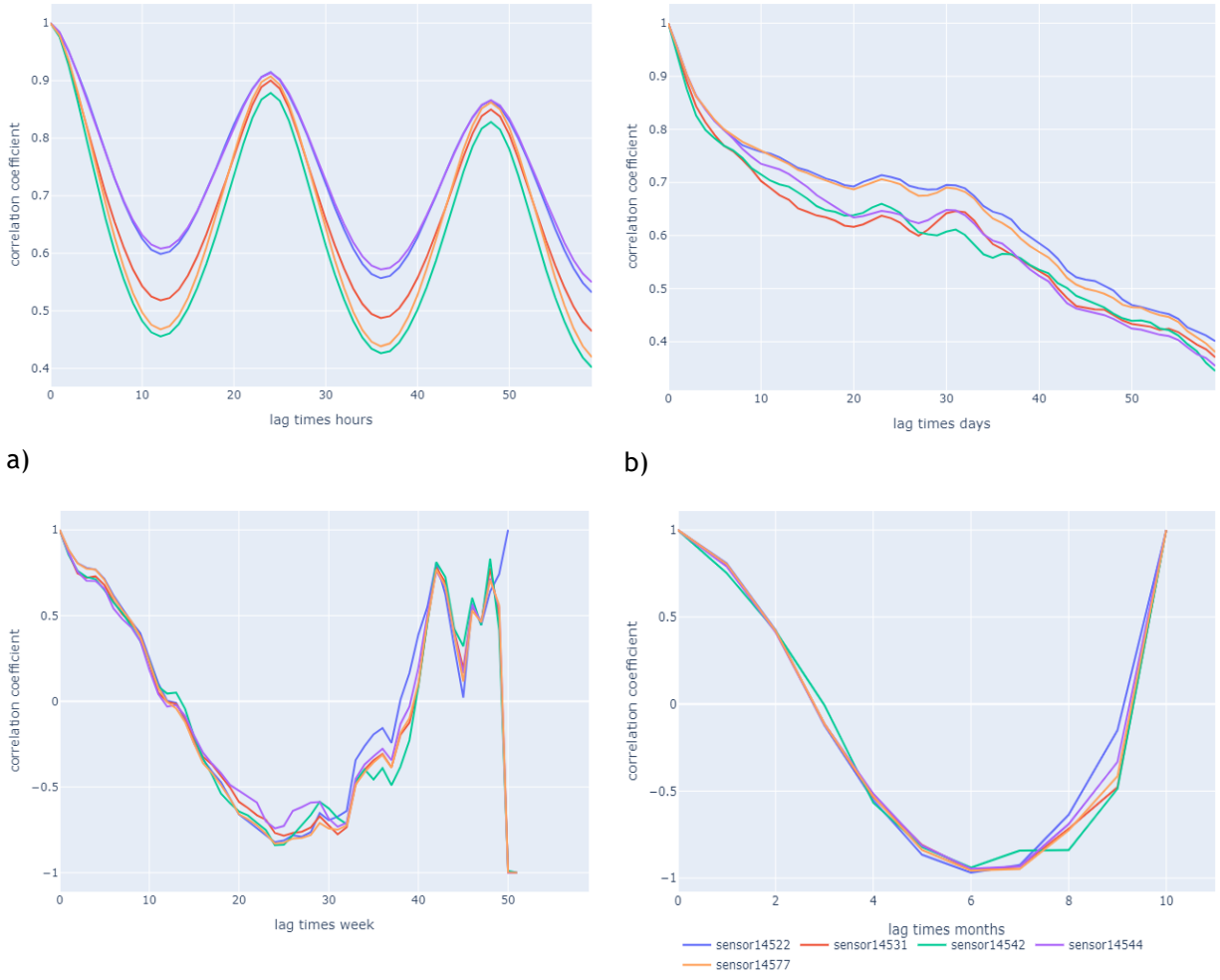

Figure 3-9. Autocorrelation coefficients for all sensors for temperature over different intervals, namely a) hours, b) days (averaged), c) weeks (averaged), and d) months (averaged). Lags taken correspond to the aggregation period, i.e. for daily data is lag=1 equal to a lag of 1 day, lag-2 equal to 2 days, etc.

For the cross-correlation analysis, the hourly values were used. Correlation values are presented in Figure 3-10 in a similar fashion as was done in the MJS data exploration (Figure 3-4):

- all sensors have a high cross-correlation coefficient with all sensors. Measurements made by a certain sensor could thus be checked and, if needed, corrected with the use of temperature data measured by other sensors
- sensors located near main roads are as well cross-correlated with sensors in neighbourhoods as with other sensors near main roads, and vice versa

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

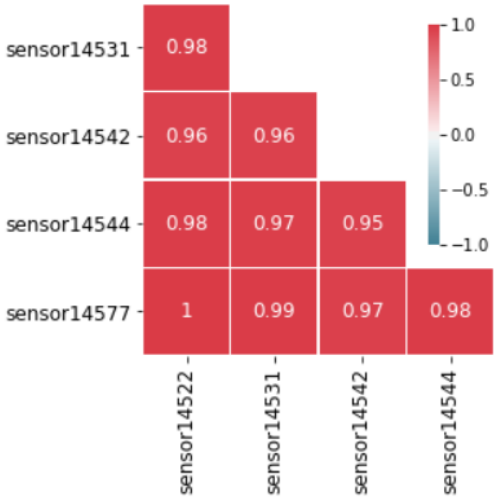

Figure 3-10. Cross-correlation of the temperature data

By plotting temperature against time only, and not datetime, values at the same time of day can be compared. This provides the option to create bandwidths for certain periods of time with which new temperature values can be tested for being an anomaly. Figure 3-11 presents the temperatures measured by sensor 14577, plotted against time of day. Sensor 14522 shares a similar pattern with temperatures slowly building towards 18:00, and then steeply falling. For the other sensors, a less shared, but other pattern is observed with a steep rise in temperatures at 9:00 and a fall at 18:00. This explains the high cross-correlation coefficient value for sensors 14522 and 14577 (see Figure 3-10).

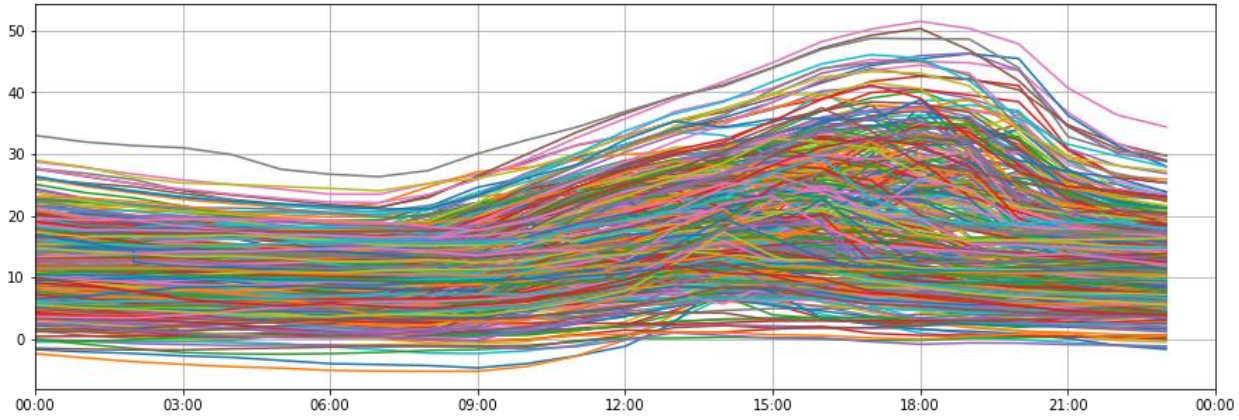

Figure 3-11. Temperature data measured by sensor 14577 plotted against time

3.2.1.3. DATA PREPARATION

It was decided to select the MJS data for further use and to not use the CoA data in the first iteration of the Heat-risk case. The reason for the second decision was the unavailability of validated CoA data. Of course, a validated dataset could be constructed following the same procedure as was used by the MJS organisation. However, the MJS organisation do have exact knowledge of how sensors are configured and thus has a better judgement of what records could be anomalies and which could not be anomalies. We lack this knowledge to do the same for the CoA data.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The main part of the raw data cleaning was appropriating the raw data for the comparison with the validated data. The data was cleaned by:

- resampling the records of the raw dataset with the same interval as is present in the validated dataset, being 15 minutes. The value of the records was determined by averaging the observed temperatures till the new timestamp plus 15 minutes
- removing records from the validated dataset that satisfied the requisites below:
    - contained missing values
    - corresponding record in raw dataset also contained missing values
    - corresponding record in raw dataset is only preceded or followed by records with missing values till respectively January 1st, 2018 or October 19th, 2018
- clipping the raw dataset to the temporal extent of the validated dataset, being January 1st, 2018 till October 19th, 2018, or to a smaller extent if the validated dataset has valid records for a shorter period

From the cleaned raw dataset, a feature dataset with 573875 records was constructed that contains the following 18 features:

Table 9. Data model used to learn – Non-nan anomaly detection in temperature data

Feature	Description	Type
Sensor	This feature is dropped in modelling, but is included in the feature dataset for evaluation purposes	Integer
DayOfYear	Day number derived from the record's timestamp	Integer
HourOfDay	Hour derived from the record's timestamp	Integer
QuarterOfHour	Quarter of the hour derived from the record's timestamp	Integer
Longitude	Longitudinal coordinate of sensor location	Float
Latitude	Latitudinal coordinate of sensor location	Float
temperature_lag	Temperature measured by the sensor at the record's timestamp (lag=0) or up to 3 timestamps prior (0<lag<=3)	Float
temperature_mean_past_hour_lag	Mean of temperatures measured by the sensor during the hour preceding the record's timestamp (lag=0) or 8 and 16 timestamps prior (corresponding with 2 and 4 hours ago)	Float

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Feature	Description	Type
temperature_mean_past_day	mean of temperatures measured by the sensor during the 24 hours preceding the record's timestamp	Float
temperature_mean_past_week	mean of temperatures measured by the sensor during the week preceding the record's timestamp	Float
temperature_variance_past_hour	variance of temperatures measured by the sensor during the hour preceding the record's timestamp	Float
temperature_variance_past_day	variance of temperatures measured by the sensor during the day preceding the record's timestamp	Float
temperature_variance_past_week	variance of temperatures measured by the sensor during the week preceding the record's timestamp	Float
anomaly (target feature)	False if temperature measured by the sensor at the record's timestamp in the raw data corresponds with the temperature at the record's timestamp in the validated data, True if otherwise. N.B.: if at the record's timestamp the raw and validated temperature is a NaN-value, the instance is not marked as an anomaly, as the goal of the algorithm is not to detect existing data gaps.	Boolean

Missing values in the constructed feature dataset were inputted by replacing these with means of the features. The imputed values are marked by a mask. The mask is a Boolean array with the same size as the feature dataset, which contains a True value at locations corresponding with the missing value locations in the feature dataset array.

The last transformation of the dataset is a standardised scaling. For each feature, the mean was deducted from the records and the results were divided by the standard deviation of the feature's records.

3.2.1.4. MODELLING & EVALUATION

The AI objective of the data driven classification model to be designed in the current case is to detect and flag outliers and other non-NaN (Not a Number) value anomalies in temperature timeseries. To make sure that not all available data was used for model fitting and calibration and that 'unseen' data was left to test the model with training and test datasets were produced. The test dataset contained samples from September and October 2018. The training set contained samples for the eight preceding months (January to August 2018). Since this is the first iteration, only the training dataset is used to evaluate the performance and potential of the models in this use case. From the training set, smaller subsets were made according to the Time Series split technique (see Annex 3) and were used as training and validation datasets in cross-validation.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Model performances were evaluated using two metrics, being the recall and the precision scores. The former is the ratio between the number of correctly detected anomalies and the number of the observed anomalies, whereas the latter is the ratio between the number of correctly detected anomalies and the number of the predicted anomalies.

In the first iteration, two classifiers types have been used to build algorithms, being the histogram-based Gradient Boosting classifier (HBGB) and the AdaBoost classifier. The HBGB classifier is different from the regular Gradient Boosting classifier as feature values are first binned before the classifier is fit by the training set (Sci-kit learn, 2019). Because a separate bin is reserved for missing values, the HBGB classifier can cope with missing values in the input data set, whereas the regular Gradient Boosting classifier could not.

Table 10 presents the evaluation metric scores for both used classifiers. These scores are the weighted averages across the cross-validation iterations, with weights being proportional to the number of samples in the training set per cross-validation iteration. The HBGB classifier did to some extent correctly detect only anomalies (precision), but was even worse at detecting all anomalies (recall). The AdaBoost classifier only detected non-anomalies, thus the precision score was zero and the recall did not provide a number (division by zero).

Both classifiers do not meet the objective of 80% accuracy in anomaly detection.

Table 10. Evaluation metric scores for used classifiers

Algorithm	Recall Score	Precision Score
HBGBC	0.04	0.20
AdaBoost	0.0	n/a

Observing these results, the design of a new data model will be a first solution to improve the ability of detecting non-NaN temperature values with machine learning, before finetuning the used classifiers or using other classifiers or even moving on to model evaluation. The reason being that it was discovered after internal discussions that temporal, but more importantly, spatial information is included in an erroneous way in the data model. The data model could be reformed in the first iteration to include temporal information in a better way, but not the same could be said for the spatial information.

Rather than using coordinates, closely correlated sensors will be 'linked' directly by including temperature values of neighbouring sensors in a dataset entry. Since the measurements of almost all sensors are quite well correlated with each other (see Figure 3-10), the four closest sensors are selected to be included in a sample (given of course that they are well correlated (R>0.98) with the sensor of interest). Table 11 presents the design of the data model that will be used in the second iteration.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 11. Data model used to learn in second iteration – Non-NaN anomaly detection in temperature data

Feature	Description	Type
Sensor	This feature is dropped in modelling, but is included in the feature dataset for evaluation purposes	Integer
temperature_lag	Temperature measured by the sensor at the record's timestamp (lag=0) or up to 3 timestamps prior (0<lag<=3)	Float
temperature_mean_past_hour_lag	Mean of temperatures measured by the sensor during the hour preceding the record's timestamp (lag=0) or 8 and 16 timestamps prior (corresponding with 2 and 4 hours ago)	Float
temperature_mean_past_day	mean of temperatures measured by the sensor during the 24 hours preceding the record's timestamp	Float
temperature_variance_past_week	variance of temperatures measured by the sensor during the week preceding the record's timestamp	Float
neighbour_sensor_#_temperature_lag	Temperature measured by the closely correlated neighbouring sensor # (0-3) at the record's timestamp (lag=0) or up to 3 timestamps prior (0<lag<=3)	Float
anomaly (target feature)	False if temperature measured by the sensor at the record's timestamp in the raw data corresponds with the temperature at the record's timestamp in the validated data, True if otherwise. N.B.: if at the record's timestamp the raw and validated temperature is a NaN-value, the instance is not marked as an anomaly, as the goal of the algorithm is not to detect existing data gaps.	Boolean

Besides a new data model, also the performance of the Robust Covariance and Isolation Forest classifiers will be assessed alongside with the performance of the HBGB and AdaBoost classifiers, as they are mentioned in the SciKit-learn documentation on Outlier detection (or non-NaN value anomaly detection).

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

3.3. GROUND WATER / SOIL MOISTURE: OPTIMISATION

A third effect of climate change to the urban environment is drought: long periods without significant precipitation. Urban vegetation (parks, trees), historically adapted to a moderate climate with an average monthly precipitation depth of 60 mm, suffer from these elongated dry periods, leading to limited growth and increased vulnerability to plagues or illnesses. This has a negative effect on the urban livelihood.

COA has several options to mitigate the effects of droughts, e.g. watering trees and parks or changing to other, more drought resistant vegetation. As these measures are expensive, COA wants to investigate an optimal strategy to fight droughts.

3.3.1. ITERATION 1

3.3.1.1. BUSINESS UNDERSTANDING

Trees and other vegetation take up water with their roots from the 'vadose' or unsaturated zone. This is the upper layer of the soil, between the surface level and the ground water table. In the City of Amersfoort and its surroundings, the thickness of the unsaturated zone typically varies between 0.5 - 3 m.

Soil moisture is the water that is contained in the unsaturated zone between the soil particles. The amount of soil moisture available to plants depends on the soil type (e.g. clay, sand, silt), the amount of precipitation and the depth of the groundwater table.

Soil moisture sensors can be very accurate, but have an extremely low geographical reach, due to the heterogeneity of the soil. Typically, one sensor has a reach of 0.1 - 0.2 m. Therefore, several soil moisture sensors are usually deployed in a vertical line, providing a 'soil moisture horizon' between the surface level and the ground water table at one location.

Research in the Netherlands has shown a relatively strong correlation between soil moisture content (observed by a cluster of vertically placed sensors) and the ground water table, especially in rural areas. The latter hydrological variable is much easier to observe and has a wider geographical reach - i.e. a less dense network is required to capture the variations within an area.

The COA and the MJS platform want to investigate the correlation between soil moisture available to vegetation and the ground water table in the urban area of Amersfoort. With this correlation established and quantified, the existing ground water observation network can be used to characterise the vulnerability to drought for different areas within the city (e.g. vulnerable - moderate - robust). This information can serve as a basis for a drought mitigation plan and as a source of information for urban planning.

As a first iteration, a regression model is derived to estimate soil moisture content based on ground water level observations.

The soil moisture network is currently being installed as part of deliverable D4.17. The business case described above will be elaborated in deliverable D2.5.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

4. BARCELONA CASE

4.1. SEDIMENT LEVEL PREDICTION ON SEWAGE SYSTEM

Sewer systems are among the most critical urban infrastructures, suffering from a tremendous variety of problems. Sewer chokes are blockages typically caused by faulty human behaviour such as discharging fats, oils, and wet wipes, which must be recycled, and also produced by natural factors such as tree roots. A blockage may lead to uncontrolled overflows into public or private property.

Sewer systems have old regions, new regions, regions that have difficult access due to structural restrictions, and regions at high risk toward blockages. Not only the sewer conditions themselves are involved in the potential blockages, but also the geographical location, such as a location with a lot of bars and hotels, or a location with a high concentration of trees and plants. To prevent blockages, Water and Sewerage Companies (WaSCs) carry out inspections into the sewerage system, but because of the high monetary cost, there is a huge interval between inspections. Furthermore, to identify each blockage, the number of inspections needed in the sewer system is high, causing an inevitable risk of blockages.

4.1.1. ITERATION 1

4.1.1.1. BUSINESS UNDERSTANDING

To reduce the risk of blockages, the city council of Barcelona follows maintenance and cleaning routines over all the sewage system in the city. One of the key factors to decide if a cleaning is needed is the sediment level, which is the accumulation of sand, oils, fats, or other objects that may obstruct the sewer.

The city council spends a lot of money in each maintenance done, not only on the sewer cleaning but on the revisions. The business objective is to reduce the number of revisions needed by predicting the actual sediment level in a section, so the manager can decide which sections have preference of being reviewed and how fast it needs to be done.

The AI objective is to predict the height of sediment level in a concrete section, obtaining another indicator to decide if the section should have maintenance. The approach will focus on spatial prediction, using the physical properties and sediment levels of the nearest sections to predict the actual level of the section to evaluate.

In this study, the dataset used contains information about a sewer grid in Barcelona, in the neighbourhood of Poblenou. The data contains the physical properties of different sections in the sewer and historical sediment levels extracted during maintenance routines.

There are a couple of criteria to achieve a successful model, one being low error in our predictions when evaluating the model and the second to keep predicting well in the future, having a model with good generalization when different data is used.

4.1.1.2. DATA UNDERSTANDING

To understand the sewer data, an exploratory data analysis was done with fixed guidelines. At the beginning of an exploration, it is important to identify which are the variables available and the type of each one to decide which are the analytics that can be applied. The merged dataset contains physical information about each section in the sewer grid and sediment levels, resulting in historical data of 23 columns and 2452 entries. At a first glance, the low amount of entries indicate the models created are not good enough, so the team decided to analyse and extract features that could assure having better models while including new entries, but also have a model that provides competent predictions at the beginning.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The datasets received until now come from different files sent by email. Table 12 identifies each datasource and addresses some details about the location, Table 13 describes each file and provides the length and format of the file, and Table 14 gives more detail on each feature in the dataset.

Table 12. Details about data sources – Sediment level prediction on sewage system

Datasource	Location	Method used to acquire	Problems
Node	Local directory	Received by email	-
Section	Local directory	Received by email	-
Odour problems	Local directory	Received by email	Only 16 registers, small quantity which cannot be analysed
Sediment measures	Local directory	Received by email	-
Materials catalogue	Local directory	Received by email	-
Property details	Local directory	Received by email	-
Section type	Local directory	Received by email	-

As explained in Table 12, the files have been received by email. There is no problem with these files since the training of the models can be done with batch data and in the future when the model gets deployed, the registers can be delivered in real time one by one.

Table 13 shows low registers in all the datasets, and only 2453 registers as sediment measures. The team would like to include more registers in the future to secure the model generalization.

Table 13. General details about available data sources – Sediment level prediction on sewage system

Data Source	Description	Format	# Registers	# Feature
Node	Structural union of different sections and structural change of a section	CSV	405	4
Section	Portion of the sewer between two nodes	CSV	444	10
Odour problems	Point where an issue was recorded	CSV	16	3
Sediment measures	Historical measures done in a section	CSV	2453	5
Materials catalogue	Materials that compose a section	CSV	38	2
Property details	Not all the sewer in Barcelona pertains to the city council. This file points out each manager for a section.	CSV	63	2
Section type	Dimensions and format of the sewer	CSV	2093	9

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 14 shows that most of the features are text, which may be categorical or just explanatory. Explanatory variables should be converted into categorical, and the categorical variables right now should be transformed to a correct format for the ML algorithms. The other variables are numbers, which should be analysed during the next steps.

Table 14. General details about available fields – Sediment level prediction on sewage system

Field	Description	Type	UoM	Data Source
Id	Identifier of the register	Numerical	-	Section
Section	Text identifier of the section	Alphanumerical	-	Section
Length	Section length	Numerical	m	Section
Material	Material identifier of the section walls	Numerical	-	Section
Velocity	Residual water velocity	Numerical	m/s	Section
Water height	Height occupied by the water	Numerical	m	Section
Flow	Flow of the water	Numerical	m^3/s	Section
% Occupied	Percentage occupied by the water	Numerical	%	Section
Special property	Indicates an important property of the section	Categorical	-	Section
Id	Register identifier	Numerical	-	Odour problems
Date	Date of gathering	Date	DD/MM/YYYY	Odour problems
description	Citizen explanation	Textual	-	Odour problems
id	Register identifier	Numerical	-	Sediment measures
parameter	Type of sediment measured	Categorical	-	Sediment measures
Value	In the cases a measure is needed, this is the value of it	Numerical	Depends on the sediment, mostly is meters	Sediment measures
Element id	The section identifier of the measures	Numerical	-	Sediment measures
Maintenance date	Date of the gathering	Date	dd/mm/yy	Sediment measures

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Field	Description	Type	UoM	Data Source
Id	Identifier of the register	Numerical	-	Node
Street quota	The street absolute quota or the projection of the street surface	Numerical	m	Node
Sewer quota	Absolute quota of the sewer bottom	Numerical	m	Node
Node type	Physical element of the node	Categorical	-	Node
code	Material code	Numerical	-	Materials catalogue
Concept	Material description	Textual	-	Materials catalogue
Code	Register identifier	Numerical	-	Property details
Concept	Property manager	Textual	-	Property details
Code	Type identifier	Alphanumerical	-	Section type
Section size	Section area	Numerical	dm^2	Section type
Height	Section height	Numerical	m	Section type
Width	Section width	Numerical	m	Section type
Bucket width	Width of the bucket in the sewer	Numerical	m	Section type
Bucket depth	Depth of the bucket in the sewer	Numerical	m	Section type
Contact width	Width of the sewer on the street height.	Numerical	m	Section type
Pavement number	Some zones with	Numerical	-	Section type
Perimeter	Perimeter of the section	Numerical	m	Section type
Typology	Identifier of the typology used	Categorical	-	Section type

The next step is the analysis of basic metrics. A set of calculations are done to analyse the shape of the received data. The mean to understand the arithmetic centre of each variable, the standard deviation (σ) to understand the difference in value between the registers, the quantiles to understand better the shape of the feature distribution and finally the minimum and maximum points to indicate possible outliers.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 15. Statistical basic metrics – Sediment level prediction on sewage system

Feature	Count	Mean	σ (SD)	Min	Q1	Q2	Q3	Max
Section size	2452	88.33	90.77	1.7	28.8	60	131	806
Height	2452	1.008	0.43	0.15	0.6	1	1.4	2.5
Width	2452	0.91	0.56	0.15	0.6	0.7	1.15	4.5
Bucket width	1543	0.76	0.49	0.4	0.4	0.6	1.4	3.65
Contact width	1453	1.0	0.63	0.5	0.6	0.9	1.4	4.5
Bucket depth	2452	0.05	0.14	0	0	0	0.15	1.25
Perimeter	2452	33.05	16.67	5	19	31	44	124
Length	2452	17.26	12.97	0.8	7.15	13	25.7	72.6
Velocity	2384	0.036	0.079	-0.29	0	0.015	0.071	0.32
Water height	2384	0.1	0.06	0	0.057	0.09	0.14	0.29
Flow	2384	0.002	0.007	-0.02	0	0	0.002	0.033
% Occupied	2384	0.11	0.07	0	0.07	0.1	0.13	0.58
Sediment level [cm]	2414	4.11	5.09	-1	1	3	5	60

In Table 15 some first insights can be seen. Some values of sediment level are negative (which cannot happen), velocity and flow also negative values, the section size feature contains a really big maximum value which may be related with the big perimeter value and length, and also the maximum sediment level is really far from the Q3, which may also be related to these high sizes.

Graphical Univariate Analysis came afterwards. While the previous step gave us the ability to imagine how the distribution is with just a couple of values, this step objective is to exhaustively explore the different variables.

The team has extracted valuable conclusions of these steps. First, the flow registers should always be positive, but the different values could be negative, as shown in the Figure 4-1.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

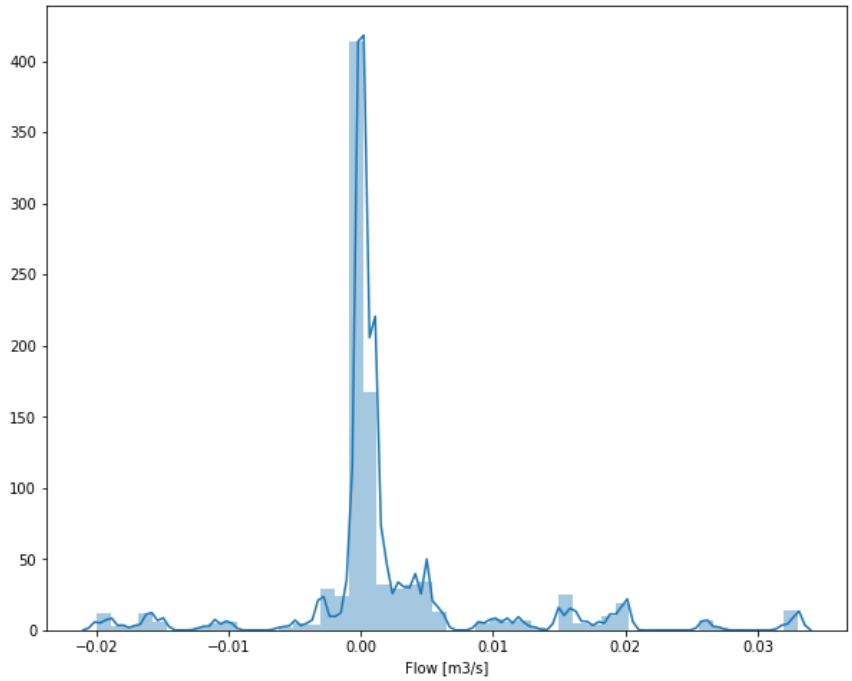

Figure 4-1. Data distribution of flow levels of the different sections – Sediment level prediction on sewage system

The team was able to notice a negative flow, which was then notified to the domain experts. The flow cannot be negative, but it indicates a direction. One of the transformations to the data model was to convert all the flow recordings to positive values.

The sediment level was also analysed, and since it was our objective variable, the information extracted was valuable. The Figure 4-2 shows the distribution of the different sediment levels, having most of its values between 0 and 10, with a couple of registers being higher.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

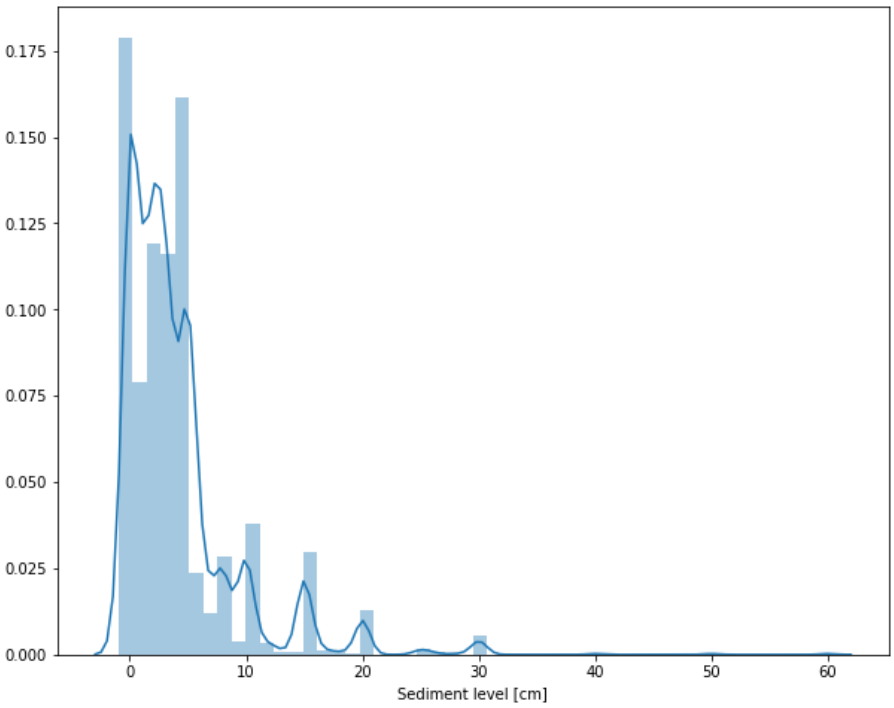

Figure 4-2. Data distribution of sediment level of each section- Sediment level prediction on sewage system

The team noticed that the higher quantiles of the distribution were not homogeneous, for example, value 20 and 30 had some registers, but the number of registers between these two values were too low. This meant the data gathering could have had problems. After checking with domain experts, it was confirmed that the extraction of the sediment level was done by a worker using a basic tool.

An important part of the sediments is which type of sediments can get accumulated into the sewer.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

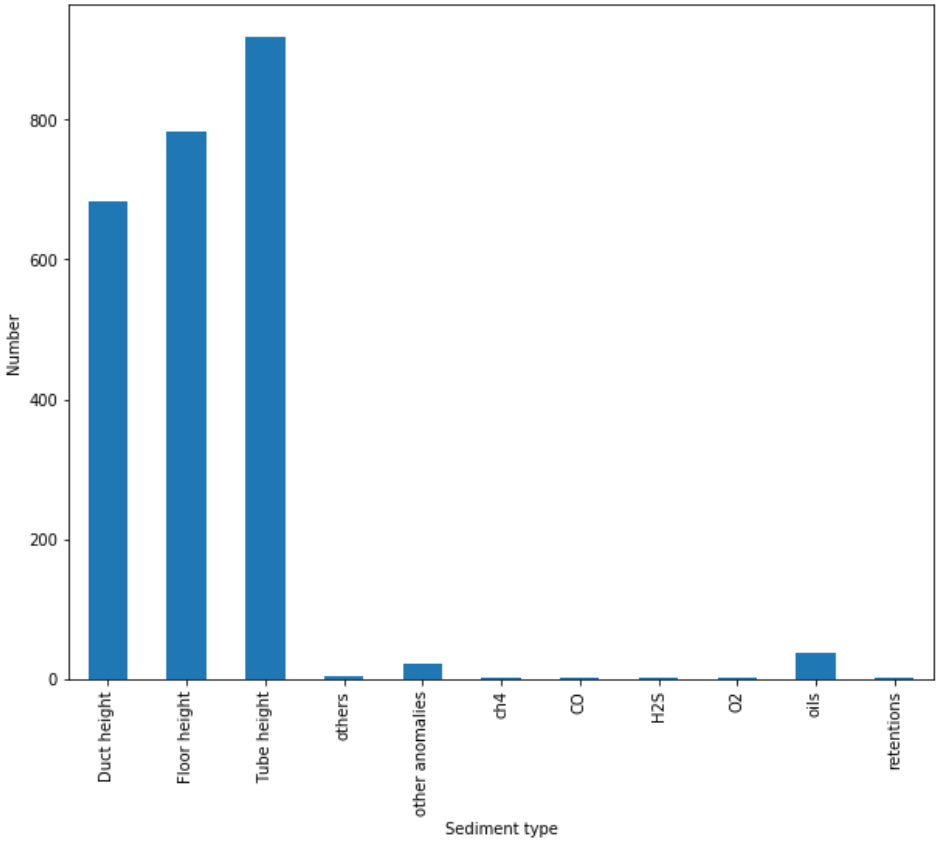

Figure 4-3. Types of sedimentation – Sediment level prediction on sewage system

The majority of sediments comes from 3 different types (see Figure 4-3), which is a domain aspect they are the same, accumulation of sand. Usually, the different types of sedimentation will affect each other, so the team is going to consider all sediment types equal for the first iteration.

The number of sediment level gathered for each section is variable, the team needed to know which is the range of registers to build a data model that can represent the trend and relations between the sediment in the near section. The Figure 4-4 shows the shape of the variable data.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

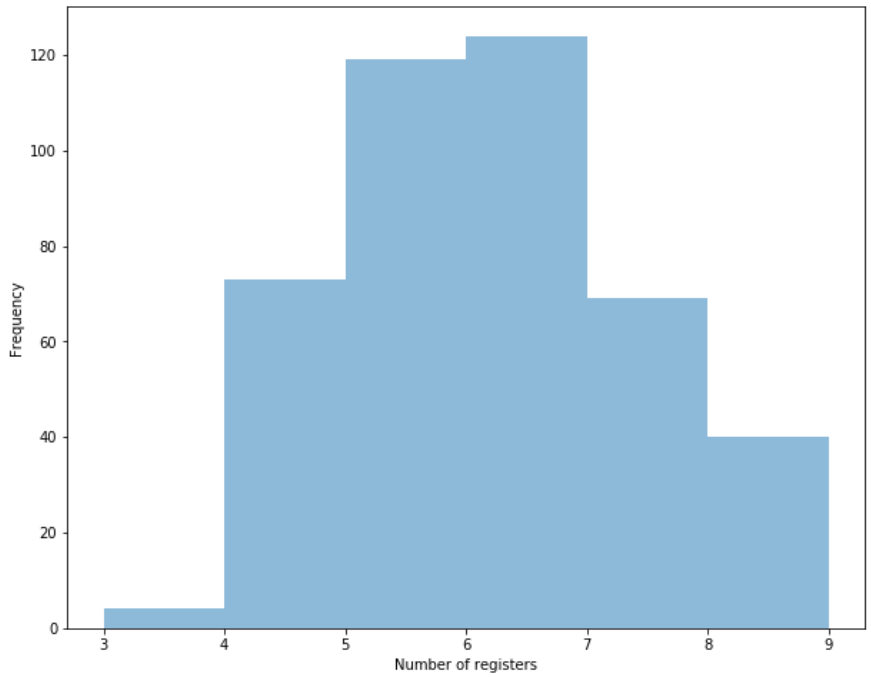

Figure 4-4. Number of sediment level gathered – Sediment level prediction on sewage system

As can be seen in the Figure 4-4, most sections contain 6 registers, but there are some with only 3 and some others with 8. When constructing the data-driven model, the team needs to have in mind this difference in each section.

After univariate analysis, the multivariate analysis comes to solve the hypothesis about the dataset and help understand the relation of the different variables. The team focused on identifying the main point during this step: to analyse which is the sediment growth (the sediment level difference between two registers) taking into account other possible variables.

The first hypothesis is that sediment level is dependent of different section sizes, different slopes and different flows. The section size plays an important role since the sediment accumulation can change given different sizes. Figure 4-5 demonstrates that the section size does not impact the growth of sediment directly. For example, some small-sized sections have bigger or lower sediment growth, while the majority have neutral growth.

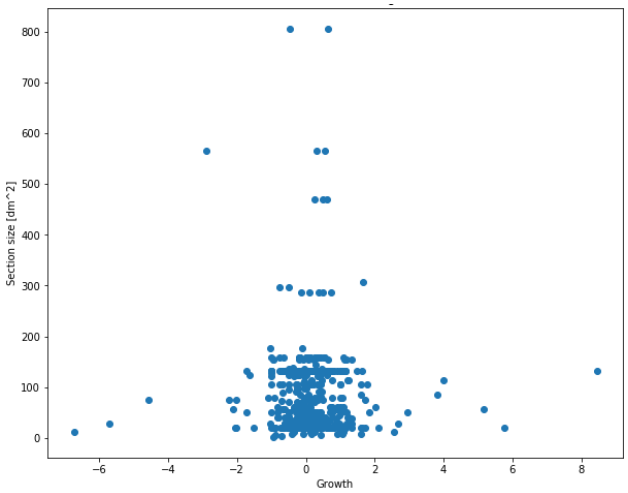

Figure 4-5. Growth in centimetres of the sediment given a section size – Sediment level prediction on sewage system

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The slope is represented by the mean water velocity of the residual water in a section. Let us see the relationship between the velocity and the sediment growth. Again, the influence of the water velocity over the sediment growth is not important as shown Figure 4-6. There are some sections with a lot of sediment growth that is nor affected by the velocity and the high velocity does not affect the sediment growth at all.

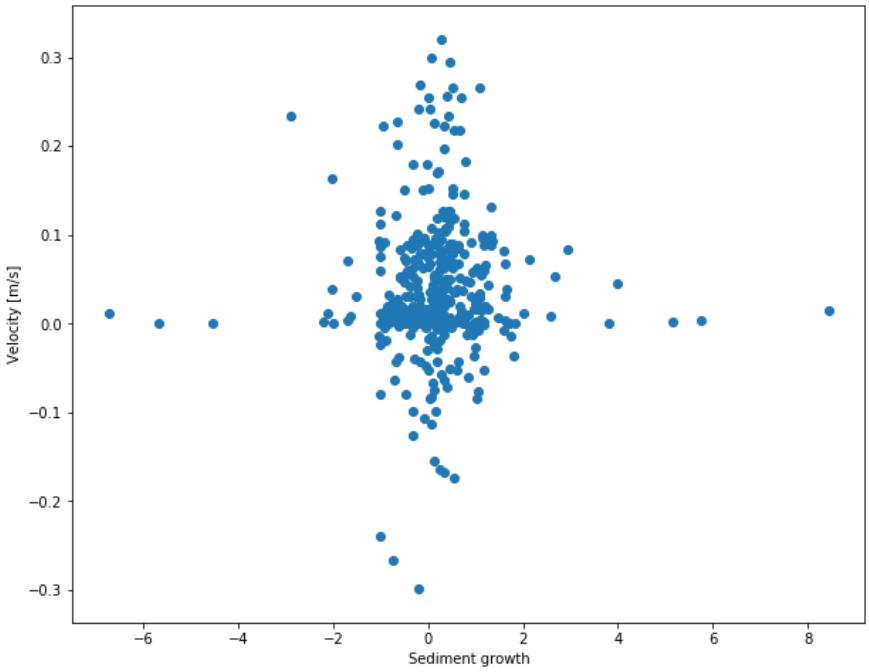

Figure 4-6. Growth of sediment in centimetres given a mean velocity in a section – Sediment level prediction on sewage system

A new hypothesis is that the passing of time affects sediment growth. A direct comparison between the sediment growth and the number of days since the last inspection was done in a section to validate the possible direct relation. Figure 4-7 shows almost no direct relationship between the time and growth since the growth is almost 0 in a lot of points. One important aspect is the big amount of points in the 0 point of the x-axis. When cleaning is applied, the sediment before and after the cleaning is gathered and the days since the last inspection is 0 between those two registers, so they should be ignored in the analysis. It is important to say that maybe the growth is 0 when more than 200 days have passed because some raining has happened between or other events that the team does not know in this iteration.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

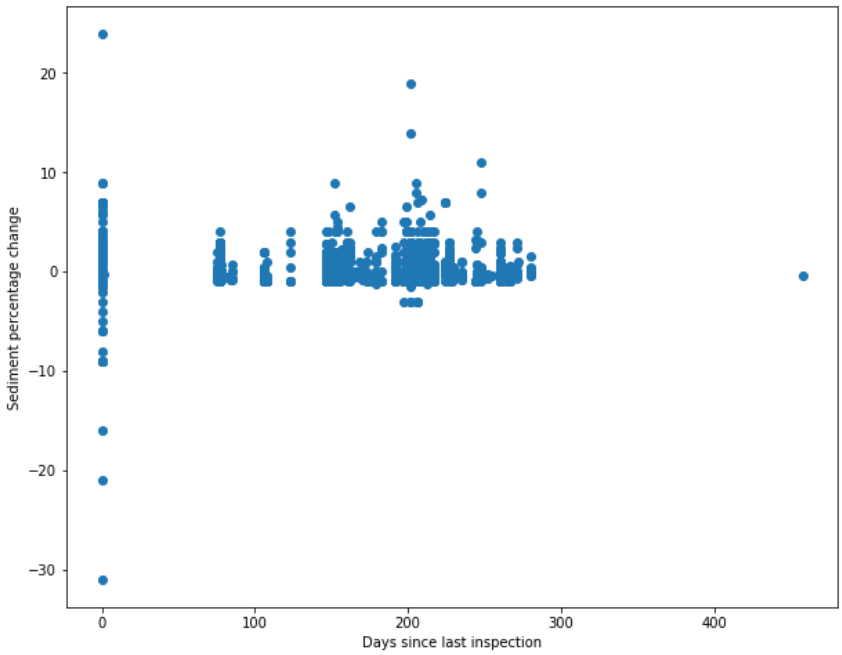

Figure 4-7. Sediment growth since the days of inspection – Sediment level prediction on sewage system

Finally, it is important to say that the impact of raining periods is not included in this first iteration, meaning in following iterations will be important to add the dry-rain periods and the intensity of rain.

Since the data available had spatial properties, it was important to also analyse in a map view which is the hot points and the similarities between near points. To encounter similarities between sections, different metrics were analysed from a spatial perspective.

The proportion of the section occupied by the sediment was one of the statistics to consider. The figure below shows the maximum proportion occupied in each section, showing some clusters with the same proportion and some points alone.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

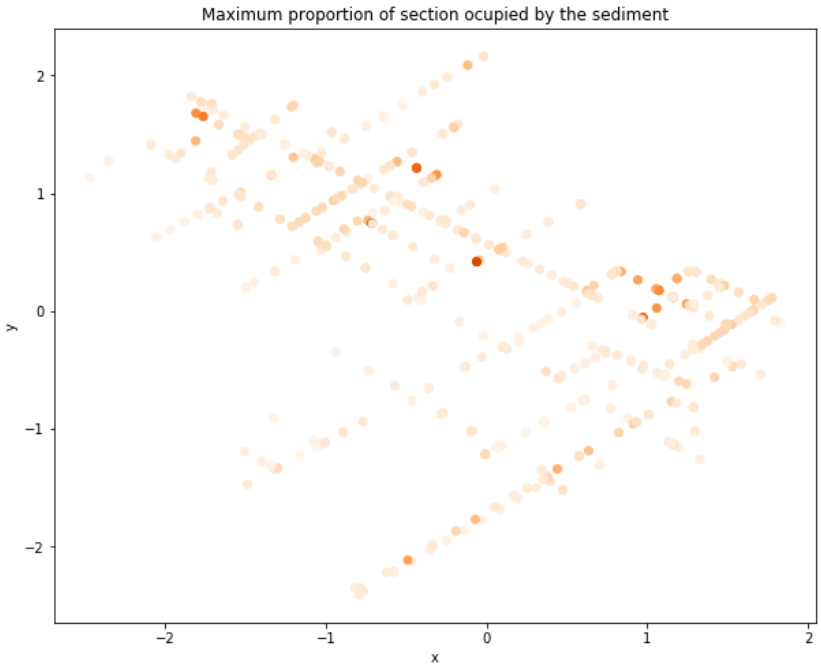

Figure 4-8. Two-dimension map of a neighbourhood in Barcelona. Maximum percentage occupied by sediments on each section, being red the maximum coverage – Sediment level prediction on sewage system

Some sections may have a different percentage occupied but the same sediment level. The size of the section plays an important role, not only in this case but also when measuring the sediments, the physical properties affect the sediment level in a domain aspect.

The mean proportion occupied of a section is also calculated and shown in the Figure 4-9 giving similar results as the maximum proportion.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

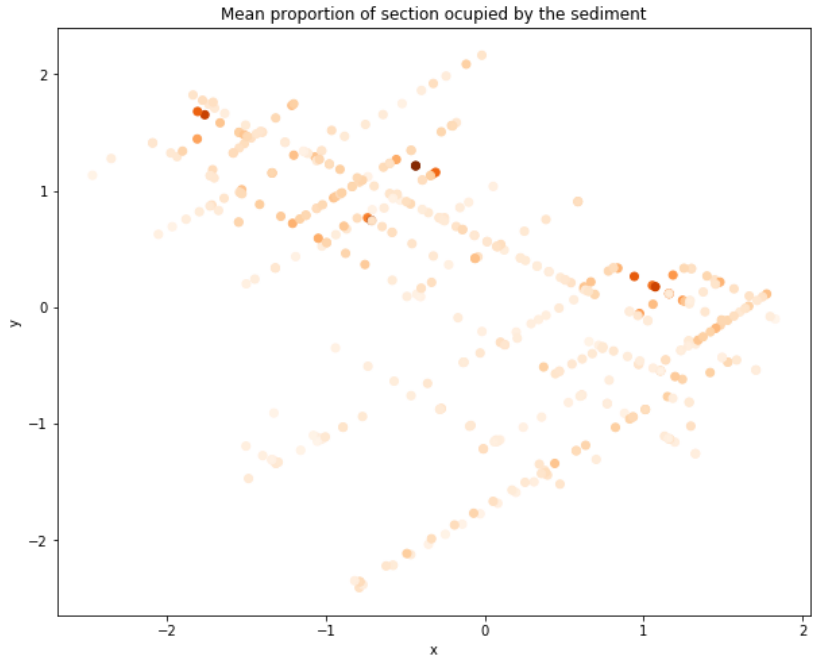

Figure 4-9. Mean percentage occupied by sediment of each section, red being the bigger mean – Sediment level prediction on sewage system

It is interesting to analyse the sediment level and see which patterns it adopts. The maximum sediment level is shown in the Figure 4-10.

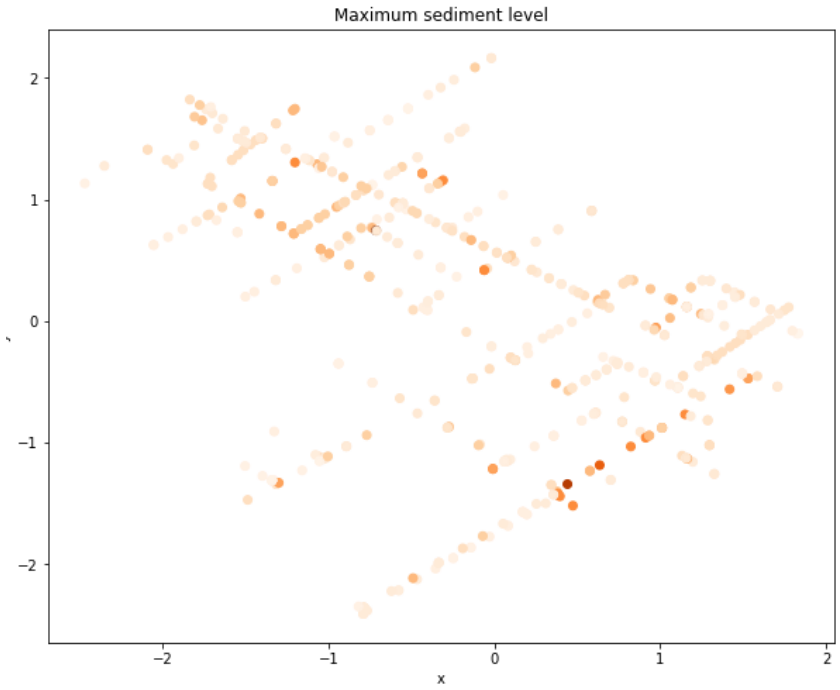

Figure 4-10. Maximum sediment level in each section, red being the bigger maximum value – Sediment level prediction on sewage system

The clusters encountered before do not appear in this new calculation, but a lot of the section interconnected in a straight line shows the same pattern, meaning the sediment accumulation may have a cascade effect. To validate it the mean sediment level needs to be calculated. The Figure 4-11 shows it.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

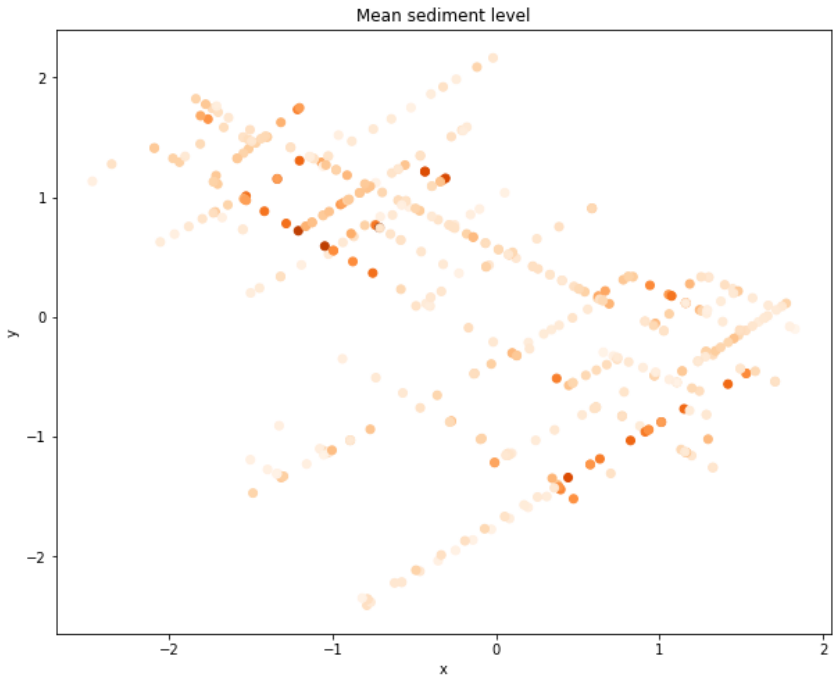

Figure 4-11. Mean sediment level in each section, red being the bigger mean – Sediment level prediction on sewage system

The mean metric confirms the team hypotheses, most of the similar points are next to each other, showing the cascade effect.

The final step of all the process is the correlation analysis (see Annex 1) between the different features. The following feature shows the relation between the features of the same section.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

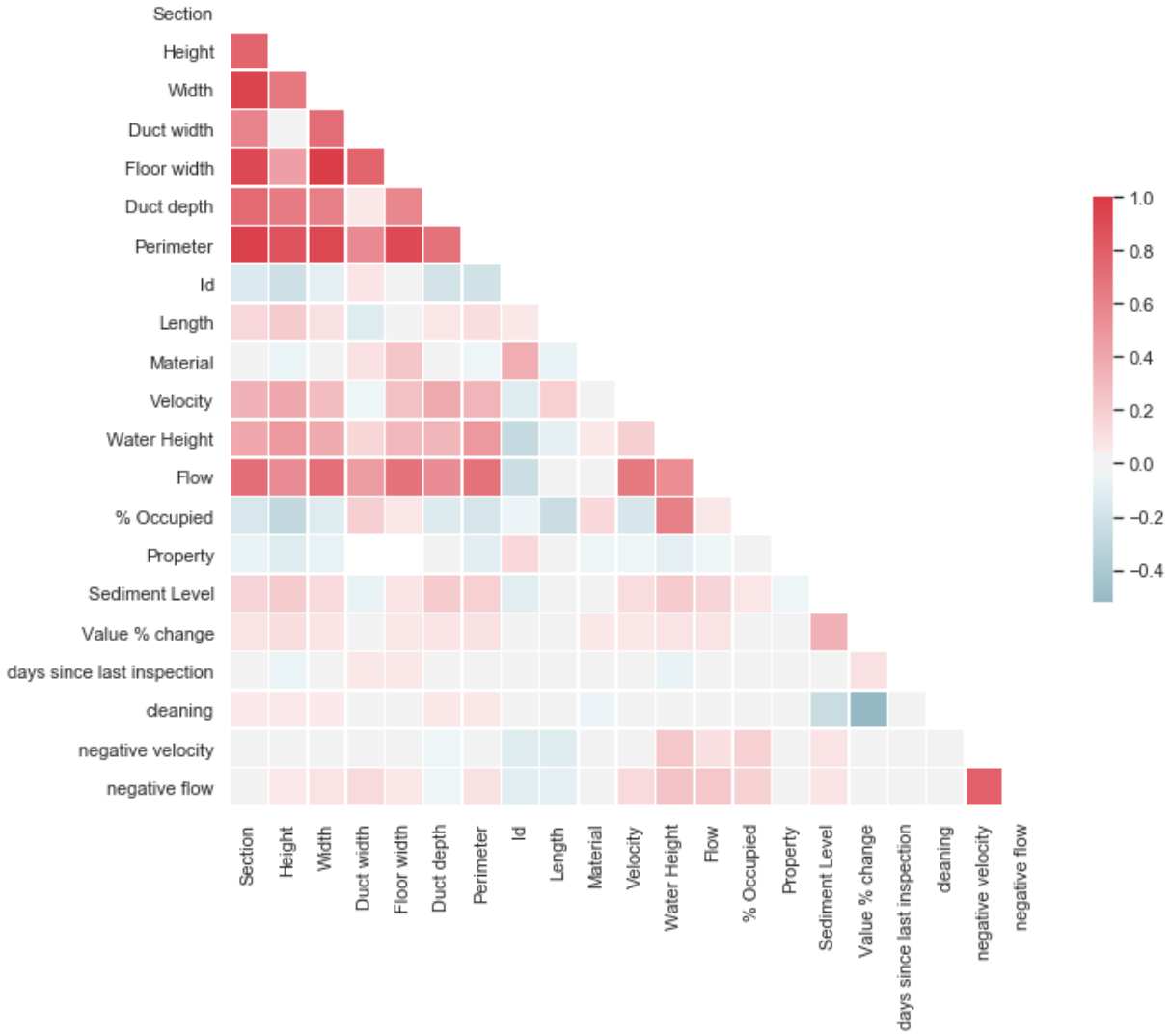

Figure 4-12. Correlation matrix between features of a section – Sediment level prediction on sewage system

As can be seen, there is a high correlation between the physical properties, but low correlation between the sediment level and the other properties. After the data preparation, which is going to be explained in the following section, the team decided to identify the correlation between the different features. Figure 5-13 shows the correlation between the sediment level of the near sections.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

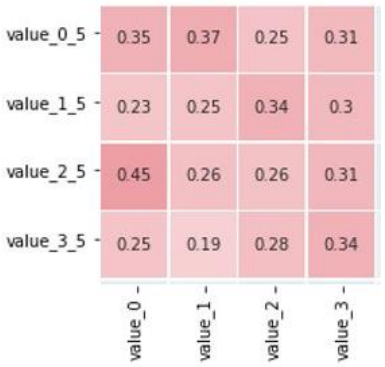

Figure 4-13. Correlation between sediment level of near sections. X labels are 4 historical gatherings of the evaluated section and Y labels are 4 historical gatherings of nearby section number 5 – Sediment level prediction on sewage system

The sediment level has a little bit of correlation with near sections sediment level.

4.1.1.3. DATA PREPARATION

Since the received data was distributed into different CSV and QGIS datafiles, a merging process was needed. Since the sections were separated row by row, the team decided to create a proximity function that calculates which are the similar sections to the one being inspected. The function considers the mean sediment level of the nearer sections and selects the top 5 more similar to the one being evaluated.

Taking into account the previous analysis, the team designed a data model to be fit into a machine learning algorithm. The designed data model contains a set of extracted features, explained in the Table 16.

Table 16. Data frame used to learn – Sediment level prediction on sewage system

Feature	Description	Type
Perimeter	The perimeter of the section	Float
Cubicle width	Width of the sewer section cubicle	Float
Section width	Width of the section	Float
Section height	Height of the section	Float
Mean velocity	Mean velocity of the residual water during the dry season	Float
Mean flow	Mean flow of the residual water during the dry season	Float
Material	The material of the section walls	Categorical
Sediment level lags 0 to 3	Sediment level in 4 different timestamps	Float
Days between maintenances	Days between the sediment level gathering	Integer
Cleaning applied	Indicates if cleaning was applied during the maintenance session	Boolean

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Feature	Description	Type
Size of nearer sections	Size of each of the nearer sections	Float
Mean velocity of nearer sections	Mean velocity of the residual water during the dry season on each of the sections	Float
Sediment level of nearer sections	Sediment level in 4 different timestamps, for each of the sections	Float
Days between maintenances of nearer sections	Days between sediment gathering, for each of the sections	Integer
Cleaning applied of nearer sections	Indicates if cleaning was applied during the maintenance session, for each of the sections	Boolean

Each section contains a different number of historical sediment gatherings. The number of routines oscillates between 3 to 9, so the team decided to use 4 for each section, interpolating the missing one if the length is 3.

The final data model contains a small size of 500 registers and 93 features. This low number of registers is not beneficial for the model, and the trained model could end being under fitted, not understanding the rules behind the dataset. When training and evaluating the models, different feature combinations will be used to study if a reduction in dimensionality produces better results.

The objective variable in the model is the sediment level in timestamp 0, which is the actual sediment in a section.

4.1.1.4. MODELLING & EVALUATION

To predict sediment level in the sewage network, a batch of regressive algorithms, such as *Linear Regression* (*LR*), *Ridge Lasso*, *ElasticNet*, *K-Neighbours Regressor* (*KNR*) and *Gradient Boosting Regressor* (*GBR*), is tested and compared. The algorithms used work well with low quantity of data, so instead of using neural networks, which have high success with a lot of data, more traditional algorithms are going to be used.

The first training iteration has been done without optimizing the hyperparameters, setting the most used configurations for each algorithm and using 70% of data for training and another 30% of the data for the testing. The results are shown in the Table 17. More detailed information about the Scoring metrics on Annex 2.

Table 17. Main results of the initial modelling – Sediment level prediction on sewage system

Algorithm	MAE	MSE	R^2 Score
LR	2.28	14.02	0.38
Ridge	2.3	14.18	0.37
Lasso	2.2	13.10	0.42
ElasticNet	2.23	13.28	0.41
KNR	2.45	17.7	0.21

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Algorithm	MAE	MSE	R^2 Score
GBR	2.35	13.31	0.41

The results show the models have high prediction error. The coefficient of determination (R^2) is low in all cases, being 0.42 the highest when the best value can be 1. The minimum mean absolute error is 2.2 and the minimum mean squared error is 13.10. Considering that the range of sediment level is mostly between 0 to 15, with some registers being bigger with a maximum of 30, we can see that most of our error is coming from the high values. Below, Figure 4-14 shows the comparison between the real and predicted values.

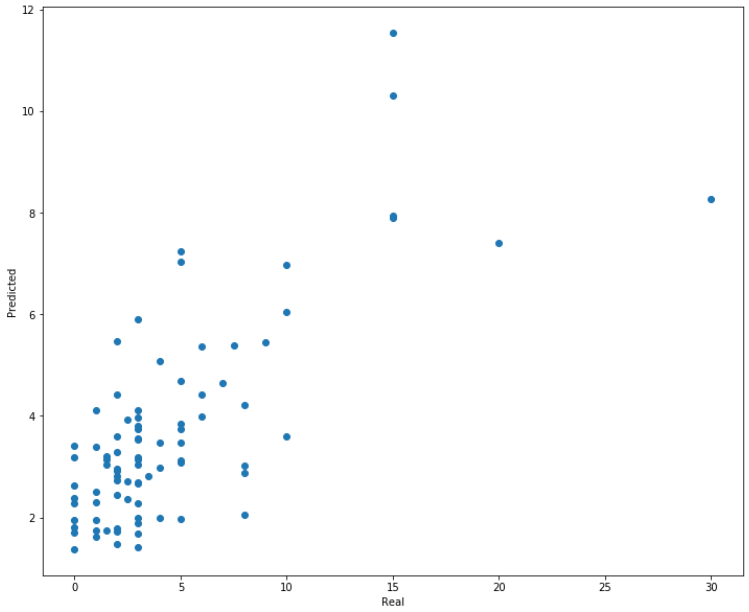

Figure 4-14. Prediction vs Real sediment levels – Sediment level prediction on sewage system

The predicted values are lower than the real values in most cases, being higher the difference if the real values are bigger. Table 18 shows the results after optimizing the hyperparameters of the algorithms that worked better: *Lasso*, *ElasticNet* and *Gradient Boosting*.

Table 18. Main results of the initial modelling after fine-tuning the hyperparameters

Algorithm	MAE	MSE	R^2 Score
Lasso	2.19	13.10	0.4214
ElasticNet	2.21	13.07	0.4228
GradientBoosting Regressor	2.35	13.31	0.4122

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The results after optimizing the hyperparameters got slightly better, showing improvements in the Lasso and *ElasticNet* models. Not only the hyperparameters but also the number of features used was changed. The best option in all cases was to use the physical properties of the predicted section and all the features containing historical sediment levels.

Finally, the *Lasso* and *ElasticNet* models have similar metric values, both will perform equally. The *Lasso* model will predict better the low sediment levels while the *ElasticNet* model will predict better the bigger sediment levels, but the difference in the predictions is not big enough to decide which model is the best.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

5. GOTHENBURG CASE

5.1. EARLY WARNING SYSTEM FOR WATER POLLUTION EVENTS ON CONSTRUCTION

Construction industry is one of the major sources of pollution, responsible for around 4% of particulate emissions, more water pollution incidents than any other industry, and thousands of noise complaints every year (Gray, 2020).

Without careful management, discharge of process water and runoff of stormwater from construction sites may cause significant negative impact on adjacent waterbodies, especially during wet periods. Typical activities which cause water pollutions at construction sites are excavation, drilling, blasting and ground stabilisation with cement (jet-grouting). Other day-to-day activities which involve chemicals such as paint, solvents, fuel and concrete also present further risks.

The most common source of pollution on construction sites is suspended solids. When a construction site removes the topsoil, the remaining surface has no shield or binding element to protect it from rainfall and run-off. With no plants and with the surface compacted by the use of heavy machinery, the rate of run-off increases, and the effect is aggravated. Then, rainy conditions release and move soil particles that become suspended in the surface water reaching water recipients. Additionally, high pH is also a relevant problem on construction sites due to injection of cement during soil stabilisation activities and washing of concrete mixers and tools.

Currently, practical steps are being taken on Gothenburg to minimise such silt and pH pollution by installation of portable and monitored treatment stations. Nevertheless, the application of pre-emptive techniques to anticipate the problems, that is, prepare for the unexpected can be key to face pollution events. In this section, the design of data-driven models to provide an early warning system for water pollution events on construction is addressed.

5.1.1. ITERATION 1

5.1.1.1. BUSINESS UNDERSTANDING

The business objective of this study case is to minimize the impact of pollution events on construction sites. Currently, water quality parameters (conductivity, pH, turbidity and flow) are accessible on-line and real-time through a platform.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

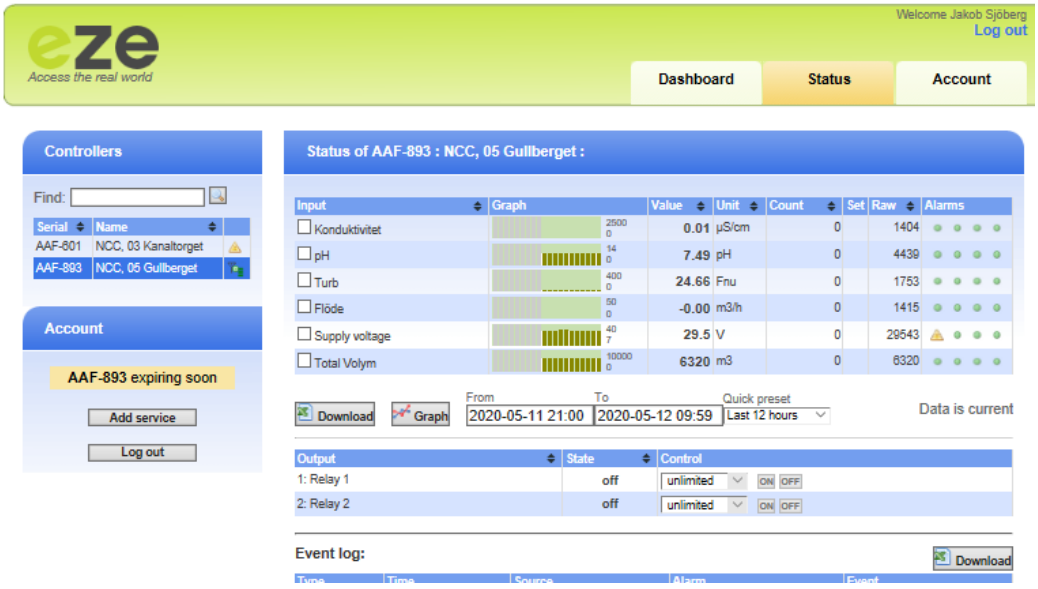

Figure 5-1. Platform available to access water quality parameters

Concerning the AI goal, the aforementioned business goal can be translated to "*Predict a contamination event in advance, taking advantage of historical and real-time water quality parameters (pH, conductivity and turbidity)*".

The most relevant criteria for a successful prediction are to provide a certain level of predictive accuracy and anticipation.

5.1.1.2. DATA UNDERSTANDING

This study case only has one data source, which contains water quality parameters. Data is available through a platform, whose URL cannot be shared in this deliverable by security reasons. Data can be exported on CSV standard. Some problems appeared in the webpage when trying to retrieve large datasets (more than 2 months). Then, all the data were downloaded monthly and later, the data were integrated by using a python script. It is important to note that it is a minor problem because once the data is downloaded, this is no longer necessary again. Below, Table 19 summarizes it.

Table 19. Details about data source acquisition - Early Warning System (EWS) for water pollution events on construction

Data Source	Location	Method used to acquire	Problems
Water quality parameters	n/a (URL cannot be shared)	Download CSV through webpage	The webpage is blocked when trying to retrieve large datasets (2 months)

Additionally, a brief description of each data source is provided, including its format, its number of records and fields.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 20. General details about data sources - Early Warning System (EWS) for water pollution events on construction

Data Source	Description	Format	# Registers	# Fields
Water Quality Parameters	Minute-by-minute information about water quality parameters (conductivity, flow, pH and turbidity), operation parameters (voltage) and alarms	CSV	545701	19

Table 21 enhances the collected information about the data sources, describing the fields part of the data sources.

Table 21. General details about fields - Early Warning System (EWS) for water pollution events on construction

Feature	Description	Type	UoM	Data Source
Time	Date and time of the measurement	Date	YYYY-MM-dd HH:mm:ss	Water Quality Parameters
Conductivity	Instantaneous conductivity measurement	Numerical	µS/cm	Water Quality Parameters
Conductivity count	Not used	Numerical	n/a	Water Quality Parameters
Conductivity Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters
pH	Instantaneous pH measurement	Numerical	pH	Water Quality Parameters
pH Count	Not used	Numerical	n/a	Water Quality Parameters
pH Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters
Turbidity	Instantaneous turbidity measurement	Numerical	FNU	Water Quality Parameters
Turbidity Count	Not used	Numerical	n/a	Water Quality Parameters
Turbidity Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters
Flow	Instantaneous flow measurement	Numerical	m^3/s	Water Quality Parameters
Flow Count	Not used	Numerical	n/a	Water Quality Parameters

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Feature	Description	Type	UoM	Data Source
Flow Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters
Supply Voltage	Instantaneous supply voltage measurement	Numerical	V	Water Quality Parameters
Supply Voltage Count	Not used	Numerical	n/a	Water Quality Parameters
Supply Voltage Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters
Total Volume	Daily total volume	Numerical	m^3	Water Quality Parameters
Total Volume count	Not used	Numerical	n/a	Water Quality Parameters
Total Volume Alarm	Alarm based on a threshold	Categorical (Void/"1")	n/a	Water Quality Parameters

Once identified the data source and their fields, an *Exploratory Data Analysis (EDA*) was done. Table Table 22 presents the analysis of statistical basis metrics. It is important to note that analysis was focused on *Conductivity*, *pH*, *Turbidity*, *Turbidity Alarm* and *Flow* fields due to these fields being strong candidates to be exploited during this study case.

Conductivity feature only contained 109170 registers, five times less compared to pH and turbidity. Additionally, data presented a high spread (σ=114.6) and the maximum and minimum seemed out of range. Initially, the data quality of the conductivity feature is low and should be checked through visual analysis.

pH and turbidity contained 545701 registers. pH shown a low spread, instead turbidity presented a high spread. The presence of outliers was expected due to the minimum and maximum values giving the impression of being out of range. The minimum was negative, and the maximum was very far from the median.

Table 22. General details about features - Early Warning System (EWS) for water pollution events on construction

Feature	Count	Mean	σ (SD)	Min	Q1	Median	Q3	Max
Conductivity	109170	38.8	114.6	-617.9	0.0	4.9	76.7	1472.1
pH	545701	7.1	2.6	-3.4	7.2	7.5	8.1	14.0
Turbidity	545701	-11.3	61.7	-99.2	-25.1	-25.1	22.5	399.0
Turbidity Alarm	4211	1.0	0.0	1.0	1.0	1.0	1.0	1.0
Flow	545701	0.5	2.6	-8.7	-0.04	0.0	0.0	35.1

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Figure 5-2, Figure 5-3, Figure 5-4 and Figure 5-5 present a graphical univariate analysis of turbidity, flow, pH and conductivity measurements, respectively. Flat signal was observed for flow, pH and turbidity. Despite of flatlines would indicate no activity, they do not coincide in time. Additionally, outliers and out of range values such as negative values of flow, pH, turbidity were also observed. Conductivity did not present enough data to be exploited. Finally, the data quality was low for all the features, probably due to poor maintenance of the sensors. The major part of the data cannot be recovered by applying signal conditioning techniques. Nevertheless, despite of aforementioned data quality problems, the data sets are large enough to be split into minor data sets with a minimum of data quality. Then, these subsets could be used to build and assess an initial proof of concept of the data-driven model. It is important to note that it is strongly recommended to gather in a future turbidity and pH data with a better data quality. For that, discussion with the data owners will be taken to see what maintenance that is performed as well as how the data is used and quality assured.

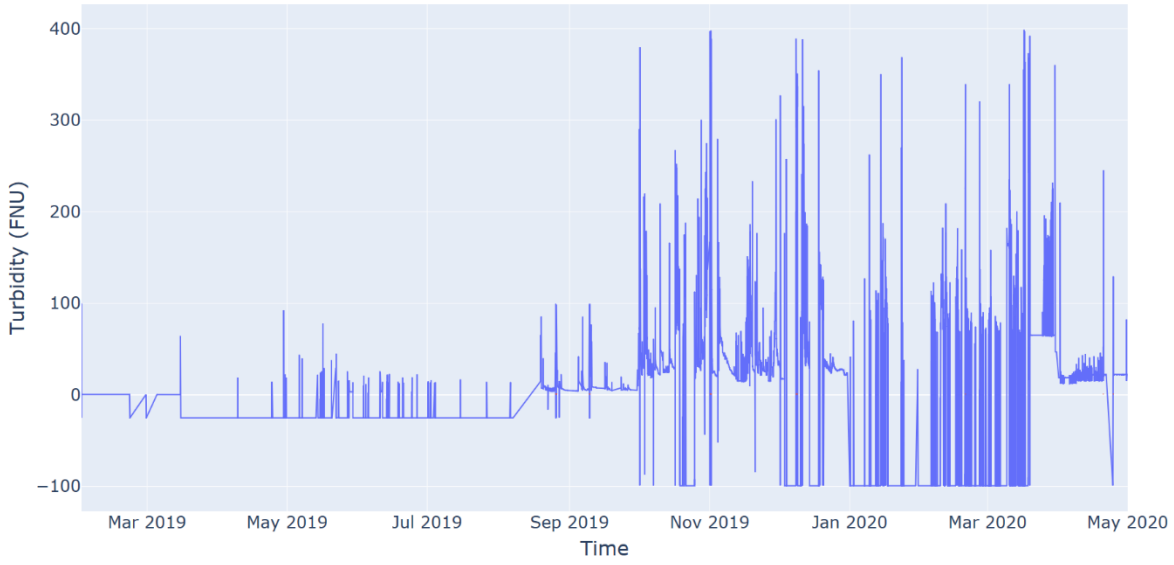

Figure 5-2. Graphical univariate analysis of turbidity - Early Warning System (EWS) for water pollution events on construction

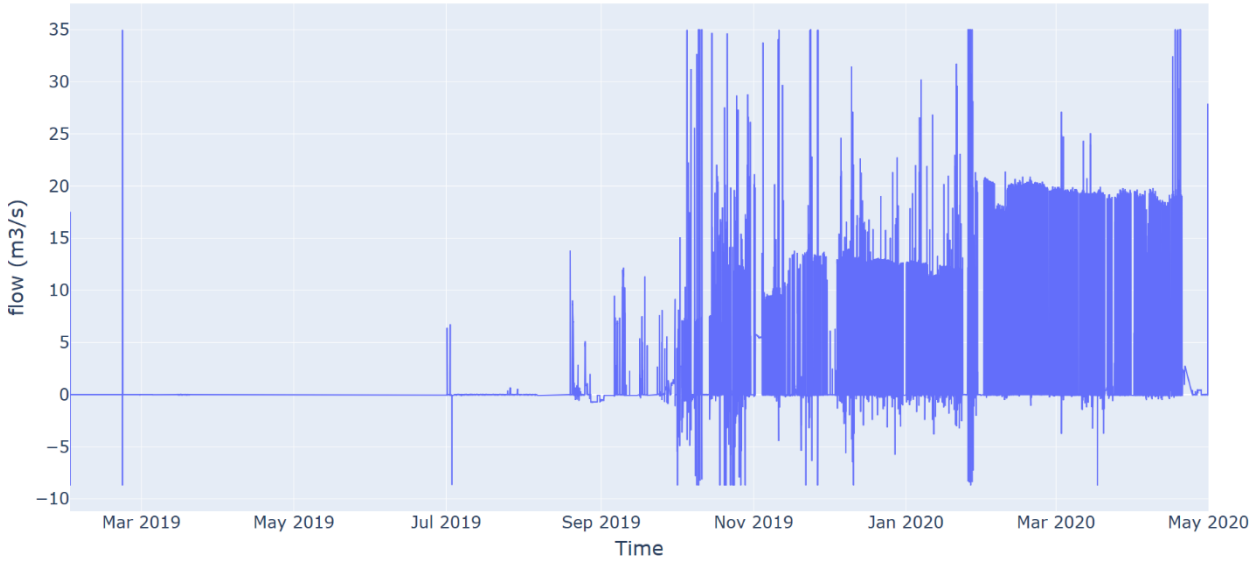

Figure 5-3. Graphical univariate analysis of flow - Early Warning System (EWS) for water pollution events on construction

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

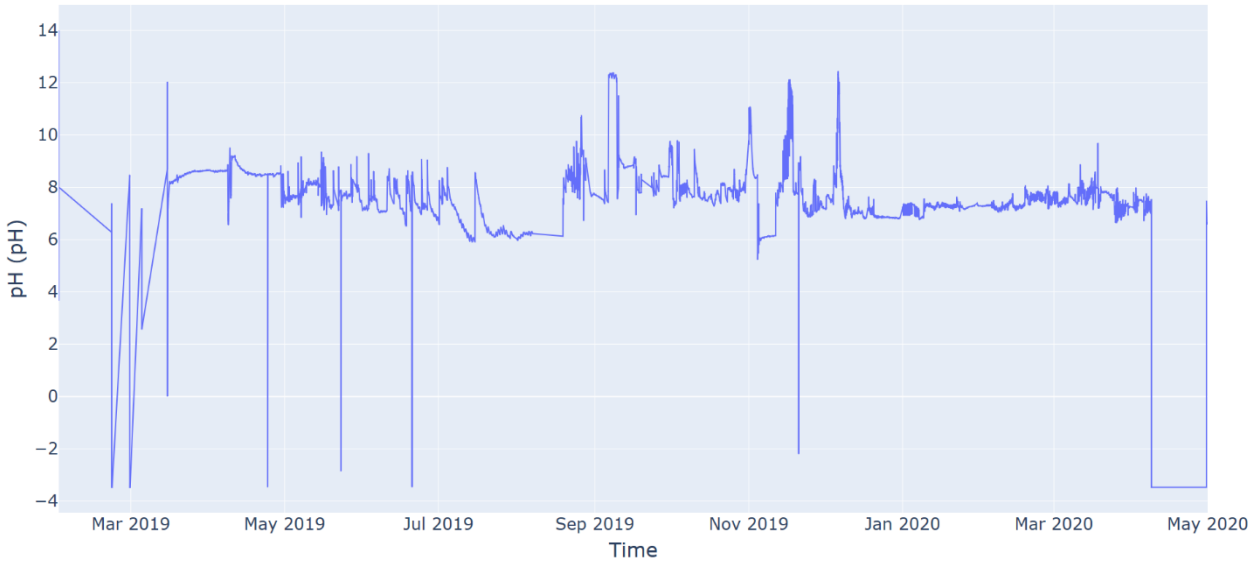

Figure 5-4. Graphical univariate analysis of pH - Early Warning System (EWS) for water pollution events on construction

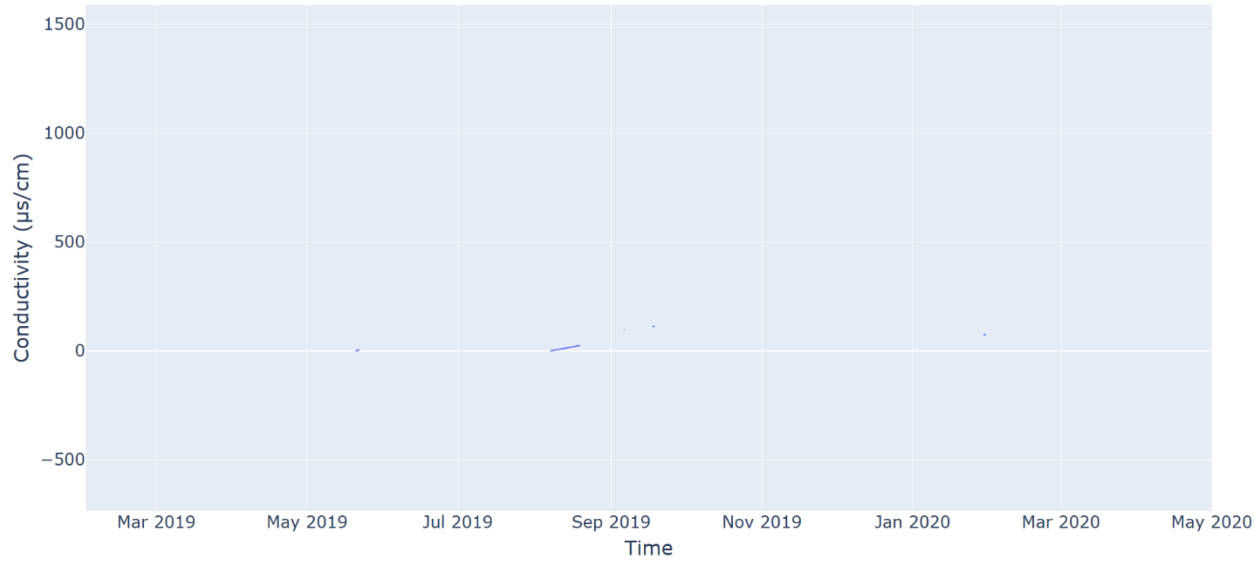

Figure 5-5. Graphical univariate analysis of conductivity - Early Warning System (EWS) for water pollution events on construction

Figure 5-6 presents a graphical detailed of some data quality issues identified in the datasets. The upper graph shows an example of flat signal (green circles) on turbidity and the lower graph shows outliers (red circles) and out of range data (orange circles).

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

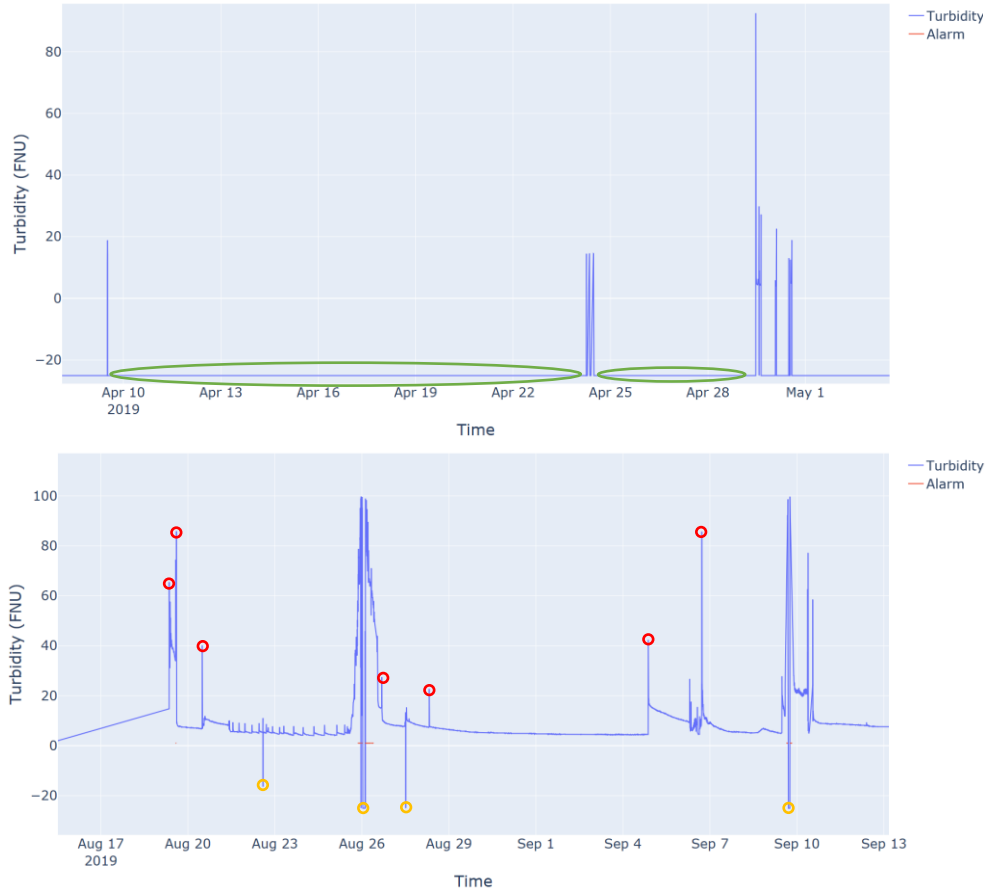

Figure 5-6. Graphical detailed of data quality issues (top: flat signal (green circles), bottom: outliers (red circles) and out of range values (orange circles)) - Early Warning System (EWS) for water pollution events on construction

Figure 5-7 presents the autocorrelation plot (see Annex 1) for the turbidity time series, that is, how the turbidity is correlated with a delayed copy of itself. The graph did not contain repeating patterns or periodic signals, hence the turbidity is not seasonal. Additionally, high correlation was observed for close prior measurements, demonstrating that moving average or similar techniques based on previous measurements could be used to detect and correct outliers.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

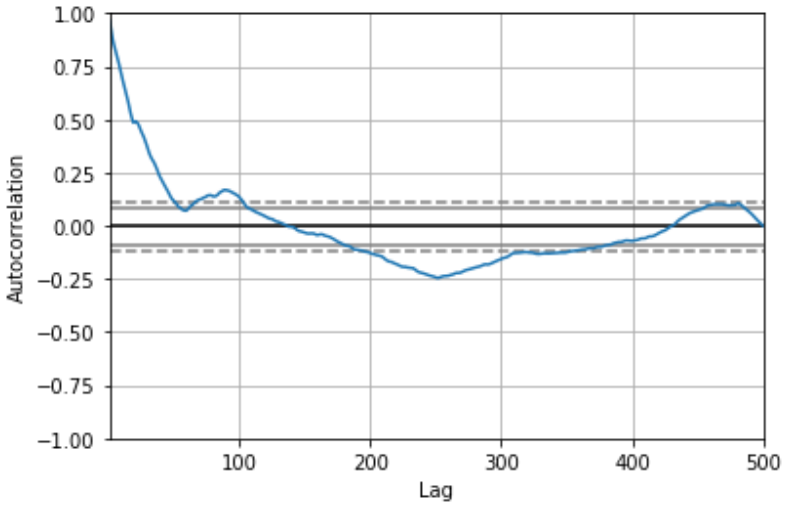

Figure 5-7. Autocorrelation plot of turbidity (lag = 15 minutes) - Early Warning System (EWS) for water pollution events on construction

Summarizing, the results of the data exploration concluded that the data quality was low due to the high amount of erroneous data contained in the features, including outliers, flat signal and out of range data. Additionally, conductivity time series only covered a few sparse data. But despite these aforementioned erroneous data, a large representative time series of turbidity can be extracted and enhanced to apply data-driven modelling. Finally, the inclusion of data from more treatment plants in the area will be studied in the next iteration of this deliverable. It will allow to generalize the solution despite differing calibrations.

5.1.1.3. DATA PREPARATION

During the data preparation, firstly, the turbidity time series data set were split, discarding data too corrupt to be pre-processed properly. The final dataset contains measurements of two months, from 20th August to 14th October. Having split the data, the empty and out of range (negative values) observations were substituted by propagating last valid observation forward, that is, the previous observation that it is not empty or negative. Figure 5-8 shows the result of this pre-processing.

Moreover, a new feature to represent the state of early warning was created in the data model with the aim of applying supervised algorithms. This feature was based on manual labelling and took advantage of alarm field of the dataset, which presented real alarms based on a threshold. Therefore, previous time instants to the alarm were labelled like abnormality (early warning). Additionally, all the alarm period was labelled as an abnormality facilitating the future learning.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

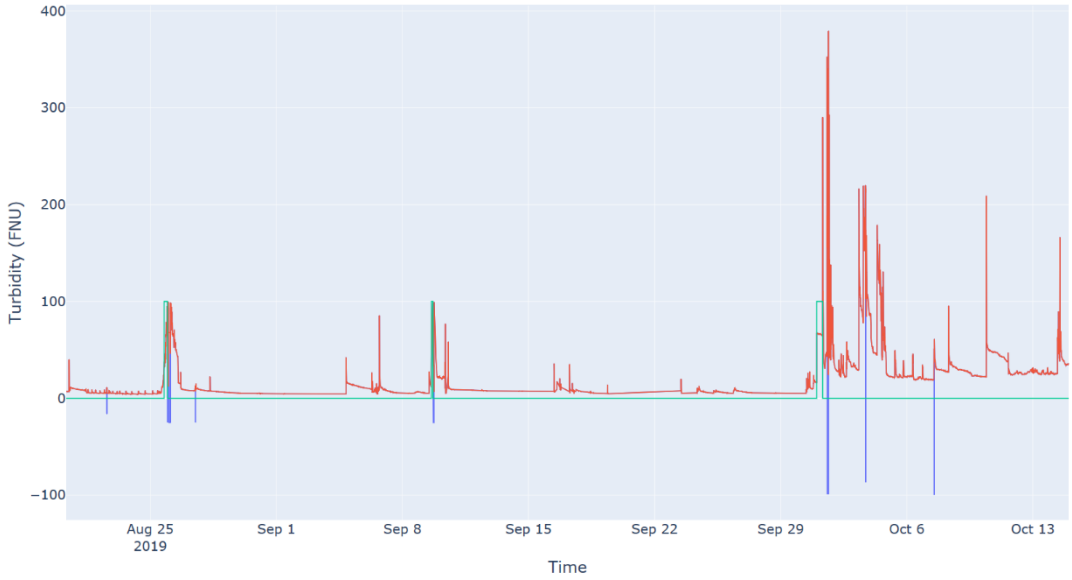

Figure 5-8. RAW and processed turbidity - Early Warning System (EWS) for water pollution events on construction (purple: RAW turbidity, red: processed turbidity, green: alarm)

The statistical measurements of the turbidity time series varied due to these changes. Below, the results are presented.

Table 23. Statistical details of pre-processed turbidity - Early Warning System (EWS) for water pollution events on construction

Feature	Count	Mean	σ (SD)	Min	Q1	Median	Q3	Max
Turbidity	66638	18.1	23.5	4.0	5.4	7.8	24.6	379.7

Also related to improving data quality, smoothing techniques like Exponential Moving Average (EMA) were applied. Exponential Moving Average is a type of Moving Average (MA) that places a greater weight and significance on the most recent data points. It is useful to smooth the signal and minimize the impact of the outliers. Figure 5-9 presents the pre-processed turbidity and the result of apply EMA with three different smooth factors (5, 10 and 40).

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

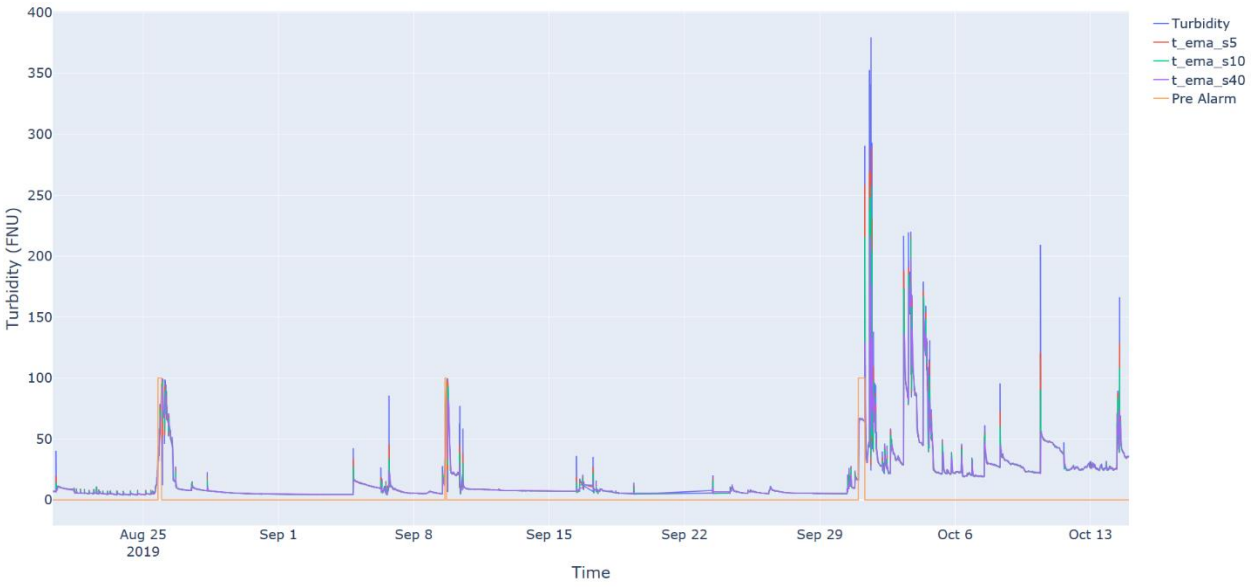

Figure 5-9. Smoothing of turbidity signal by applying EMA - Early Warning System (EWS) for water pollution events on construction

Figure 5-10 shows more clearly the impact of the smoothing in the turbidity. Higher smoothing factor minimized the impact of the new values and hence, and therefore of the possible outliers.

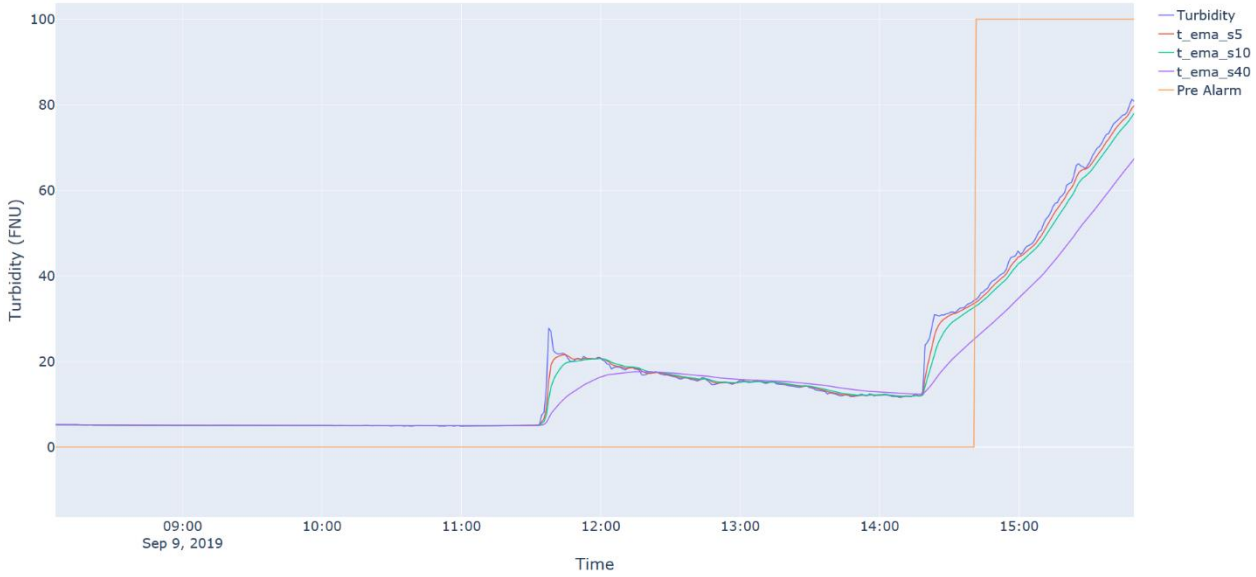

Figure 5-10. Zoom in on smoothing of turbidity signal by applying EMA - Early Warning System (EWS) for water pollution events on construction

Currently, alarms are triggered when turbidity increases and exceeds a threshold. The slope of the smoothed turbidity and the smoothed turbidity allow to characterize the behaviour of the turbidity time-series. For example, positive slope and higher smoothed turbidity values could be linked to early warnings of pollution events. Then, four new features were added to the data sets based on calculating the slope of the smoothed turbidity for four different windows, one of 10 values, other of 25, other of 50, and finally one of 100. Different window sizes are useful to capture trends at different time scales. For example, the windows of 10 values is useful to detect fast warnings, which cannot be captured with larger windows due to the slight impact of new values to the slope. Instead, large windows, like a window of 100 values, allows to capture slow but steady growths.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Figure 5-11 shows the calculated slopes for smoothed turbidity.

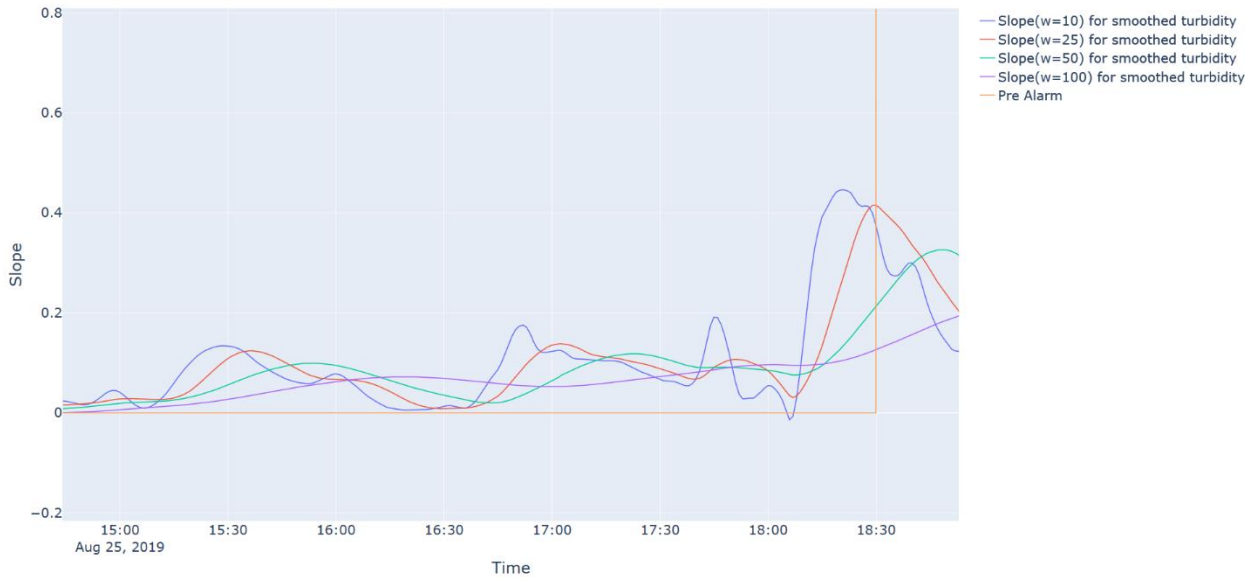

Figure 5-11. Visualization of the slopes (w=10, w=25, w=50, w=100) for smoothed turbidity- Early Warning System (EWS) for water pollution events on construction

As an initial hypothesis, smoothed turbidity and the slope of smoothed turbidity for a window size of 10, 25, 50 and 100 may be able to discriminate between warning states or normal states. Figure 5-12 presents a 3D view of three previously created features, which are the smoothed turbidity by applying EMA (smoothing factor = 40), the slope of the smoothed turbidity with a window of 10 and other slope of the smoothed turbidity, but in this case with a with a higher window (w=50). Moreover, each point is painted with green or red, depending on if it represents a condition of pre-alarm (red) or not (green). Both clusters, normal states, and warning state, are partially separated, demonstrating that these features can be useful to provide early warning.

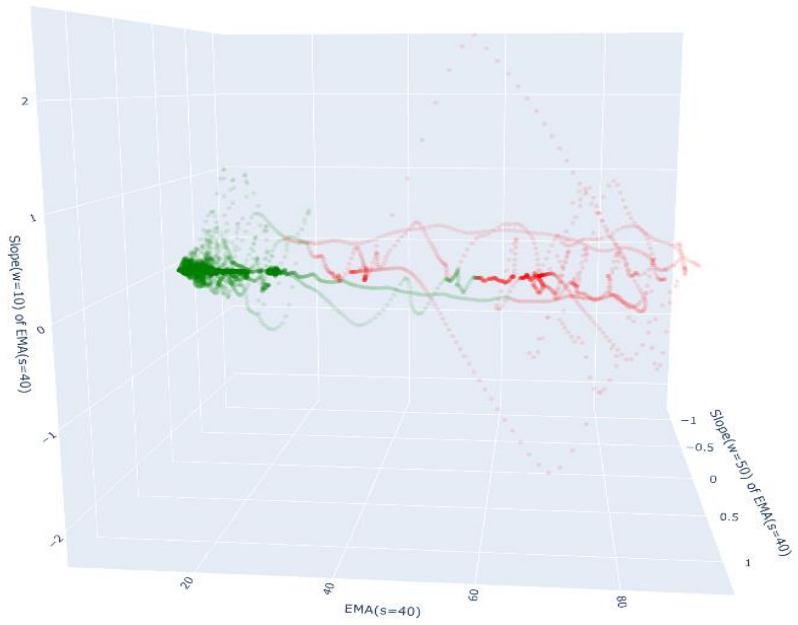

Figure 5-12. 3D view of features for class separation (red: alert and green: normal) - Early Warning System (EWS) for water pollution events on construction

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Finally, the data model used to learn was defined as represented in Table 24.

Table 24. Data model used to learn - Early Warning System (EWS) for water pollution events on construction

Feature	Description	Type
Smoothed turbidity	The EMA of the turbidity in a time period	Float
Very short trend of smoothed turbidity	Trend of the smoothed turbidity in the last 10 minutes	Float
Short trend of smoothed turbidity	Trend of the smoothed turbidity in the last 25 minutes	Float
Medium trend of smoothed turbidity	Trend of the smoothed turbidity in the last 50 minutes	Float
Large trend of smoothed turbidity	Trend of the smoothed turbidity in the last 100 minutes	Float
Abnormality	Indicate if time instant is normal or abnormal. Key feature based on manual labelled.	Boolean

5.1.1.4. MODELLING & EVALUATION

The two classes previously manually tagged, normal and abnormal state, were totally unbalanced. The warning state class represented only a 2.6% of the total observations discouraging the use supervised learning techniques. Therefore, Novelty Detection approach based on unsupervised learning techniques were tackled.

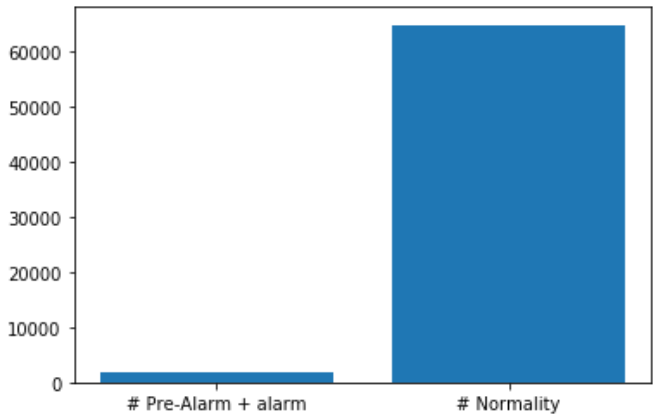

Figure 5-13. Unbalanced tagged classes of pollution events alerts - Early Warning System (EWS) for water pollution events on construction

Then novelty detection techniques allow to decide whether a new observation belongs to the same distribution as existing observations or should be considered as different, that is, abnormal or unusual observation. Then, data-driven model will be trained with normality observations in order to be able to detect if the new observations are abnormal, that is, are unusual for the trained set and hence are pre-alarm events.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The dataset was split in two data sets, one for training and another for testing. The training dataset only includes normality observations and contains 17109 observations from 9th September to 30th September (see Figure 5-14). Instead, the test dataset includes normality and abnormal observations with a total of 28987 observations from 20th August to 9th September (see Figure 5-15).

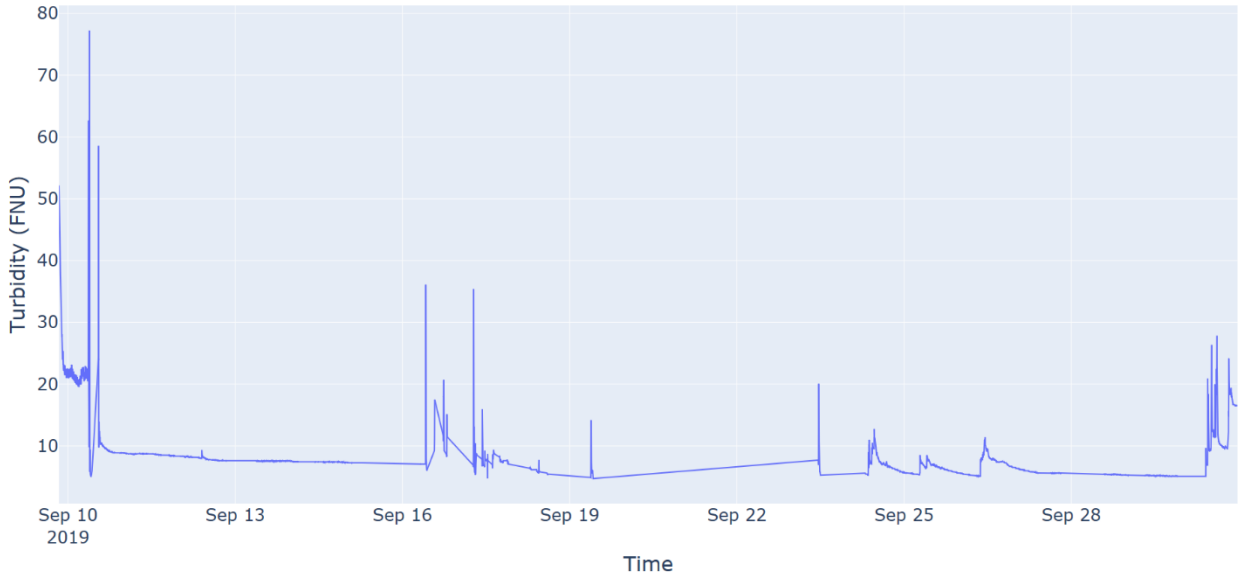

Figure 5-14. Training data set for predict pollution events - Early Warning System (EWS) for water pollution events on construction

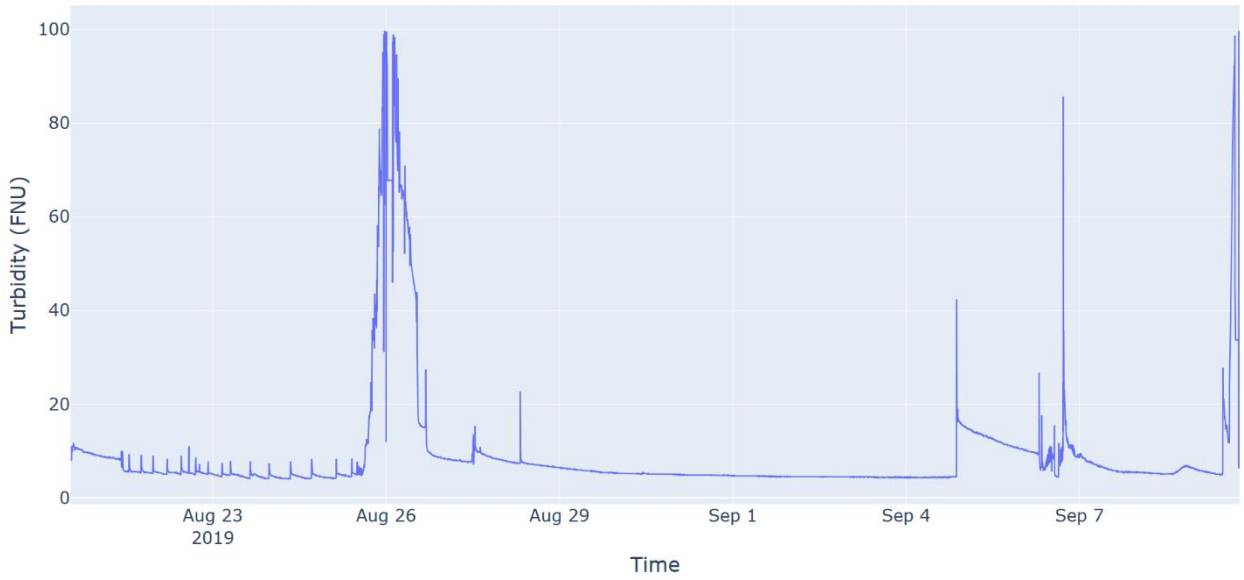

Figure 5-15. Testing data set for predict pollution events - Early Warning System (EWS) for water pollution events on construction

One-class *Support Vector Machines (O-SVM)*, *Isolation Forest (IF)* and *Local Outlier Factor (LOF)* were compared (see Annex 1 for more detailed information about the algorithms). All algorithms were trained only with normal behaviour and validated with new data. The *Numenta benchmark* and the precision metrics were adopted to evaluate the predictions of the different models and more detailed information about this scoring metrics is on Annex 2. Below, Table 25 presents the initial results of the evaluation.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 25. Results of the O-SVM, IF and LOF evaluation

Algorithm	Numenta Score	Precision Score
O-SVM	0.21	0.05
IF	0.23	0.46
LOF	0.15	0.05

Moreover, Figure 5-16, Figure 5-17 and Figure 5-18 compare tagged early warnings against predicted early warnings.

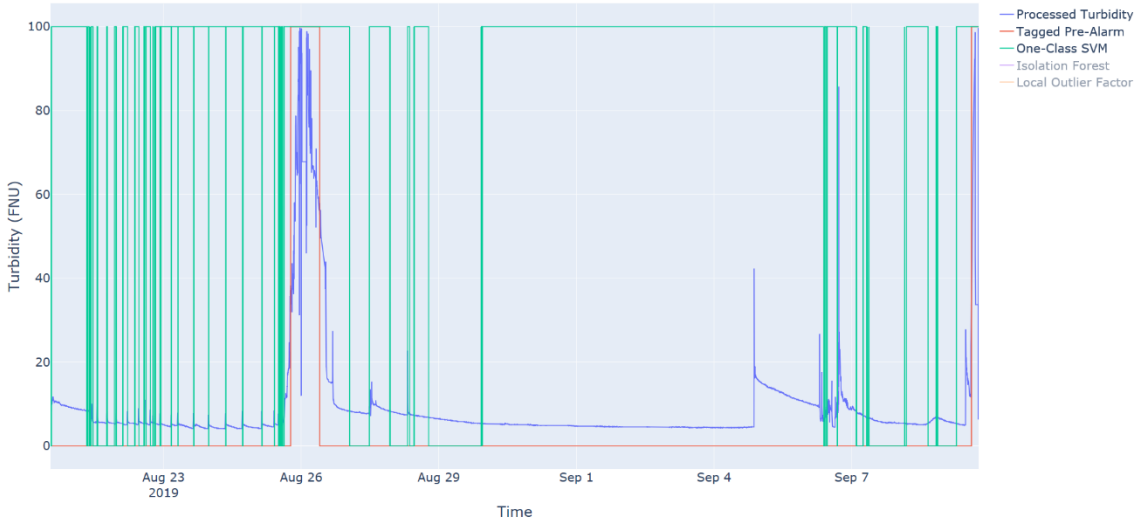

Figure 5-16. Results of prediction based on One-Class SVM algorithm - Early Warning System (EWS) for water pollution events on construction

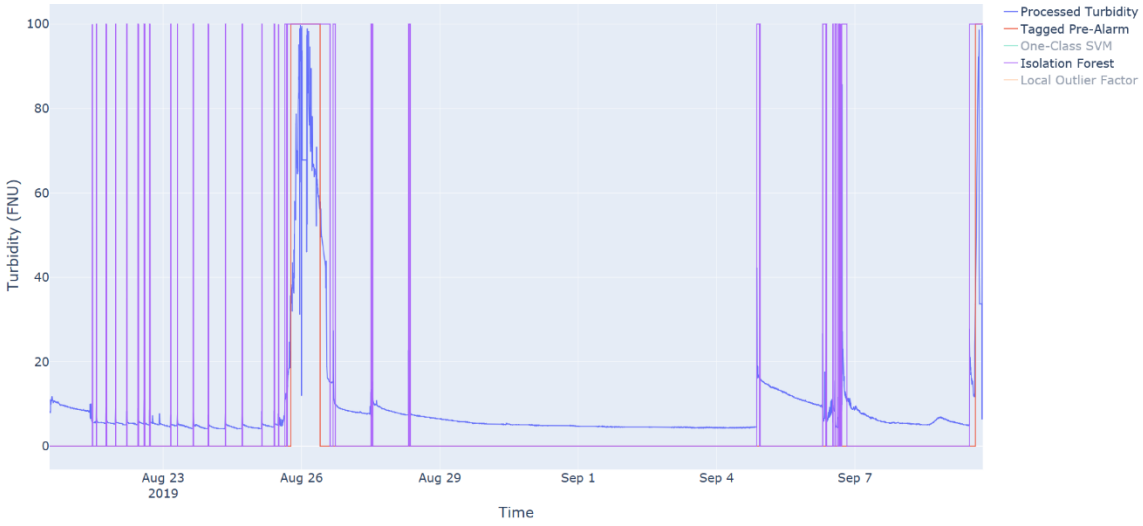

Figure 5-17. Results of prediction based on Isolation Forest algorithm - Early Warning System (EWS) for water pollution events on construction

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

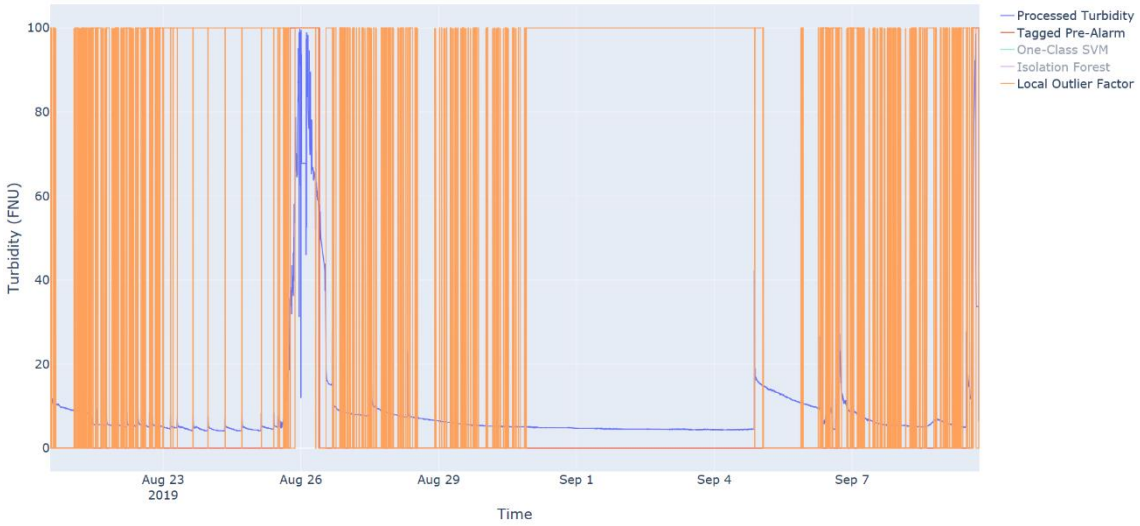

Figure 5-18. Results of prediction based on Local Outlier Factor algorithm - Early Warning System (EWS) for water pollution events on construction

The best results were linked to *IF* algorithm, obtaining a *Numenta Score* of 0.23 and a *Precision Score* of 0.46. The hyperparameters of *IF* used to learn were optimized, reaching a *Numenta Score* of 0.80 and *Precision Score* of 0.78. Figure 5-19 compares tagged early warnings against predicted early warnings, once optimized the hyperparameters.

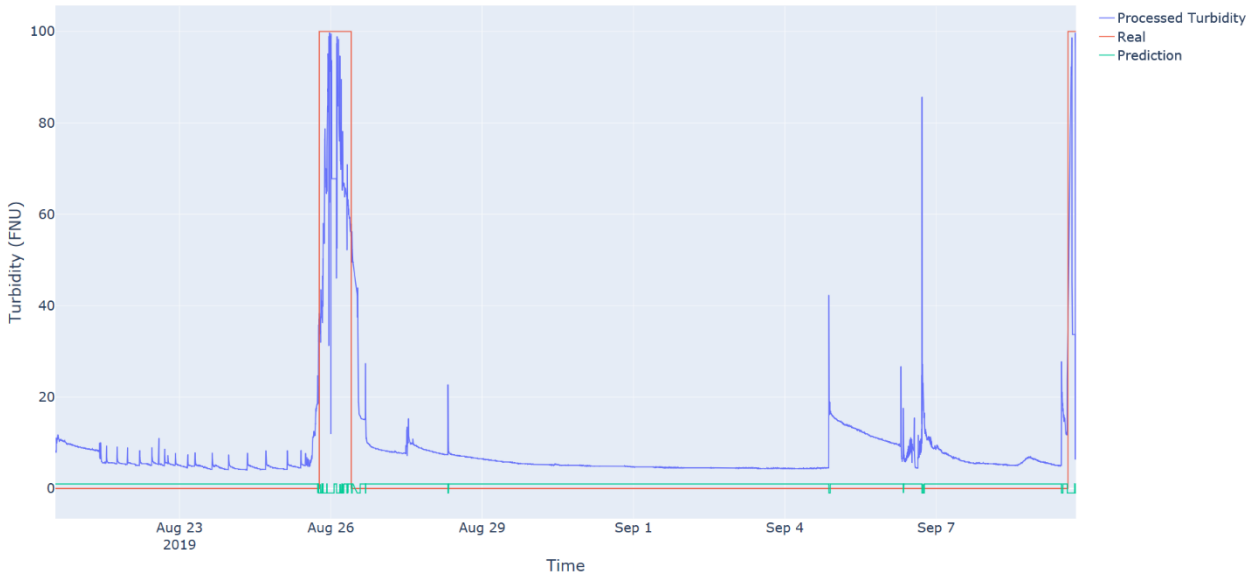

Figure 5-19. Results of prediction based on fine-tuned *Isolation Forest* algorithm - Early Warning System (EWS) for water pollution events on construction

To sum up, the initial datasets had to be fractionated to take advantage of only subsets with a higher quality, due to this, unrepresentative dataset was used for training (1 month of normality) and only two anomalies were available for validation. Despite of this aforementioned issues, *IF* presented respectable results, but more efforts should be focused on obtaining more representative and high-quality datasets. Additionally, some outliers were identified as early warnings despite of applying signal smoothing techniques like *EMA*. Therefore, a future iteration with an initial step to remove outliers could improve current results.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6. GENERAL CASE

6.1. DATA QUALITY PREDICTION FOR FLOW PATTERNS

Complex water networks, such as combined sewage systems, have monitoring systems to acquire, store and validate data from many flow meters and other sensors to achieve accurate monitoring of the whole network. A common problem is the lack of reliability of the flow meters (offset, drift, breakdowns), producing false flow data readings. These false data must also be detected and replaced by estimated data since flow data are used for several network water management tasks and tools. Additionally, the study and understating of the citizen behaviour requires flow patterns based on dry-days due to rain events distorting the flow patterns, masking the real citizen behaviour.

6.1.1. ITERATION 1

6.1.1.1. BUSINESS UNDERSTANDING

The business objective of this study case is to facilitate the study of citizen behaviour based on flow patterns, providing automatic tools to classify patterns.

Concerning the AI goal, the aforementioned business goal can be translated to "*Classify automatically flow patterns on normality and abnormality, taking into account that abnormality is any pattern not representative of a dry weather day*".

The most relevant criteria for a successful prediction provide a certain level of predictive accuracy and anticipation.

6.1.1.2. DATA UNDERSTANDING

The study case takes advantage of one of the datasets provided by the D2.1 "*Testbed data and sensor validation*" to accelerate the implementation of the data-driven model. This data set contains flow rate data recorded in a pumping station of a combined wastewater catchment in a suburb of Stockholm, Sweden.

The data set contains three different data sources, one for combined wastewater flow rate, another for accumulated precipitation and the last for outdoor temperature. All the data sources are available through IVL SharePoint platform in the WP2 folder. No problems have been identified in accessing them. Table 26 summarizes this.

Table 26. Details about data source acquisition - Data quality prediction for flow patterns

Datasource	Location	Method used to acquire	Problems
Combined wastewater flow rate	IVL Sharepoint (WP2)	Download file from Sharepoint	No problems identified
Accumulated precipitation	IVL Sharepoint (WP2)	Download file from SharePoint	No problems identified
Outdoor temperature	IVL Sharepoint (WP2)	Download file from SharePoint	No problems identified

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

As was commented previously, the *Combined wastewater flow rate* data source contains the flow rate data recorded in a pumping station of a combined wastewater catchment in a suburb of Stockholm. The information is gathered on the excel document, which contains 39261 registers and 2 fields (*Time* and *Flow Rate*) in a resolution of 15 minutes. *Time* feature corresponds to the date and time of the measurement on format YYYY-MM-dd HH:mm:ss and *Flow Rate* to the measured value of flow rate in m^3/h.

The outdoor temperature data source contains the temperature, gathering on TSV document where the data are separated by tabs. The document is structured in three features (*Time* and *Temperature*) with 39688 registers. Similarly to *Combined wastewater flow rate* data source, *Time* feature also corresponds to the date and time of the measurement on format YYYY-MM-dd HH:mm:ss and *Temperature* feature corresponds to the measured value of outdoor temperature.

The last data source, *Accumulated Precipitation*, contains the rainfall accumulation in an Excel file. 2436 registers of *Time* and *Rainfall* features are stored. *Time* storages the date and time in format YYYY-MM-dd HH:mm:ss when 0.2mm of rainfall is accumulated, hence the distance between registers is not stable. *Rainfall* is constant for all registers, containing the value 0.2mm.

Table 27 and Table 28 summarize previous information.

Table 27. General details about available data sources - Data quality prediction for flow patterns

Data Source	Description	Format	# Registers	# Feature
Combined wastewater flow rate	Flow rate data recorded in a pumping station of a combined wastewater catchment in a suburb of Stockholm	Excel	39261	3
Outdoor temperature	Temperature data recorded outdoor the pumping station	TSV	39688	3
Accumulated precipitation	Instant of time when rainfall accumulation reaches 0.2mm	Excel	2436	2

Table 28. General details about available features - Data quality prediction for flow patterns

Feature	Description	Type	UoM	Data Source
Time	Date and time of the measurement	Date	YYYY-MM-dd HH:mm:ss	Combined wastewater flowrate
Flow rate	Flow rate	Numerical	m^3/h	Combined wastewater flowrate
Time	Date and time of the measurement	Date	YYYY-MM-dd HH:mm:ss	Outdoor temperature
Temperature	Outdoor temperature	Numerical	ºC	Outdoor temperature
Time	Date and time when 0.2mm of rainfall is accumulated	Date	YYYY-MM-dd HH:mm:ss	Accumulated precipitation
Rainfall	Rainfall accumulation (0.2mm)	Numerical	mm	Accumulated precipitation

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Once identified the data source and their features, an initial *Exploratory Data Analysis (EDA)* was carried out. Table 29 presents the statistical basis metrics of the features. Features like *Time* are not included in this analysis due to them not being numerical.The presence of outliers was expected due to the minimum and maximum given the impression to be out of range. Minimum was negative and maximum was very far from the average.

Table 29. Statistical basis metrics of features - Data quality prediction for flow patterns

Feature	Count	Mean	σ (SD)	Min	Q1	Median	Q3	Max
Flow Rate	39621	60.7	26.2	-8.3	43.0	60.7	74.8	378.8

Figure 6-1 presents a graphical univariate analysis of flow, including the entire time series (from September 2018 to October 2019). Figure 6-2 and Figure 6-3 presents some identified problems on the data such as flat signal (red circle), outliers (green circle) and out of range values (orange circle).

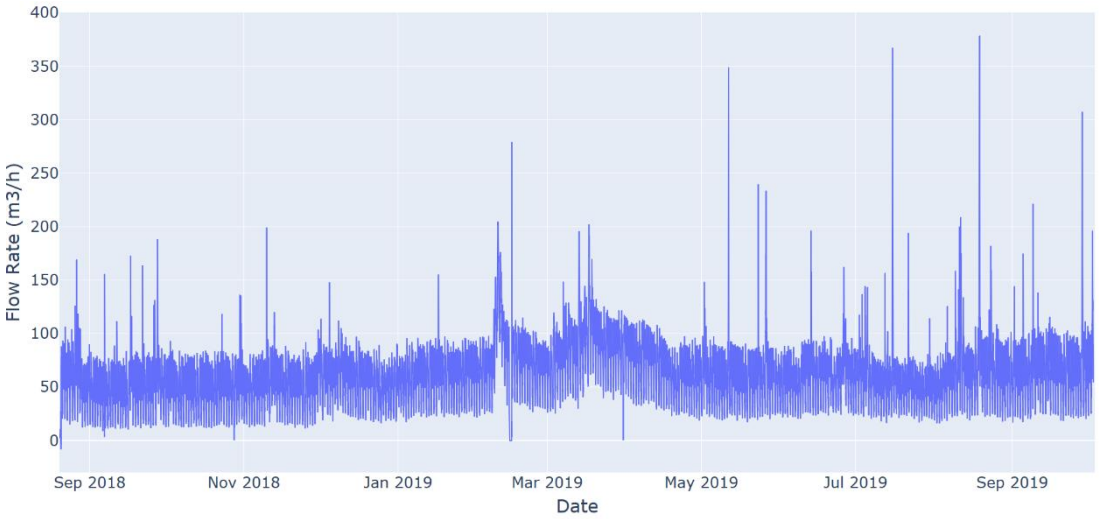

Figure 6-1. Flow rate times series - Data quality prediction for flow patterns

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

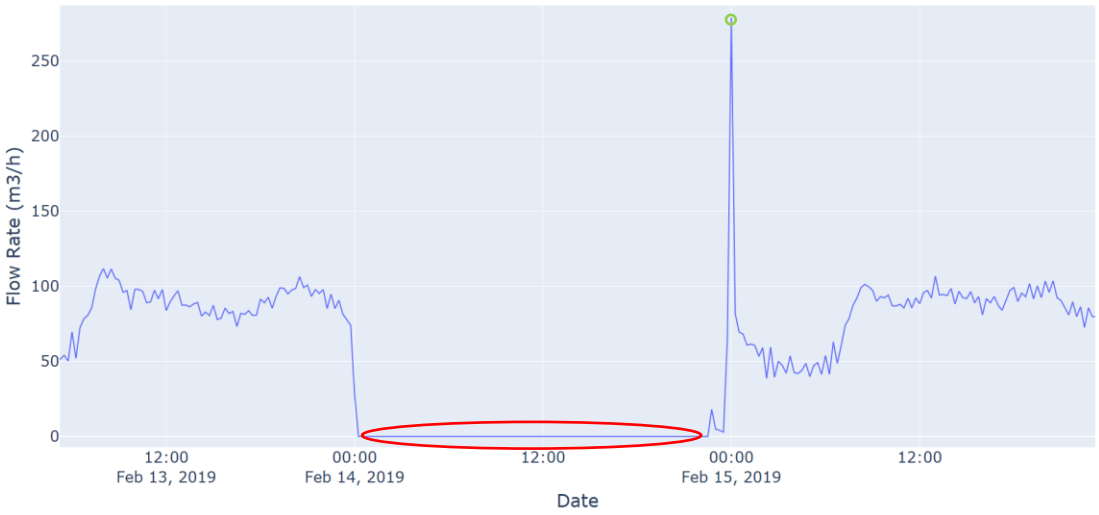

Figure 6-2. Example of flat signal and outlier on flow time series - Data quality prediction for flow patterns

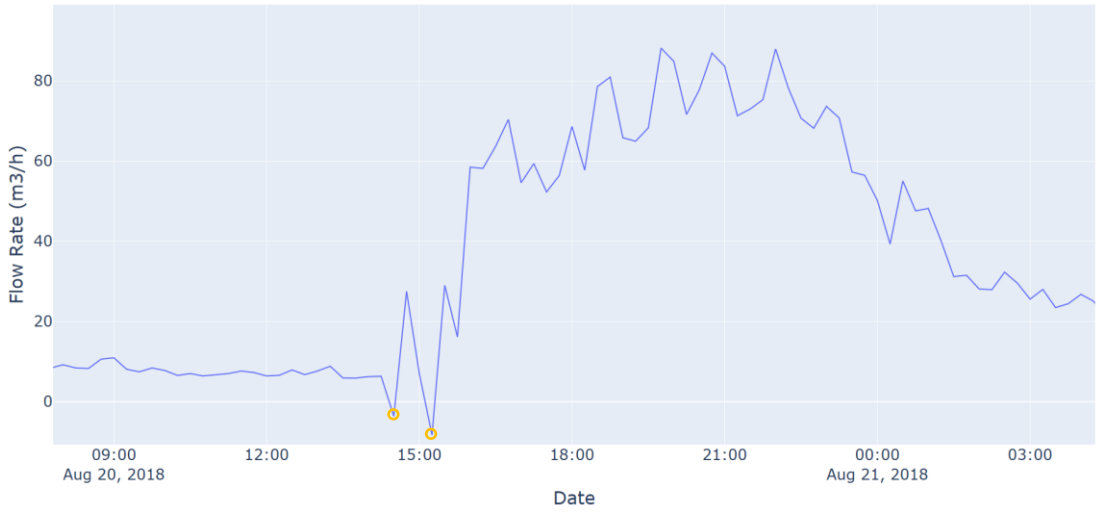

Figure 6-3. Example of out of range values on flow time series - Data quality prediction for flow patterns

These anomalous data, such as flat signal, outliers or out of range values, was not repaired due to the aim of these study case is detect abnormality on the data, which could be generated by anomalous data or anomalous patterns (for example, rainy days).

Figure 6-4 presents the autocorrelation plot (see Annex 1 for more conceptual information about the autocorrelation analysis) for flow, how the flow is correlated with a delayed copy of itself. The graph shows that previous measurements are highly correlated to the current and hence, they can be used to detect and correct outliers. Additionally, the local maximum was observed every 96 lags (24 hours) demonstrating that the signal was seasonal, and the pattern was repeated each day.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

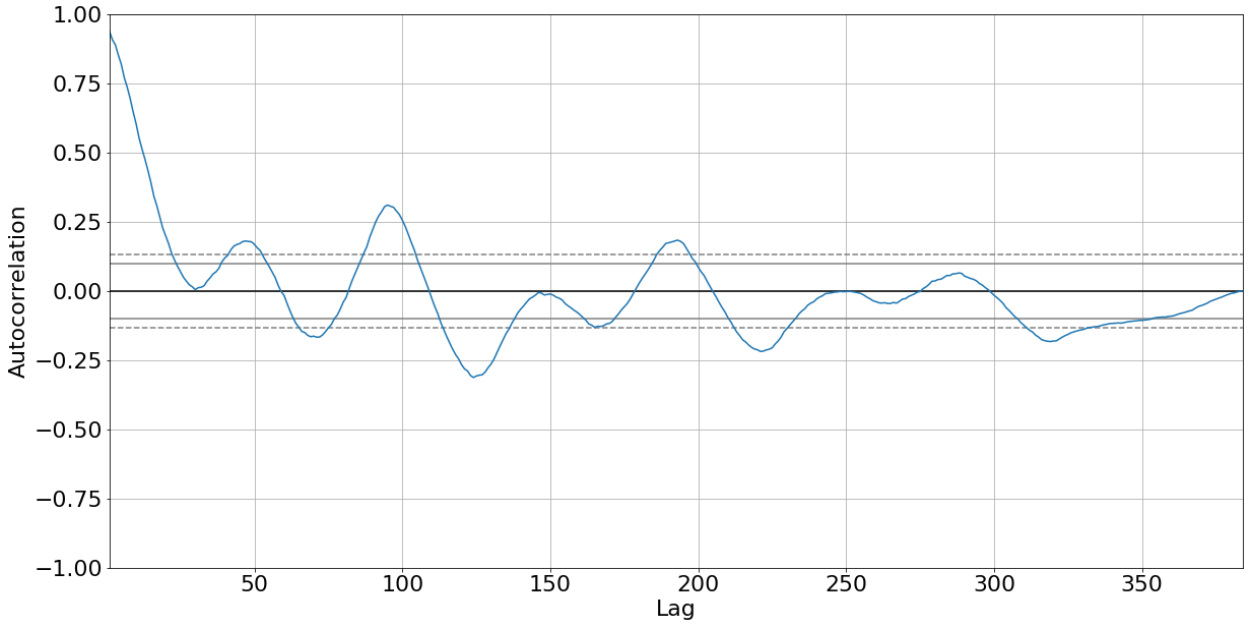

Figure 6-4. Autocorrelation plot of flow for 4 days (lag=15 minutes) - Data quality prediction for flow patterns

Summarizing, the results of the data exploration concluded that the data quality is high despite some erroneous data. Autocorrelation analysis demonstrated the seasonality of the flow, validating an initial approach based on univariate analysis. Precipitation data will be exploited in further iterations to refine the model if necessary.

6.1.1.3. DATA PREPARATION

Firstly, the data were grouped into daily time series by applying windowing techniques. Figure 6-5 presents the daily time series. The pattern is clearly visible by the overlapping of time series and from now on we will consider as normality. The flow is maintained stable during the night and starts to grow at the first hour of the morning, coinciding with people's waking hours. Later, the pattern decreases and remains stable during the working day, growing slightly again when citizens return home. The time series away from this pattern contains rain events, possible sensor failures or special events which from now on we will consider as an abnormality.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

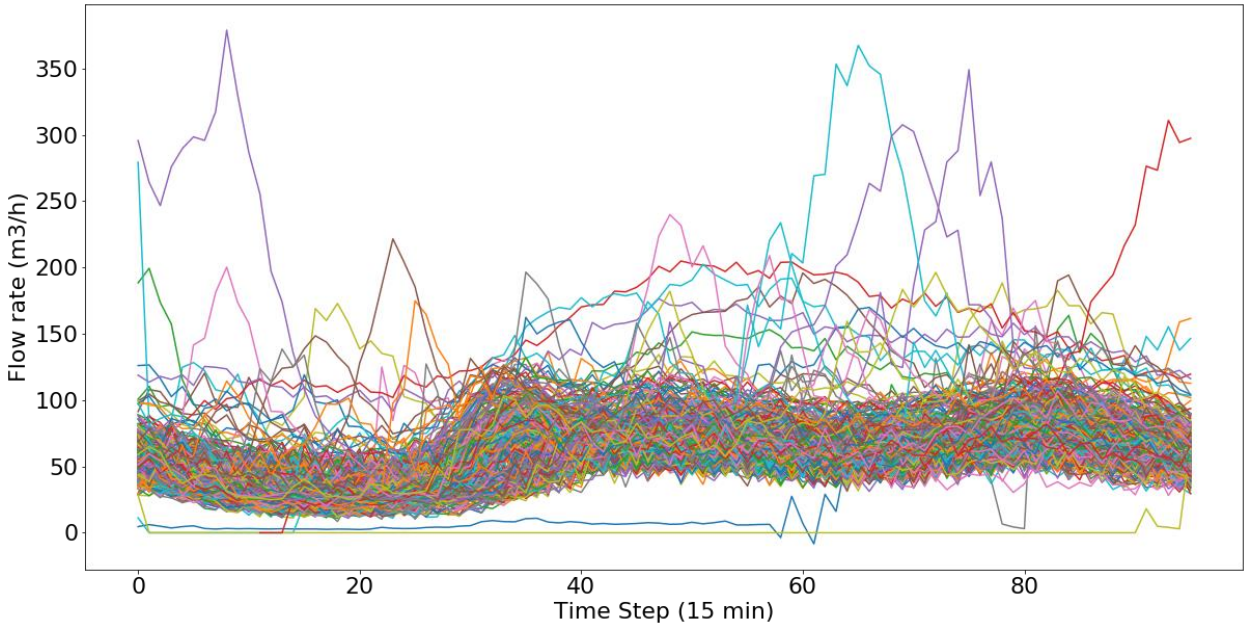

Figure 6-5. Flow data split into daily time series - Data quality prediction for flow patterns

The pattern is more clearly visible on following scatter plot (see Figure 6-6). Additionally, this view allows to discriminate slightly two different patterns, which are aligned with working days and holidays. On holidays, there was a delayed and mild slope of the flow during the morning (see from step 25 to 40 -from 06:15 to 10:00).

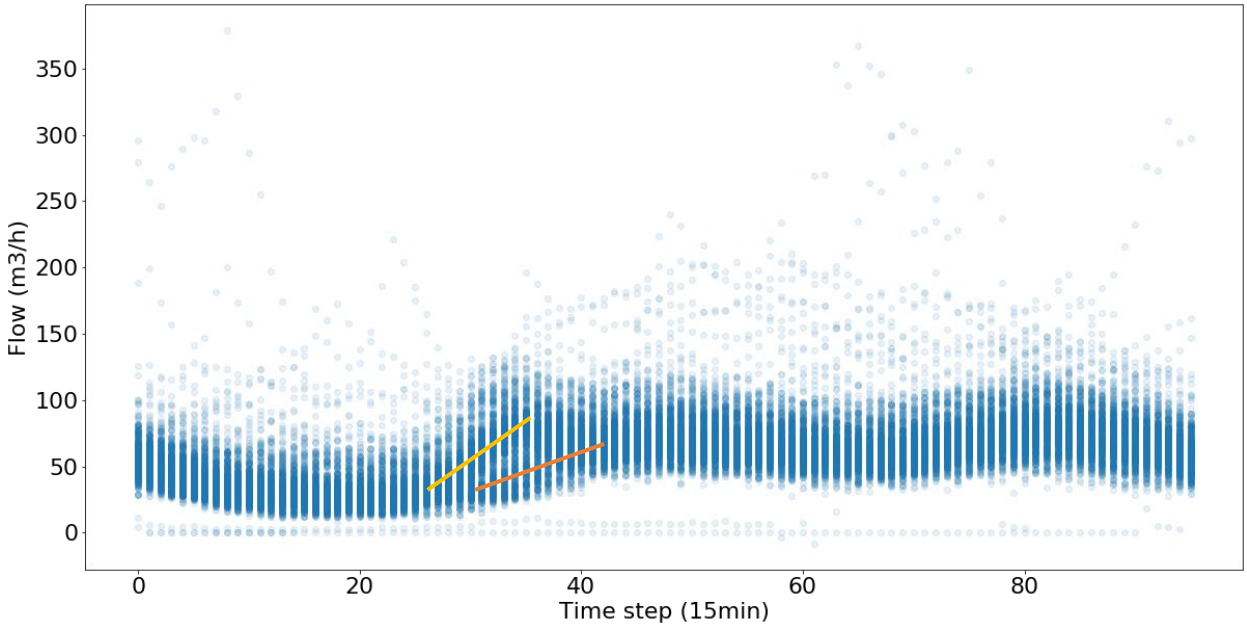

Figure 6-6. Density of flow time series (yellow: workday trend, orange: holiday trend) - Data quality prediction for flow patterns

The data were tagged manually by analysing and discriminating visually the time series considered normal from the abnormal. It was not applied automatic process due to the low amount of data and the need for quality tagging, which is essential to extract patterns of the data. Figure 6-7, presents the result of the manual tagging. Blue and green time series represents normal time series, however, blue lines are workday and green holidays. Instead, red lines presented abnormal daily time series, that is, time series with rainy events, erroneous data or special flow events. This tagging helps to understand what happened in the data and can be useful if supervised learning techniques are applied.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

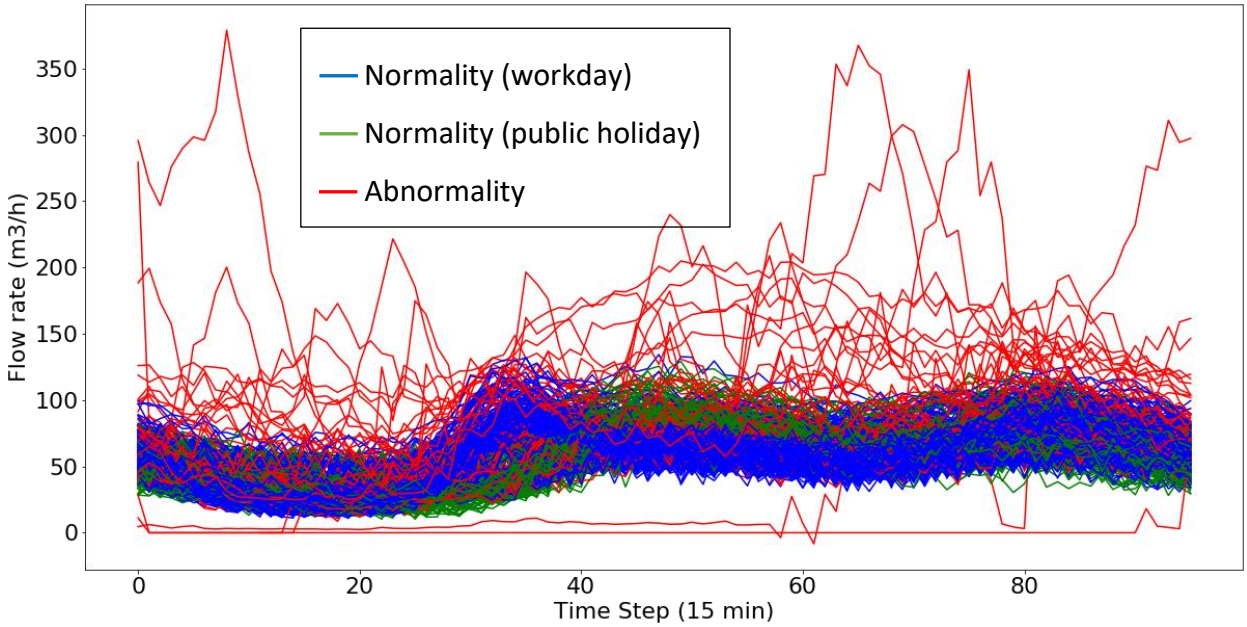

Figure 6-7. Manual tagging of normality and abnormality flow (blue: normality (work day), green: normality (holiday); red: abnormality) - Data quality prediction for flow patterns

Taking advantage of tagging data, data distribution of flow time series was analysed for understanding the data and determine the creation of new features. Figure 6-8 presents the data distribution of six different time series, four representative time series of normality (two of workdays and two of holidays) and two representative time series of abnormality. Similar data distribution patterns were observed for normality days, varying only the density of some ranges. Instead, abnormality patterns were totally different from normality patterns, they presented less homogenous and large data distributions. Therefore, statistical measurements related to data distribution (for example, percentiles) could provide relevant information to discriminate between normality and abnormality patterns.

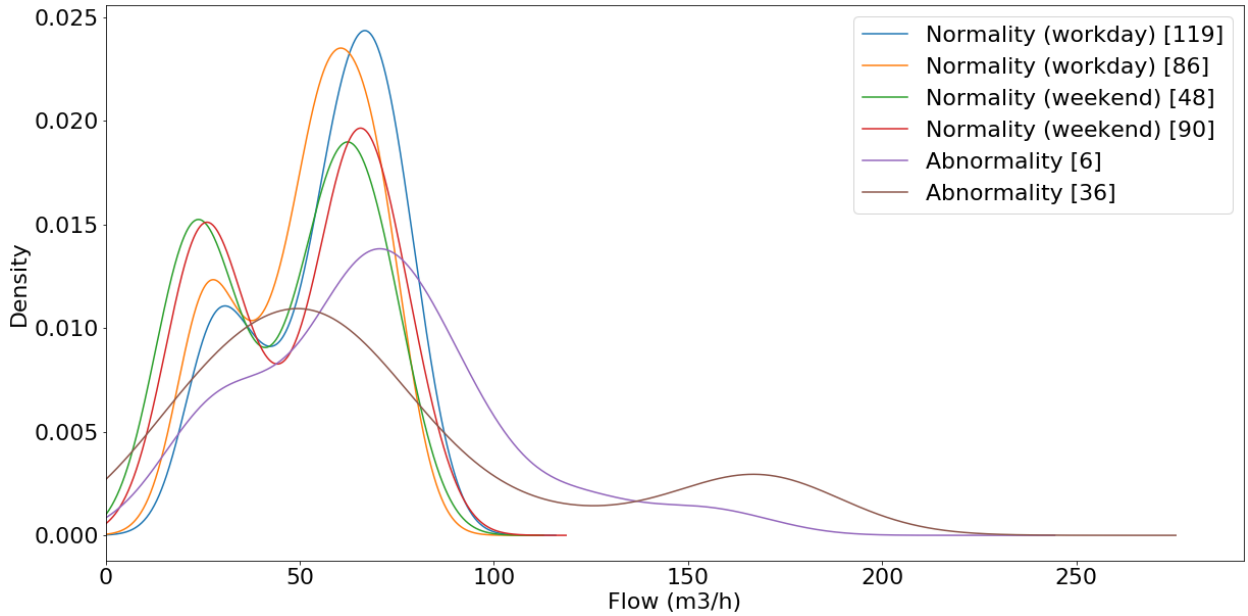

Figure 6-8. Density plot of flow time series including normality and abnormality - Data quality prediction for flow patterns

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

As an initial hypothesis, percentile 10, median (percentile 50) and percentile 90 may be able to discriminate between normal and abnormal patterns. Therefore, these percentiles were calculated for each day. Figure 6-9 shows a visual validation of the hypothesis where each percentile is presented on one axis. Percentile 90 on axis X, percentile 50 on axis Y and percentile 10 on axis Z. The percentiles of normal days form a cluster, which can be separable from abnormal days. Therefore, these features could be initially used to discriminate the daily time series.

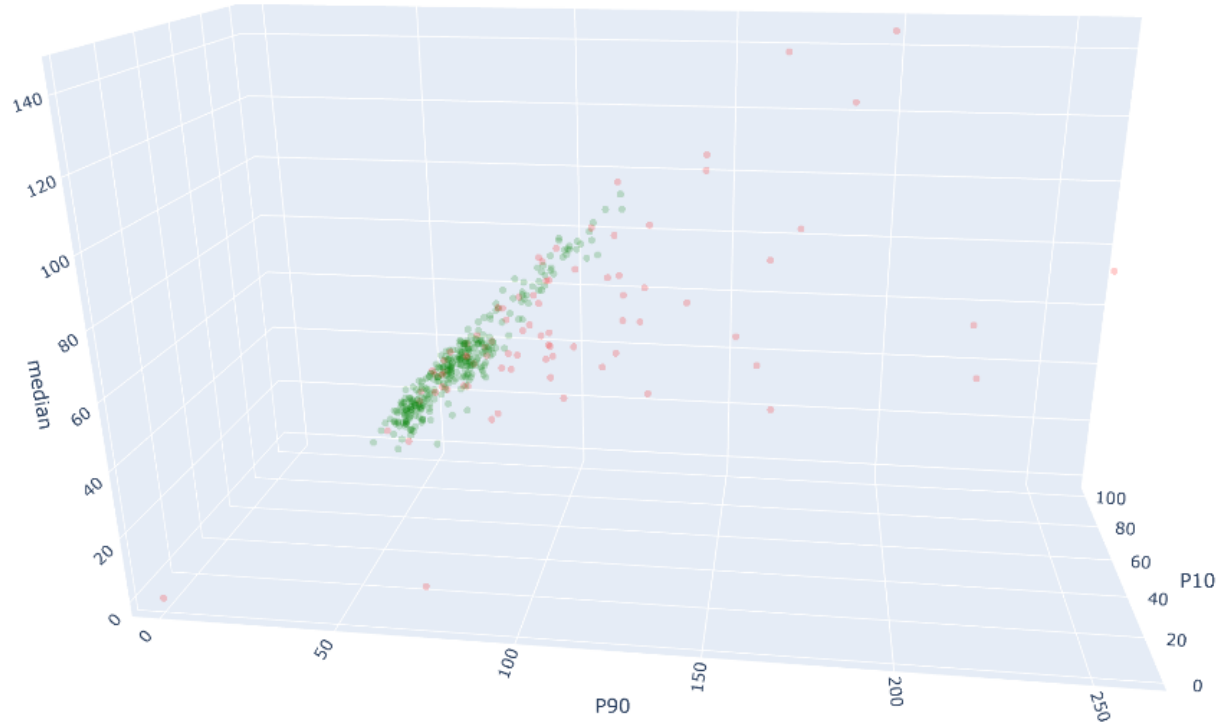

Figure 6-9. 3D visualization of percentile 10, percentile 50 and percentile 90 of daily time series (green: normal days, red: abnormal days) - Data quality prediction for flow patterns

Finally, the data frame used to learn was defined as represented in Table 30.

Table 30. Data frame used to learn

Feature	Description	Type
Percentile 10 of Flow	The percentile 10 of the flow values gathered during a day	Float
Percentile 50 of Flow	The percentile 50 of the flow values gathered during a day	Float
Percentile 90 of Flow	The percentile 90 of the flow values gathered during a day	Float
Abnormality	Indicate if the timeseries is normal or abnormal. Key feature based on manual labelled.	Boolean

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.1.1.4. MODELLING & EVALUATION

Cross-validation was applied to evaluate the model and estimate how accurately is our predictive model, avoiding to use the same data to train and test, and hence validating against independent and yet-unseen data set. More specifically, the Time Series Split technique was used (see Annex 3).

For anomaly detection, the goal is to identify all and only anomalies. *Recall Score* measures how well our algorithm identified *all* anomalies. Instead, *Precision Score* measures how well our algorithm identifies *only* anomalies. Then, *Recall* and *Precision Score* will be used to measure the performance of the algorithms. Nevertheless, it is important to note that *Recall Score* relevance is higher than *Precision Score* relevance because the priority is detecting all anomalies, allowing us some false positive.

Table 31 presents the accuracy, precision and recall results (see Annex 2 for more detailed information about this scoring metrics) of applying Time Series Split cross-validation with three splits for multiple classification algorithms such as *K-Nearest Neighbours (KNN)*, *Decision Tree Classifier (DTC)*, *Random Forest Classifier (RFC)*, *AdaBoostClassifier*, *Gradient Boosting Classifier (GBC)*, *Gaussian NB*, *Linear Discriminant Analysis (LCA)* and *Quadratic Discriminant Analysis (QDA)*.

A high true positive rate (*Recall Score*) was observed for *LCA*, *KNN* and *QDA* (see Table 31), showing that these algorithms are the most efficient to detect all anomalies. LCA and KNN algorithms are not the best to detect only the anomalies (*Precision Score*), nevertheless both algorithms are candidates to be exploited more accurately, jointly with QDA.

Table 31. Recall and Precision results of the initial modelling - Data quality prediction for flow patterns

Algorithm	Recall Score	Precision Score
KNN	0.62	0.71
DTC	0.43	0.77
RFC	0.53	0.77
Ada-boost Classifier	0.54	0.83
GBC	0.45	0.71
GaussianNB	0.49	0.74
LCA	0.67	0.74
QDA	0.54	0.86

Concerning confusion matrix of *LCA*, *KNN* and *QDA*, Figure 6-10 presents the final results of the cross-validation. Therefore, these matrixes contain the accumulation of confusion matrix generated during the cross-validation process with *Time Series Split* technique.

LCA presented better results, predicting properly 214 normal days and 39 abnormal days. Instead, 30 days were falsely predicted as abnormal days and 20 days as normal days.

It is important to remark that these results were affected by the different iterations of the cross-validation, impacting in the high number of False Positive and False Negative.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

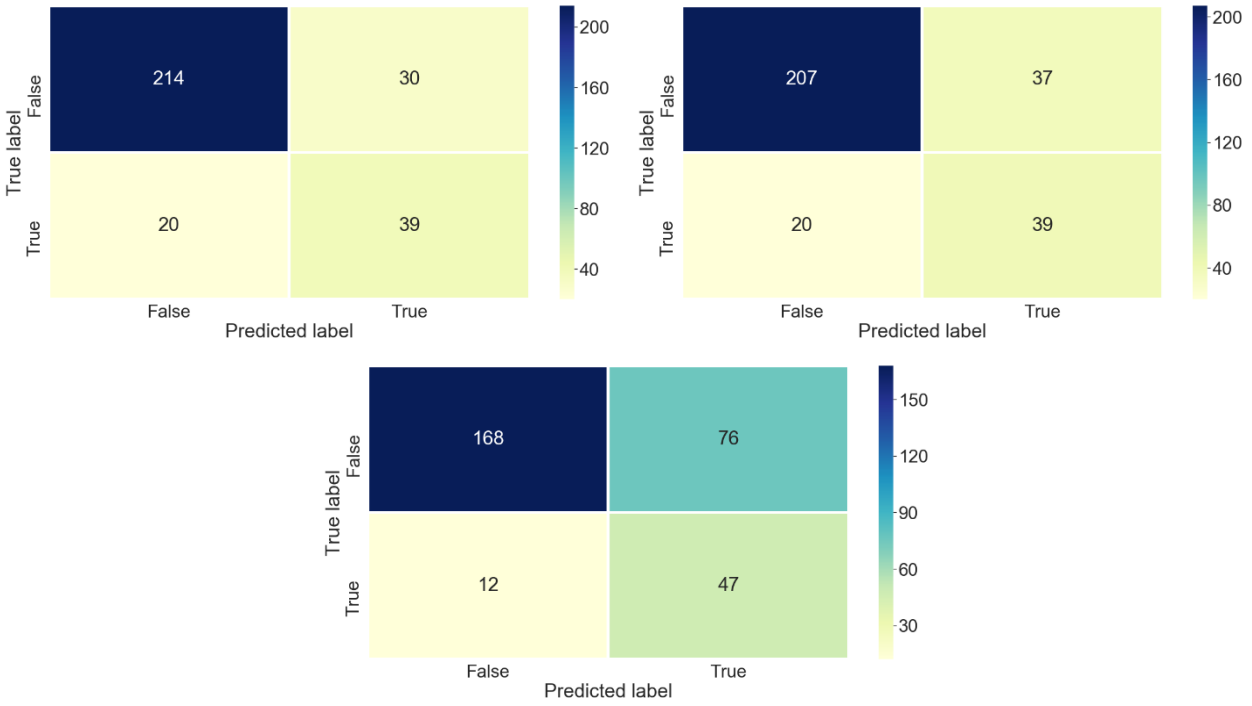

Figure 6-10. Confusion matrix resulting from applying the LCA(upper left), KNN (upper right) and QDA (lower centre) - Data quality prediction for flow patterns

False Positive and False Negative were analysed in order to determine how to enhance the first model based on percentiles. Figure 6-11 compares the data distribution of some erroneously predicted time series with normality. As shown in the figure, the data distribution of abnormal days is higher due to higher maximum flow values by rainy contribution. Additionally, accumulated flow also increases by this rainy contribution. But this reason, it is considered to add new features to represent the maximum, minimum and accumulate flow for each day.

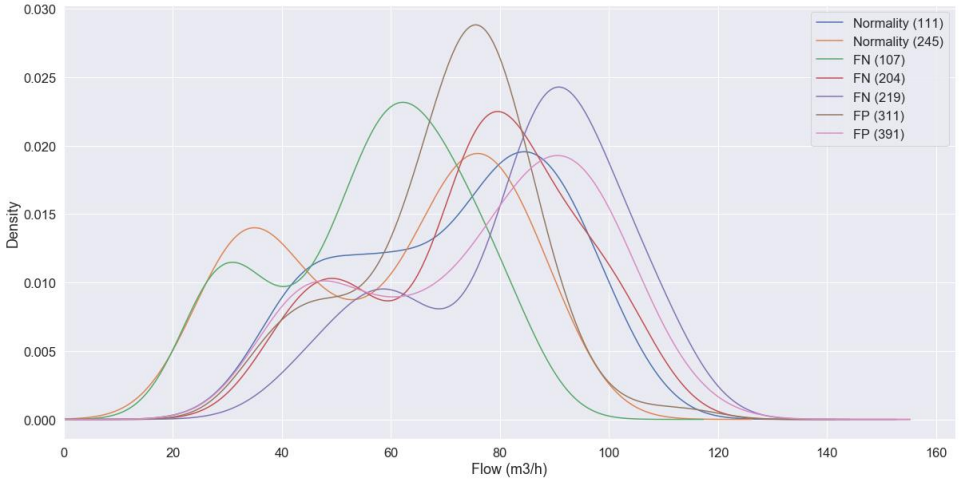

Figure 6-11. Density plot of False Positive and False Negative - Data quality prediction for flow patterns

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.1.2. ITERATION 2

6.1.2.1. DATA PREPARATION

The hypothesis of iteration 1, based on percentiles, was enhanced by including also maximum, minimum and accumulated flow. These new features were calculated for each daily time series and included in the dataset. Figure 6-12 presents these features.

If we look at the time series in detail, the day 81, which is tagged as abnormal day, presents a percentile 90 that matches with normality. Nevertheless, the new feature related to maximum flow is far from the expected normal values. Therefore, new features such as maximum, minimum and accumulated could be useful to discriminate the data sets. For example, maximum and accumulated flow to detect rainy events and minimum to detect erroneous data sets.

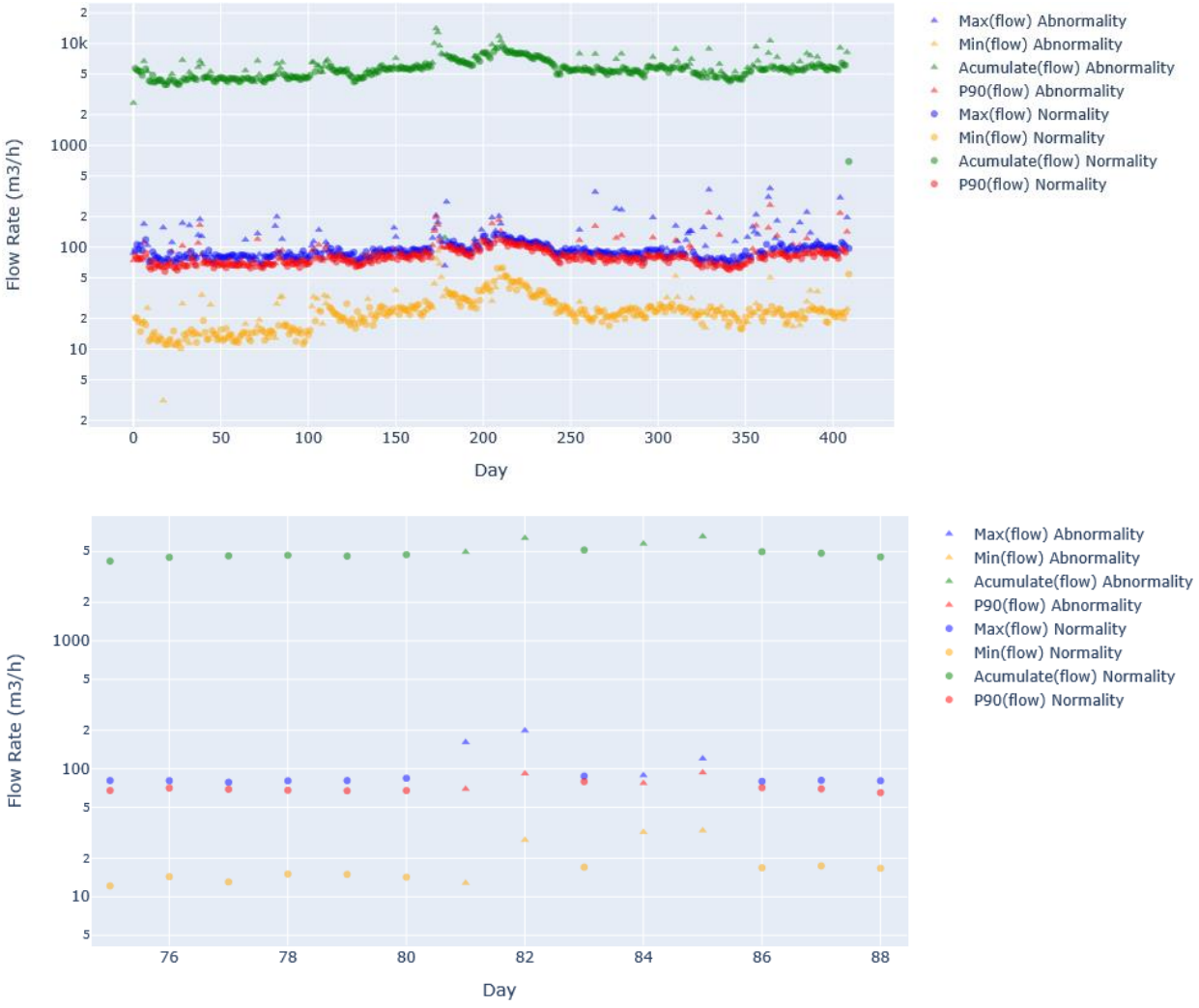

Figure 6-12. Maximum and minimum flow, accumulated flow and percentile 90 of data distribution for each daily time series - Data quality prediction for flow patterns

Figure 6-13 validates visually if the new features discriminate abnormal from normal days. Three features are presented in the graph, accumulated flow on axis X, maximum flow on axis Y and percentile 90 on axis Z. The normal days were clustered, separating them from abnormal days. Therefore, the new approach can improve the previous hypothesis.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

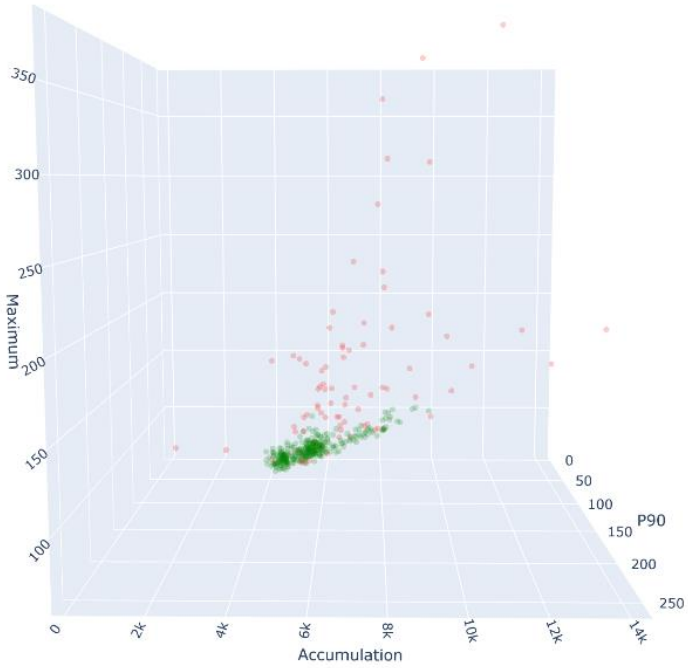

Figure 6-13. 3D visualization of percentile 90, maximum flow and accumulated flow of daily time series (green: normal days, red: abnormal days) - Data quality prediction for flow patterns

Finally, the data model used to learn was defined as represented in Table 32.

Table 32. Data frame used to learn

Feature	Description	Type
Percentile 10 of Flow	The percentile 10 of the flow values gathered during a day	Float
Percentile 50 of Flow	The percentile 50 of the flow values gathered during a day	Float
Percentile 90 of Flow	The percentile 90 of the flow values gathered during a day	Float
Min	The minimum of the flow values gathered during a day	Float
Max	The maximum of the flow values gathered during a day	Float
Accumulated	The accumulated of the flow values gathered during a day	Float
Abnormality	Indicate if the timeseries is normal or abnormal. Key feature based on manual labelled.	Boolean

6.1.2.2. MODELLING & EVALUATION

Previous models based on *LCA*, *KNN* and *QDA* were evolved, including minimum, maximum and accumulated flow. Table 33 presents the new Recall and Precision Score for them. Figure 6-14 presents the confusion matrixes for the algorithms taking advantage of cross-validation.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 33. Recall and Precision results of the initial modelling - Data quality prediction for flow patterns

Algorithm	Recall Score	Precision Score
LCA	0.69	0.69
KNN	0.47	0.33
QDA	0.94	0.43

QDA and *LCA* improved the results of the first iteration. Additionally, *QDA* presented better results than *LCA* and *KNN*, reaching a *Recall Score* 0.94 and *Precision Score* 0.43. Therefore, the data-driven model was able to predict accurately all the abnormal days. Summarizing, the *LCA* predicted properly 172 normal days and 56 abnormal days. Instead, 72 days were falsely predicted as abnormal days and 3 days as normal days (see Figure 6-14).

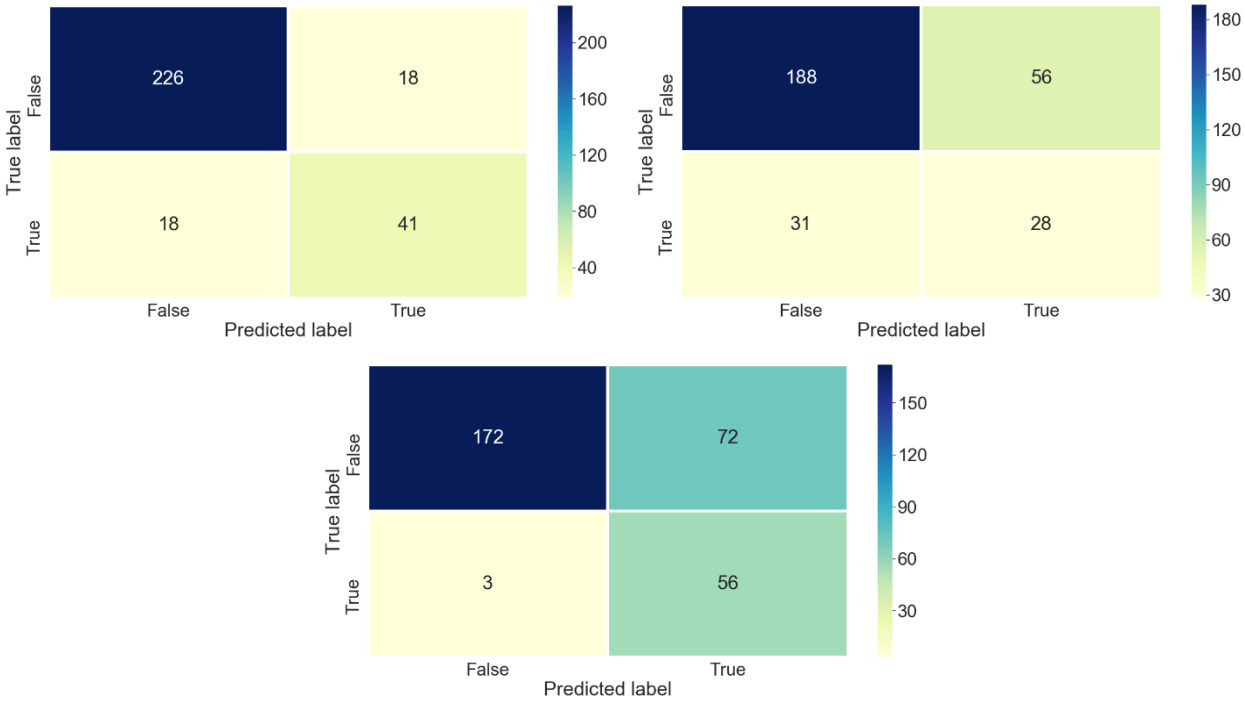

Figure 6-14. Confusion matrix resulting from applying the LCA (upper left), KNN (upper right) and QDA (lower centre) - Data quality prediction for flow patterns

In the same way of previous iteration, False Positive and False Negative were analysed to identify new possible features to enhance again the data-models.

Figure 6-15 compares False Negative with normal days (working days and public holidays). On day 327 (green line) or 391 (brown line) were observed a normal pattern, excepting the flows from step 10 to 30 and step 65 to 70, respectively. New features such as local minimum, maximum and accumulated may be useful to detect these abnormalities.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

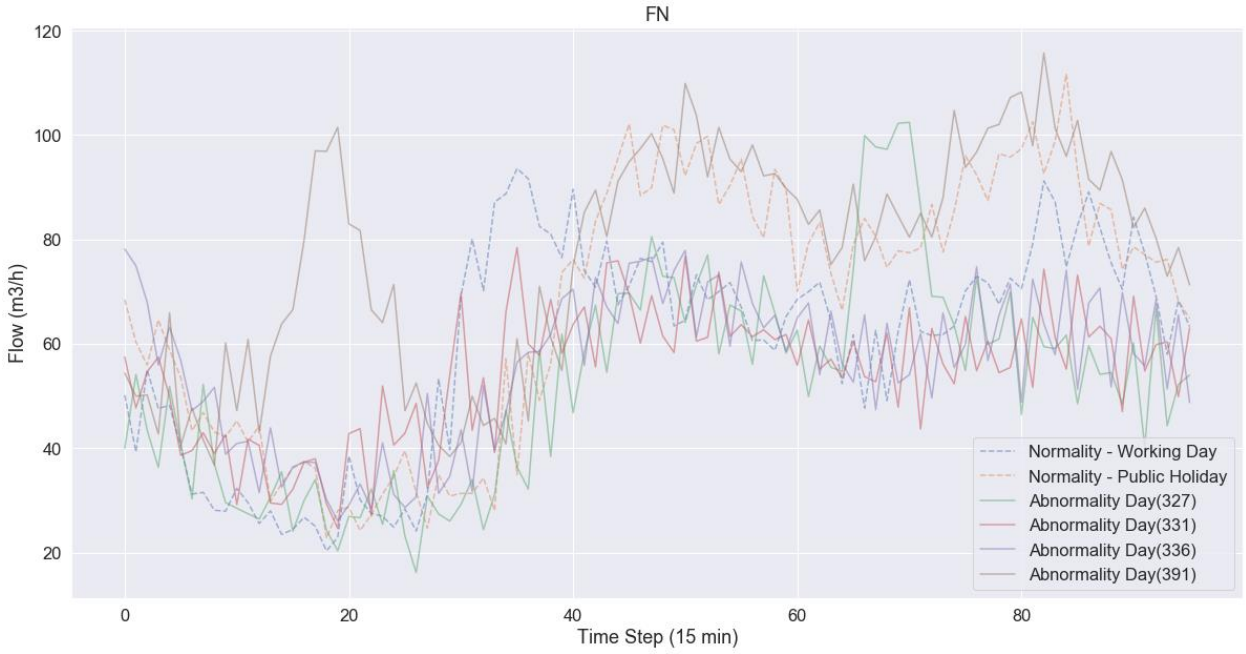

Figure 6-15. Flow rate from False Negative - Data quality prediction for flow patterns

The flow data distribution of each time windows is shown, one for each differential part of the day: one time slot from time step 0 to 30 (00:00 to 07:30) (see Figure 6-16), other from time step 30 to 70 (07:30 to 17:30) (see Figure 6-17) and the last from time step 30 to 96 (17:30 to 00:00) (see Figure 6-18).

Flow data distribution varies in line with the behaviour pattern of each time slot. Additionally, data distributions present visual differences between normality and abnormality. Therefore, local minimums, maximums, accumulated and percentiles may help to discern between normality and abnormality. For example, the data distributions related to the time series of day 327 (purple line), which is linked to abnormal day, are very similar to normality in the first two windows. Nevertheless, the data distribution of third window is totally different from normal due to its elongated (maximum value of flow). It is important to note that this maximum value was not representative until now, but it will be if local statistical measurements are applied.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

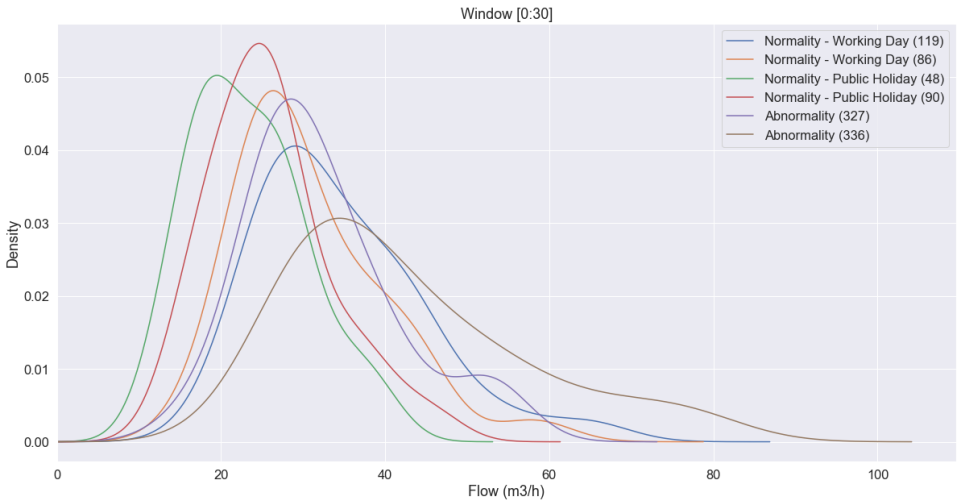

Figure 6-16. Data distribution from 00:00 to 07:30 (window 1) - Data quality prediction for flow patterns

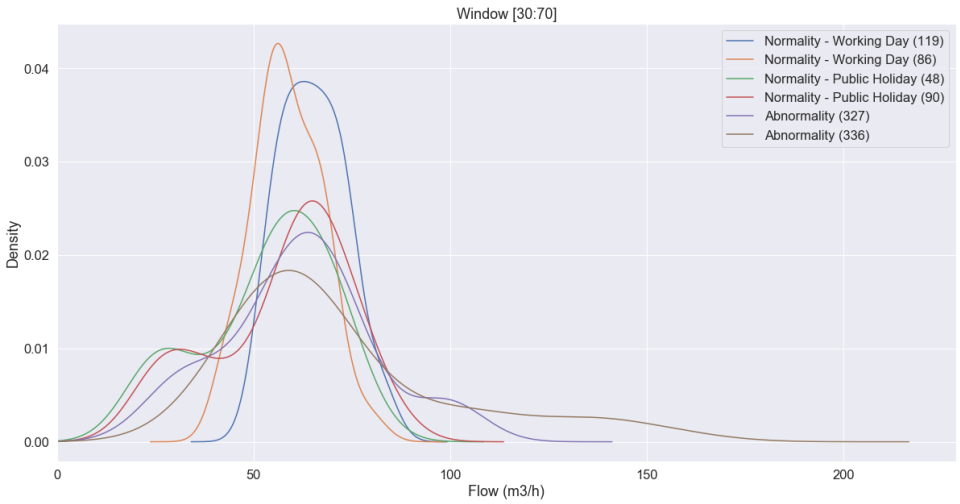

Figure 6-17. Data distribution from 07:30 to 17:30 (window 2) - Data quality prediction for flow patterns

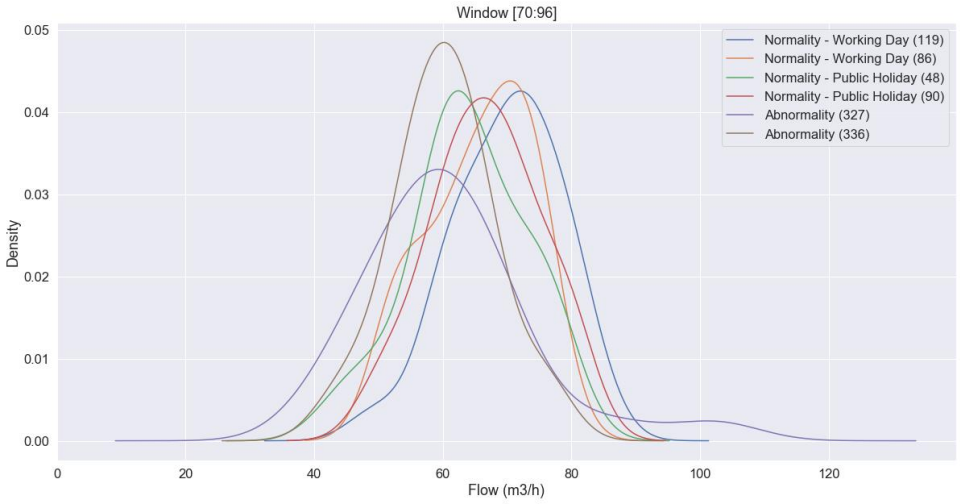

Figure 6-18. Data distribution from 17:30 to 00:00 (window 3) - Data quality prediction for flow patterns

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

To sum up, statistical measurements identified on iteration 1 and iteration 2, such as percentile 10, percentile 50, percentile 90, minimum, maximum and accumulated, are applied to three different windows on iteration 3, obtaining local statistical measurements which may increase classification capabilities of the data-driven model.

6.1.3. ITERATION 3

6.1.3.1. DATA PREPARATION

As it was concluded on iteration 3, the new hypothesis is based on including statistical local measurements as features. Therefore, percentile 10, percentile 50, percentile 90, minimum, maximum and accumulated were calculated for each window defined on the previous section.

Due to a large number of features (18 features, 6 for each window), Figure 6-19 only validates visually a set of them. Percentile 90 of window 1 on axis X, percentile 90 of windows 3 on axis Y and percentile 90 of windows 2 on axis Z are included in the graph. As the image shows, normal days (green points) were partially clustered and separated of the abnormal days (red points). Therefore, the new hypothesis is initially suitable and should be assessed through a model.

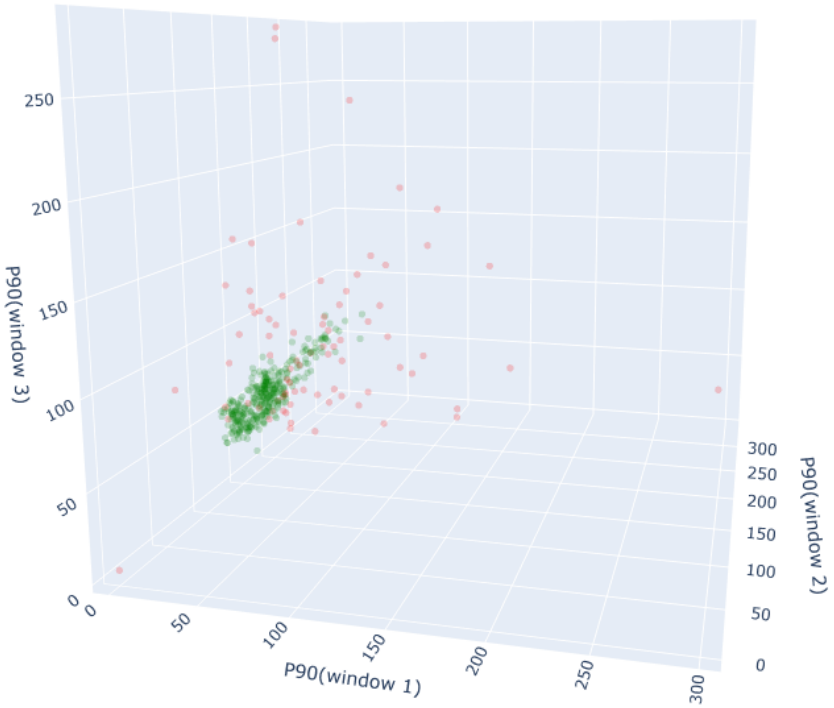

Figure 6-19. 3D visualization of percentile 90, maximum flow and accumulated flow of daily time series (green: normal points, red: abnormal points) - Data quality prediction for flow patterns

Finally, the data frame used to learn was defined as represented in Table 34.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 34. Data model used to learn - Data quality prediction for flow patterns

Feature	Description	Type
Percentile 10 of Flow on Window 1	The percentile 10 of the flow values gathered from 00:00 to 07:30 during a day	Float
Percentile 50 of Flow on Window 1	The percentile 50 of the flow values gathered from 00:00 to 07:30 during a day	Float
Percentile 90 of Flow on Window 1	The percentile 90 of the flow values gathered from 00:00 to 07:30 during a day	Float
Min on Window 1	The minimum of the flow values gathered from 00:00 to 07:30 during a day	Float
Max on Window 1	The maximum of the flow values gathered from 00:00 to 07:30 during a day	Float
Accumulated on Window 1	The accumulated of the flow values gathered from 00:00 to 07:30 during a day	Float
Percentile 10 of Flow on Window 2	The percentile 10 of the flow values gathered from 07:30 to 17:30 during a day	Float
Percentile 50 of Flow on Window 2	The percentile 50 of the flow values gathered from 07:30 to 17:30 during a day	Float
Percentile 90 of Flow on Window 2	The percentile 90 of the flow values gathered from 07:30 to 17:30 during a day	Float
Min on Window 2	The minimum of the flow values gathered from 07:30 to 17:30 during a day	Float
Max on Window 2	The maximum of the flow values gathered from 07:30 to 17:30 during a day	Float
Accumulated on Window 2	The accumulated of the flow values gathered from 07:30 to 17:30 during a day	Float
Percentile 10 of Flow on Window 3	The percentile 10 of the flow values gathered from 17:30 to 00:00 during a day	Float
Percentile 50 of Flow on Window 3	The percentile 50 of the flow values gathered from 17:30 to 00:00 during a day	Float
Percentile 90 of Flow on Window 3	The percentile 90 of the flow values gathered from 17:30 to 00:00 during a day	Float
Min on Window 3	The minimum of the flow values gathered from 17:30 to 00:00 during a day	Float
Max on Window 3	The maximum of the flow values gathered from 17:30 to 00:00 during a day	Float
Accumulated on Window 3	The accumulated of the flow values gathered from 17:30 to 00:00 during a day	Float

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Feature	Description	Type
Abnormality	Indicate if the timeseries is normal or abnormal. Key feature based on manual labelled.	Boolean

6.1.3.2. MODELLING & EVALUATION

This third iteration was also based on *LCA*, *KNN* and *QDA* algorithms, including local statistical measurements previously identified. Table 35 presents the confusion matrixes for these algorithms taking advantage of cross-validation.

Table 35. Recall and Precision results of the initial modelling - Data quality prediction for flow patterns

Algorithm	Recall Score	Precision Score
LCA	0.63	0.69
KNN	0.57	0.59
QDA	0.93	0.51

QDA and *KNN* improved the results of the first iteration. *QDA* presented the best results, reaching a *Recall Score* 0.93 and *Precision Score* 0.51. Therefore, the data-driven model was able to predict accurately all the abnormal days. Summarizing, the *LCA* predicted properly 191 normal days and 55 abnormal days (see Figure 6-20). Instead, 53 days were falsely predicted as abnormal days and 4 days as normal days.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

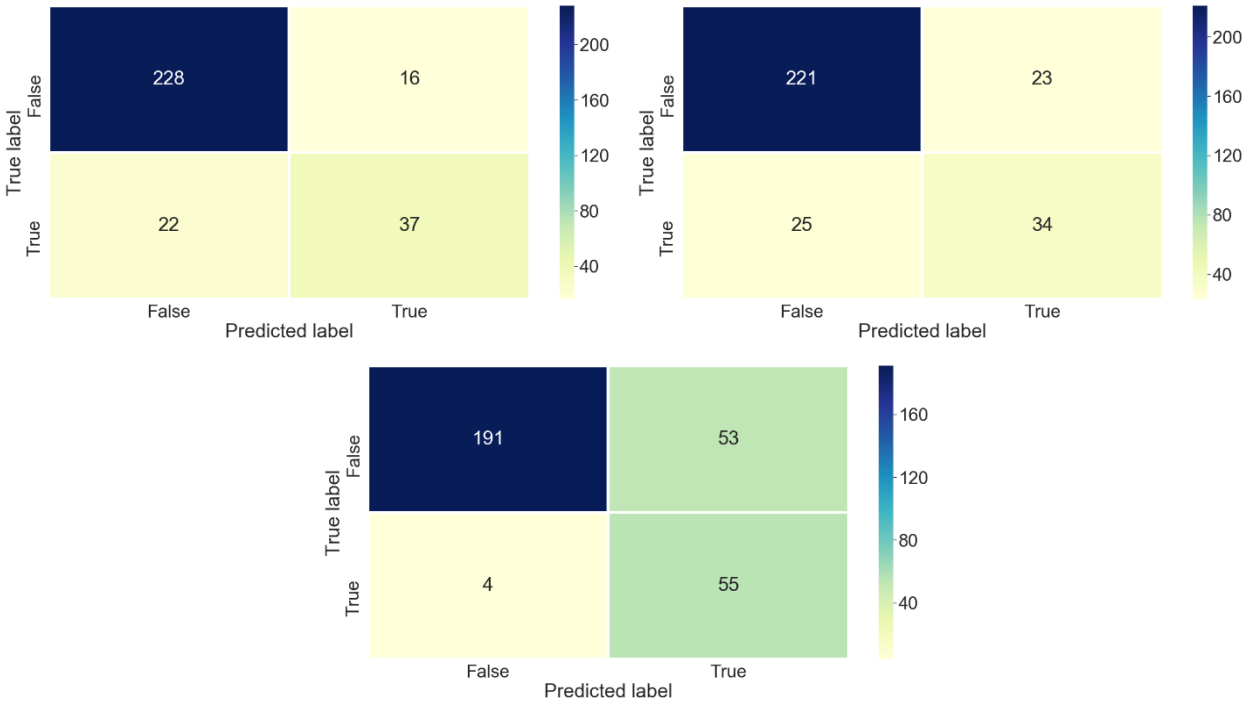

Figure 6-20. Confusion matrix resulting from applying the LCA(upper left), KNN (upper right) and QDA (lower centre) - Data quality prediction for flow patterns

Concerning the last iteration of the *QDA* cross-validation, the data-driven model reached 72 True Negative, 28 True Positive, 0 False Positive and 1 False Negative. Then, the results of the data-driven model are remarkable once it is trained with enough representative data sets.

To sum up, data-driven model built by using QDA algorithm demonstrates encouraging results, especially if it is trained with a sufficiently large and representative dataset. The optimization of hyperparameters could improve the current results, and hence it should be taken into account to a new iteration. Finally, it is important to remark that the data-driven model was trained with a dataset provided by Task 2.1 and hence, the data-driven models should be trained again to be deployed in a new location.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.2. EARLY DRIFT DETECTION

6.2.1. GENERAL DESCRIPTION

Sewer system monitoring using water quality sensors has become a significant step towards identifying issues in real-time related to possible pollution, blockages, or floods. Sensors capable of measuring water properties (e.g. turbidity, flow, pH, oxygen…) are used jointly with the pipes' physical properties to predict the different problems. To guarantee good measurements, sensors need supervision and output data handling provided by different procedures which include frequency-domain filtering and statistical evaluations. Our first part of the study develops a Machine Learning methodology to detect possible anomalous data gathered by a sensor. This solution will optimize the data gathering and reduce the sensor monitorization cost.

One of the most crucial obstacles is sensor drift, which is the increase or decrease of measurement error over time and a problem that a lot of sensors may be prone to. Each sensor has different behaviour before and during the drifting phase, meaning that usual threshold solutions do not offer enough accuracy and need a lot of checking to accomplish it. With the use of machine learning algorithms and enough historical drifting problems, a solution to predict the early stages of sensor drifting can be modelled, improving the detection of error lifetime.

6.2.2. ITERATION 1

6.2.2.1. BUSINESS UNDERSTANDING

The spectrometers gather the ultraviolet-visible spectroscopy (220 to 300 range) and calculate different water quality properties like the turbidity, COD, BOD, SST, oxygen, ammonium, etc… Like almost all sensors, when the spectrometers stay in touch with water that contains suspended solids, organic matter, nutrients, among others, the values gathered might become faulty. Not detecting (manually) on-time the drift of a sensor affects the data gathering and might create errors on sensor monitoring software. The business objective is to ensure the data quality by reducing the detection time of drift.

The AI objective in this use case is to analyse which is the behaviour of the drift and create a machine learning model capable of detecting this anomaly.

To detect the drift in a sensor, data of different spectrometers in different industrial plants were received. The next sections will go into the analysis of the data gathered and the creation of different data models which will be fitted into a machine learning model, to finally be evaluated.

6.2.2.2. DATA UNDERSTANDING

The received data is stored into two general directories, one for industrial data and the other for urban data. Industrial data directory contains datafiles from 16 different industries, and each industry has one to two months of data. Urban data directory contains 4 different sources, and each source can have between one to three months of data. Table 35 contains further details.

Table 27. Details about data sources – Early Drift Detection

Datasource	Location	Method used to acquire	Problems
Industry datafiles (16 different sources)	Local directory	Received by email	Features with number format is erroneous
Urban datafiles (4 different sources)	Local directory	Received by email	Features with number format is erroneous

As it can be seen in Table 45, the number of registers for industry is almost 4 times bigger than the urban dataset. It indicates the first exploration and experiments should be done on industry datafiles instead.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 28. General details about available data sources – Early Drift Detection

Data Source	Description	Format	# Registers	# Feature
Industry datafiles	Different directories (one for each industry), each one contains one/two months of data, and each day contains 350 registers	Directory with excel files	250000	30 to +200 (depending on the file)
Urban datafiles	Different directories (one for each urban region), each one contains one to three months of data, and each day contains 350 registers	Directory with excel files	84000	30 to +200 (depending on the file)

A spectrometer gathers the light spectrum with different wavelengths. Each wavelength has a specific domain meaning, so the behaviour differs between them slightly. Table 46 contains more details about the wavelengths.

Table 29. General details about available features – Early Drift Detection

Feature	Description	Type	UoM	Data Source
Nm190 to 750	Different wavelengths of the spectrum. They go from bigger values to lower values, and they end converted into water property values.	Number	nm	Each file in the directories

The analysis done understands which is the behaviour of the different spectres. Figure 6-21 shows a total month of data gathering in a certain industry. Each line represents a measurement, including multiple spectrums.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

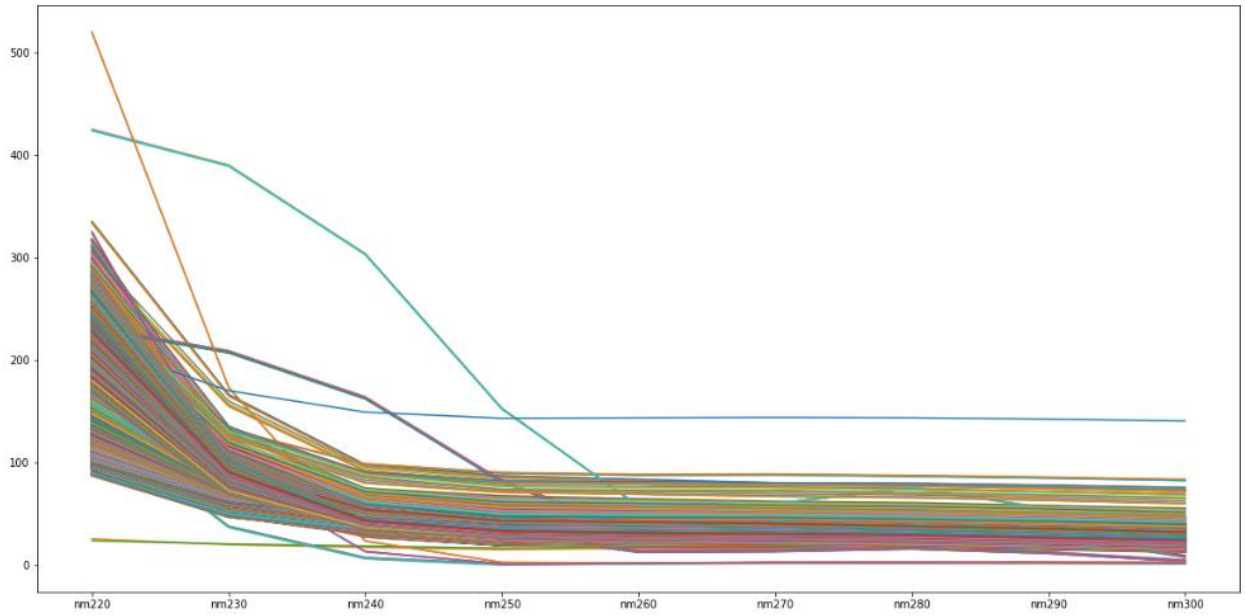

Figure 6-21. Spectroscopy data of industrial water – Early Drift Detection

The domain experts labelled the data in Figure 6-22 as anomalous. As can be seen, the deviation between each line is wide and the difference between the minimum and the maximum is also high. When a sensor is clean and maintained, the measurement of light absorbance is low. Then, light absorbance keeps increasing with the time flows due to fouling of the sensor, achieving a drift state like Figure 6-21. It is important to note that it affects mainly to the low spectre of light. The team has the idea of building this state into a data model that can be fitted into an algorithm.

As can be seen, the low spectres have bigger values than the high spectres, but a distribution plot will show way better the different distributions.

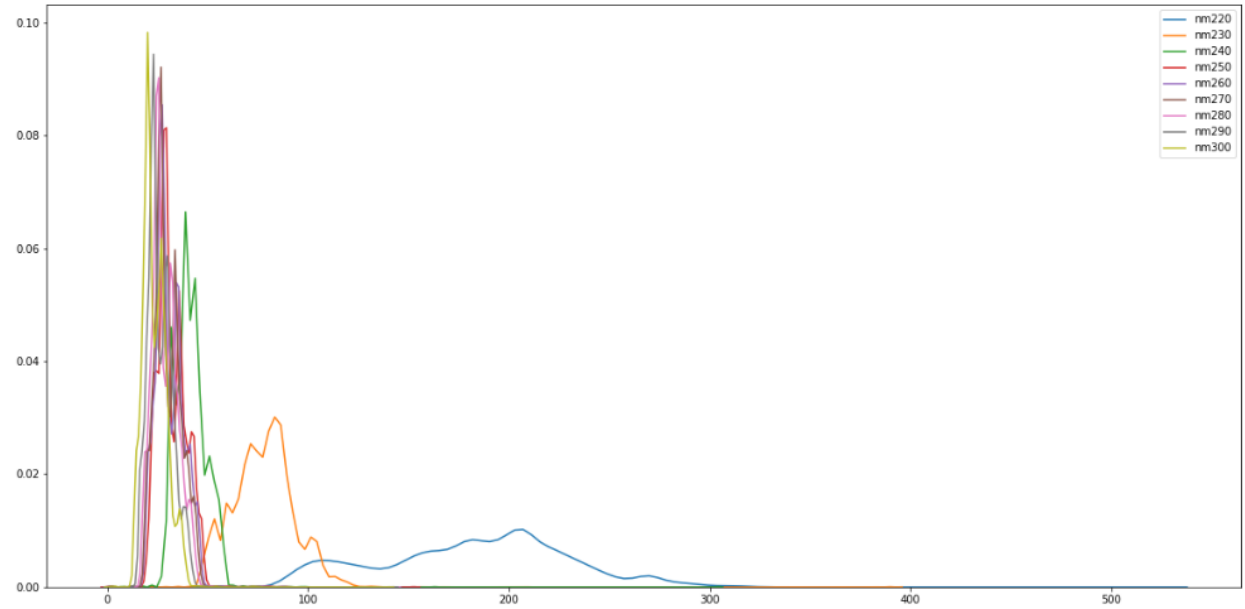

Figure 6-22. Distribution difference between wavelengths – Early Drift Detection

As the spectres gets bigger, the value scale decreases, so it is important to have into account these values need to be standardized before training a model.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.2.2.3. DATA PREPARATION

In this iteration, the data preparation did not require further analytics on the time series data. The idea behind the data model was to take representative wavelengths, calculate metrics (see Table 36) that define the different properties in a wavelength (the shape of the data) and use them to fit a model.

The wavelengths used ranged from 220 to 300. Some datasets contain more wavelengths than others, but this was the minimum gathered in all datasets and the team wanted a solution that could work on every dataset. The final data model was as defined in Table 36.

Table 36. Data model used to learn - Early Drift Detection

Feature	Description	Type
Wavelength mean	The median of the values gathered by a wavelength in a time period	Float
Wavelength quantile 0.2	Quantile 0.2 of the values gathered by a wavelength in a time period	Float
Wavelength quantile 0.8	Quantile 0.8 of the values gathered by a wavelength in a time period	Float
Slope	Slope of the gathered register by all wavelengths	Float
Wavelength Median	The median of the values gathered by a wavelength in a time period	Float

6.2.2.4. MODELLING & EVALUATION

In this first iteration, the evaluation metrics used were usual, is the accuracy, precision and recall (see Annex 2 for more detailed information about the scoring metrics). The team did not want to include complex metrics since it was the first iteration and the number of observations in this data model was low. The models used were Stochastic Gradient Descent (SGD) classifier (with an SVM), Logistic Regression, Perceptron, Feed Forward Neural Network (FFNN), K-Nearest Neighbours (KNN), Extra Tree, AdaBoost, Random Forest and Gradient Boosting.

To train and evaluate the models, a simple split of 70% train and 30% test was used, and for each model, the sets were the same. Table 37 shows the results:

Table 37. Accuracy, Recall and Precision results of the initial modelling – Early drift detection

Algorithm	Accuracy Score	Precision Score	Recall Score
Logistic Regression	0.4	0.2	0.25
SGD	0.4	0.42	0.42
Perceptron	0.4	0.2	0.5
FFNN	0.4	0.25	0.33
KNN	0.4	0.2	0.5
AdaBoost	0.2	0.5	0.4

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Algorithm	Accuracy Score	Precision Score	Recall Score
ExtraTree	0.42	0.42	0.4
Random Forest	0.4	0.2	0.5
Gradient Boosting	0.4	0.2	0.5

As can be seen, the result of the predictions by the different algorithms is poor. All the metrics are low, so the data model provided is not good enough. The team suspects one of the main problems is the lack of registers after modelling the data. There are a lot of registers in the raw data, but the team uses a register for each industry, meaning only 17 registers remain.

The next iteration will show how the team built a completely different data model, but one of the future aims is to work more on the data model for this iteration, trying to improve the idea.

6.2.3. ITERATION 2

6.2.3.1. DATA UNDERSTANDING

After iteration 1, where all wavelengths were used, the team wanted to predict drift only using one wavelength that could provide impact to the model decisions. In a domain aspect, the higher wavelengths offer more statistical explainability when trying to detect drift, Figure 6-21 shows a bigger variability in lower wavelengths than in higher ones.

The first step was to identify the different drift zones and prepare a data model considering the differential properties of the time series. The Figure 6-23 shows an example of high drift, where the values get high between the register 11000 to the register 21000, where cleaning is applied at register 14000 but the value increases again at 15500.

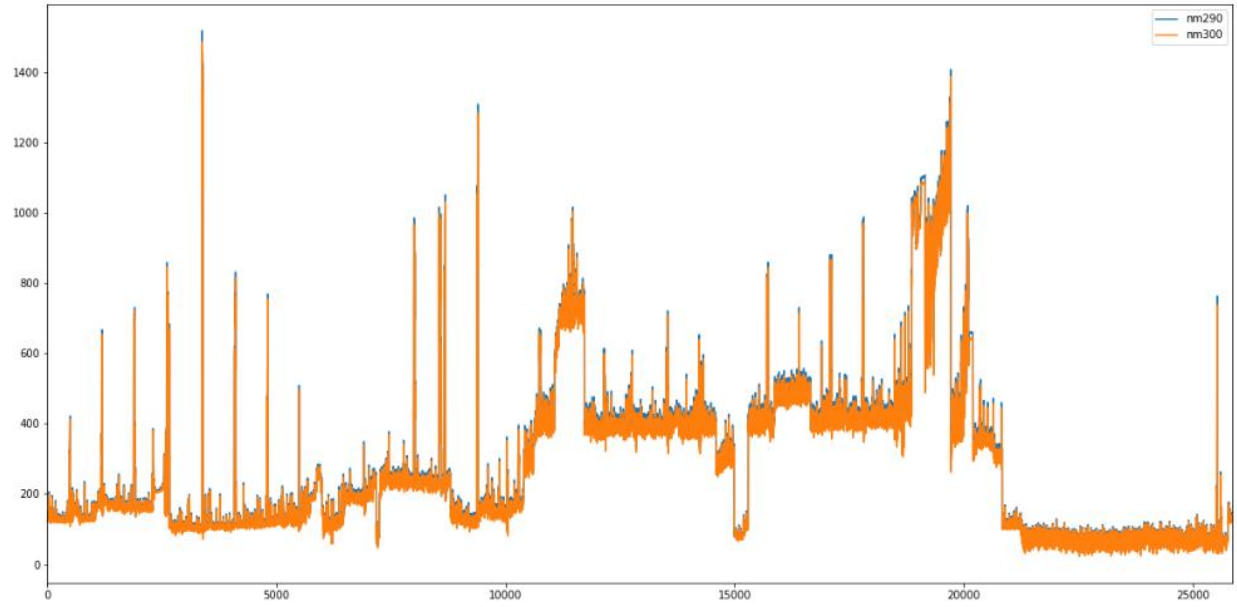

Figure 6-23. Wavelengths 290 and 300, drift example over a long time – Early Drift Detection

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The shown contextual anomaly happens in some parts of all the industrial data gathered. Each anomaly is labelled since the beginning of drift appearance until the maintenance is done. After labelling, the team decided to compare the distribution of normal behaviour and anomalous behaviour, shown in the Figure 6-24.

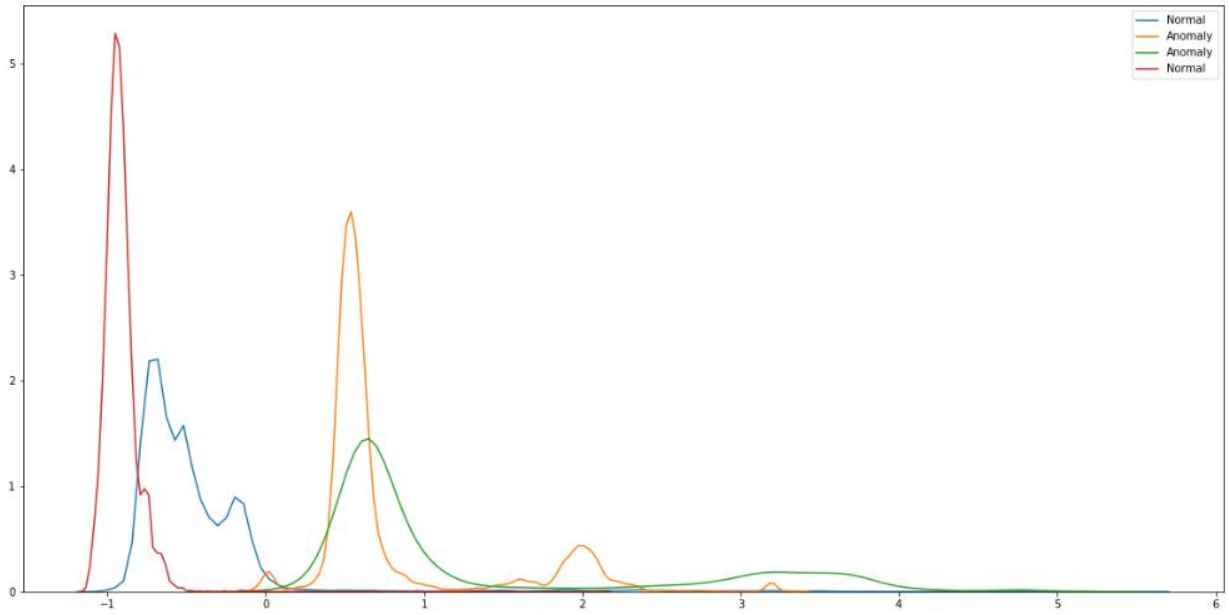

Figure 6-24. Data distribution of drift and normal behaviour – Early Drift Detection

The anomaly windows have higher values than the normal context. The plot shown is an extremist case and we need to consider that the beginning of drift is difficult to detect but having this dispersion is important.

6.2.3.2. DATA PREPARATION

The design of the data model was not complex. Since the difference between distributions could be seen using the naked eye, the features used should represent the shape of it. The Table 38 shows the features extracted from a rolling window in the time-series data.

Table 38. Data model used to learn - Early Drift Detection

Feature	Description	Type
Wavelength mean	The median of the values gathered by a wavelength in a time period of 30 points	Float
Wavelength quantile 0.2	Quantile 0.2 of the values gathered by a wavelength in a time period of 30 points	Float
Wavelength quantile 0.8	Quantile 0.8 of the values gathered by a wavelength in a time period of 30 points	Float
Wavelength minimum	Minimum value in a time period of 30 points	Float
Wavelength Median	The median of the values gathered by a wavelength in a time period of 30 points	Float

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The objective variable for the data model was if the time window was drifted or not, being boolean. The final size of the data model is 200000 rows and 180 columns.

6.2.3.3. MODELLING & EVALUATION

The problem at hand suggested the creation of a model that could detect drift in real-time, so the metric used to evaluate the model is going to be the Numenta Anomaly Benchmark (NAB) (see Annex 3).

As in other iterations and use cases, the selection of the best model involves testing different algorithms and different hyper parameters, then select the best one. On the Table 39, a set of models tested is shown with metric result for NAB with a 0.5 balance and precision apart.

Table 39. Numenta and Precision Score of the model – Early drift detection

Algorithm	Numenta Score	Precision Score
SDG	0.92	0.97
Logistic Regression	0.89	0.92
Perceptron	0.90	0.93
FFNN	0.98	0.97
KNN	0.97	0.95
ExtraTree	0.94	0.71
AdaBoost	0.97	0.95
RandomForest	0.97	0.89
GradientBoosting	0.95	0.87

The results achieved by all the algorithms tested are high, meaning the data model is good. Of all algorithms tested, feed forward neural networks work the best and have not only good NAB score, by the balance between precision and the primitive NAB score is balanced.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

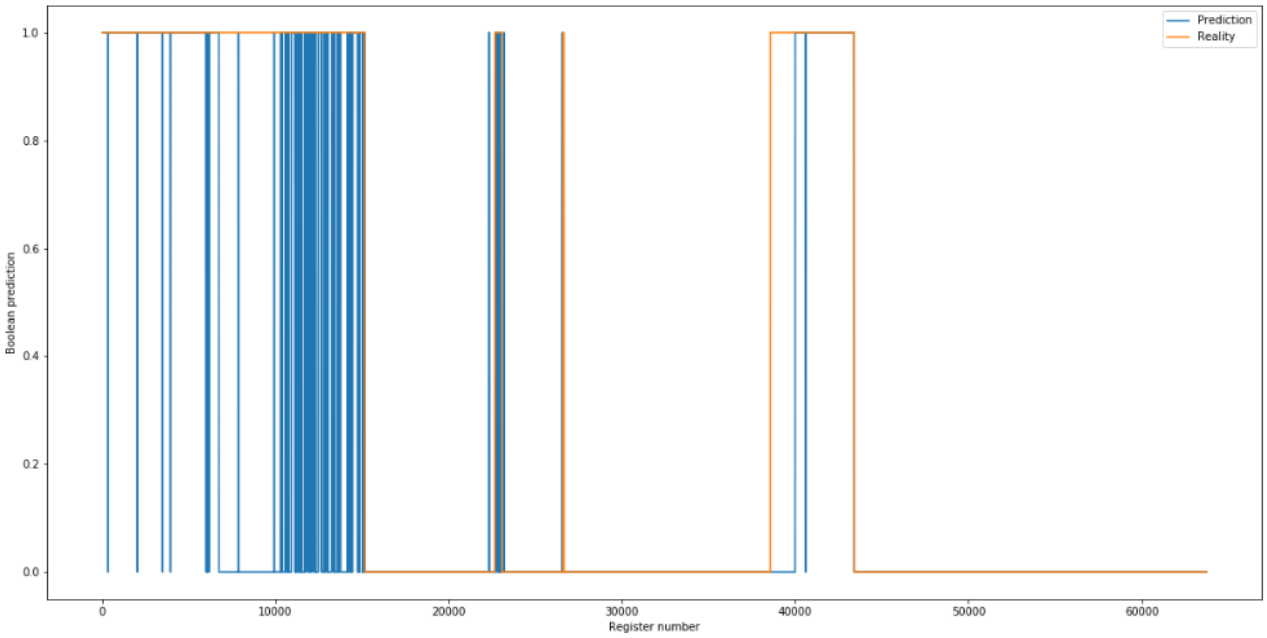

Figure 6-4. Comparison between predictions and reality

Figure 8-4 shows the comparison between the predictions done by a FFNN and the real data. As it can be seen, for most cases the anomaly is predicted early on and the number of false positives is low, only around register number 22000.

Future work needs to include residual water in an urban environment and not from industrial environment. This study can lead to fast iterations when the urban data is received in the future.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.3. GENERIC ANOMALY DETECTION

Water sensors suffer from a wide range of anomalies when gathering data as drift, high noise or instant increase/decrease in the gathering. As drift was one of the crucial points to be detected, it is also important to detect other anomalies. The team studied supervised and unsupervised approaches in two iterations to detect irregularities in the data as an early warning solution.

6.3.1. ITERATION 1

6.3.1.1. BUSINESS UNDERSTANDING

The first iteration has the AI objective of detecting water quality pattern anomalies in real-time by using supervised models. This will improve the water monitoring, resource management and will reduce the risk error of a business. The Figure 6-25 shows some of the most important anomalies, being an increase of value that has not been detected.

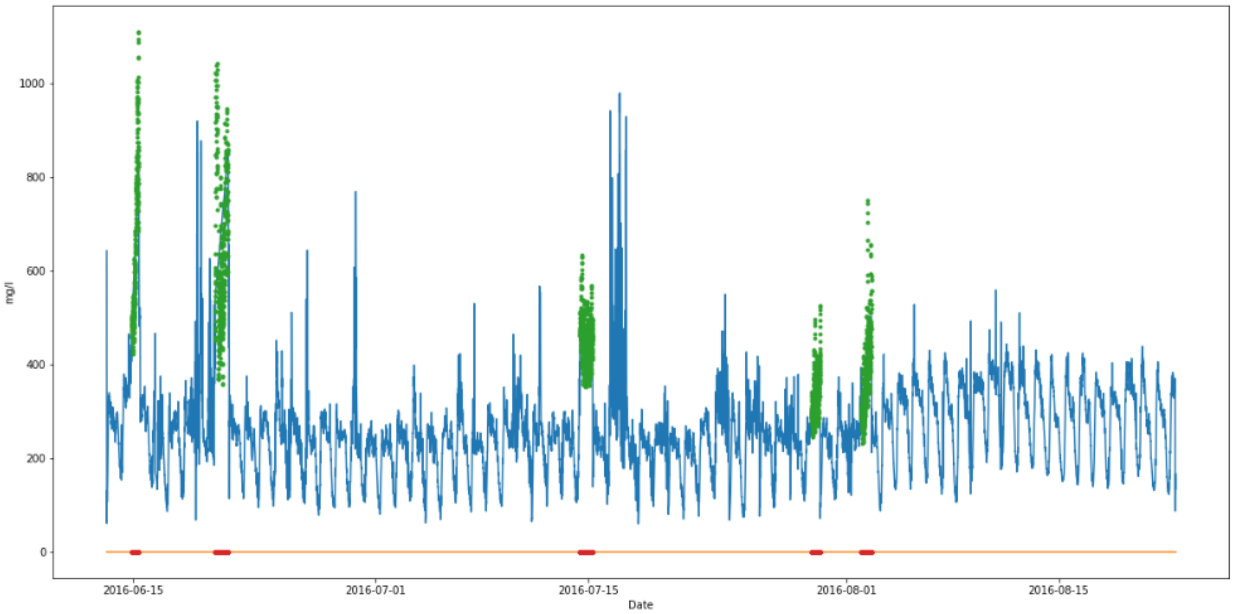

Figure 6-25. Labelling of some of the anomalies to detect – Generic anomaly detection

To solve the business objective, the work with univariate detection is going to be prioritized, so the detection of anomalies can be done without having other sensors.

6.3.1.2. DATA UNDERSTANDING

The received dataset contains data gathered by a spectrometer in a WWTP, converted into water quality values. Table 40 shows more details about the gathered data file.

Table 40. Details about data sources – Generic anomaly detection

Datasource	Location	Method used to acquire	Problems
Anomaly examples	Local directory	Received by email	Contains only two months of data

Table 41 explains it has two months of data with a size of 125000 registers, each register gathered in a 2-minute frequency, and 6 features.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 41. General details about available data sources – Generic anomaly detection

Data Source	Description	Format	# Registers	# Feature
Anomaly examples	Contains different time series, for different water quality elements	Excel file	125000	6

The data set contains 6 different water quality elements gathered. Table 42 contains further details on each feature and unit of measure.

Table 42. General details about available features– Generic anomaly detection

Feature	Description	Type	UoM	Data Source
COD	The chemical oxygen demand, an indicative measure of the amount of oxygen that can be consumed by reactions in a measured solution.	Number	mg/l	Anomaly examples
BOD	The biochemical oxygen demand is the amount of dissolved oxygen needed by aerobic biological organisms to break down organic material present in each water sample at certain temperature over a specific time period.	Number	mg/l	Anomaly examples
TSS	Total suspended solids are the dry-weight of suspended particles that are not dissolved in a sample of water.	Number	mg/l	Anomaly examples
Temperature	The temperature of the water sample.	Number	Centigrade	Anomaly examples
NH4-N	The ammonium concentration in the water sample.	Number	mg/l	Anomaly examples
pH	A measure of how acidic/basic water is, ranging between 0 to 14 and 7 being the neutral case.	Number	-	Anomaly examples

To explore the data, a set of steps have been done to understand the meaning of the values and the relationship between time-series. First, the dataset contains 6 variables which are of float type, each one having a specific behaviour that needs to be studied, so the first step is to calculate and plot the autocorrelation of each variable.

The COD autocorrelation plot (see Figure 6-26 and Annex 1 for more detailed information about autocorrelation concept) shows high correlation the first lag points, with a decreasing correlation. Around the lag 100, 200 minutes, the autocorrelation is low, meaning each value can be represented by the past 200 minutes.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

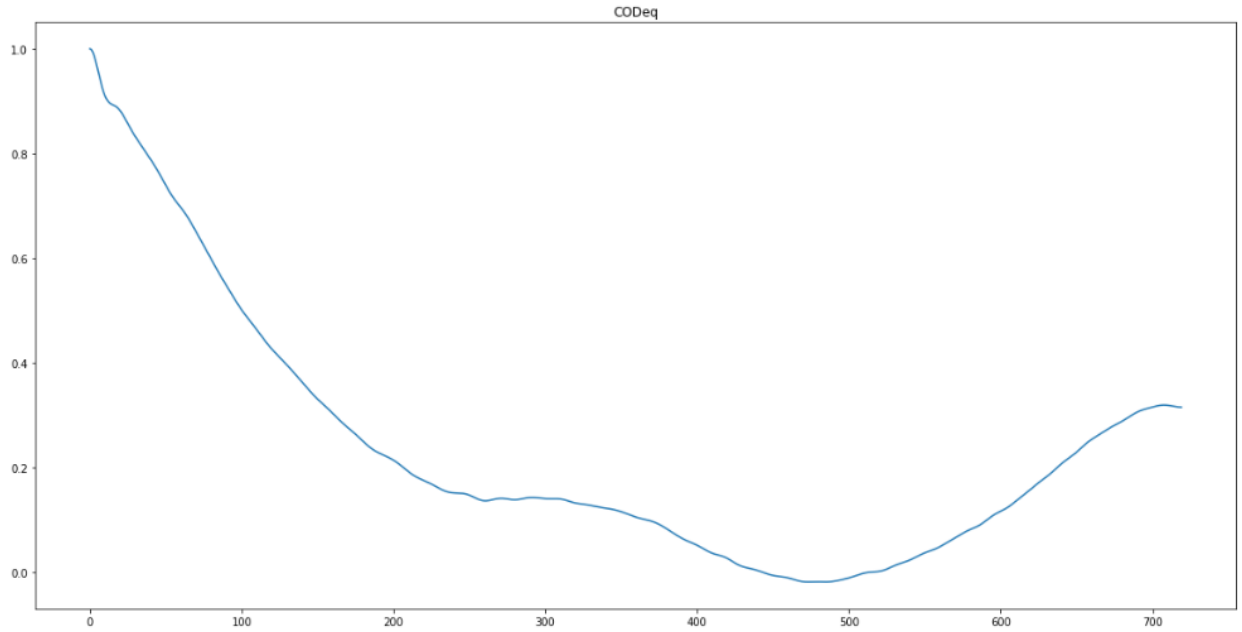

Figure 6-26. COD autocorrelation – Generic anomaly detection

The BOD autocorrelation shows different behaviour (see Figure 6-27). The correlation on the early lags is the same as the COD autocorrelation, but the latest lags have a bigger increase describing a daily seasonality of the BOD values.

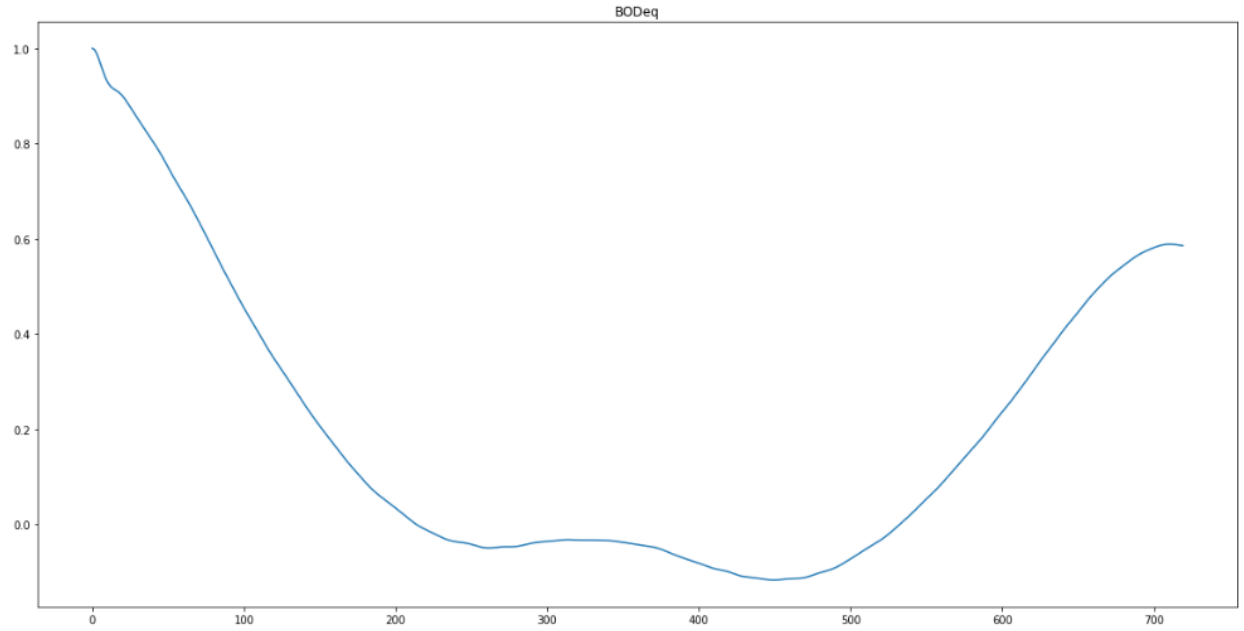

Figure 6-27. BOD autocorrelation – Generic anomaly detection

TSS autocorrelation shows high correlation on early lags with a decrease until lag 80, where the lags are not meaningful enough (see Figure 6-28)

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

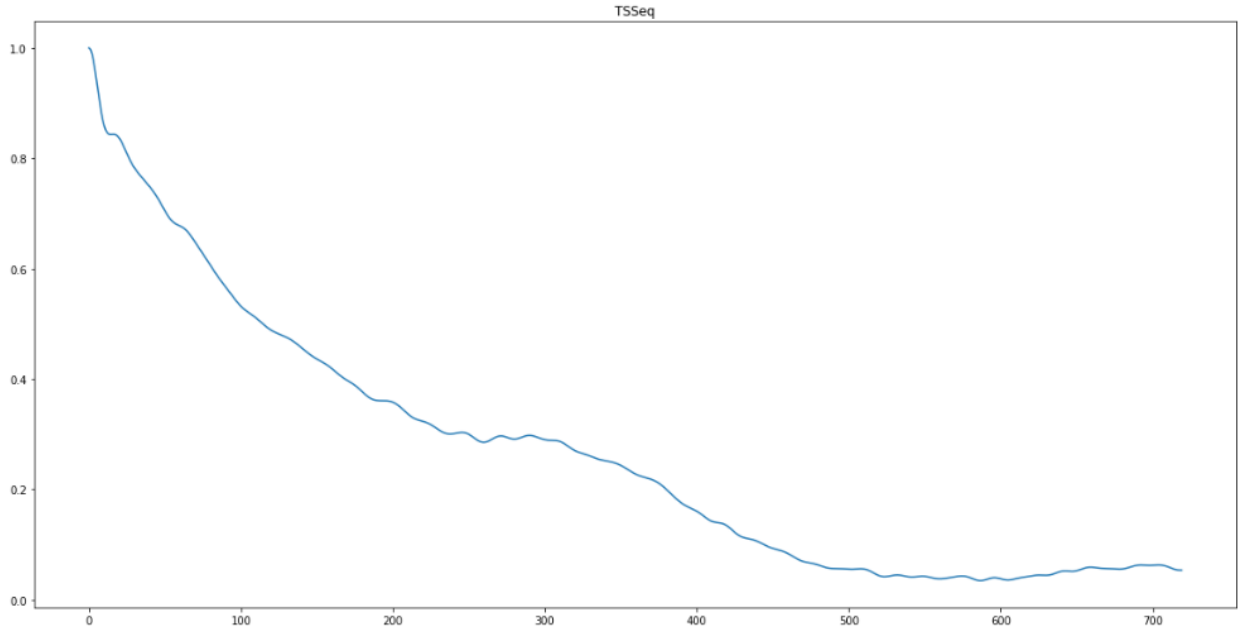

Figure 6-28. TSS autocorrelation – Generic anomaly detection

NH_4-N presents a uniform decreasing autocorrelation in all lags, being a meaningful correlation until 200 lags (see Figure 6-29).

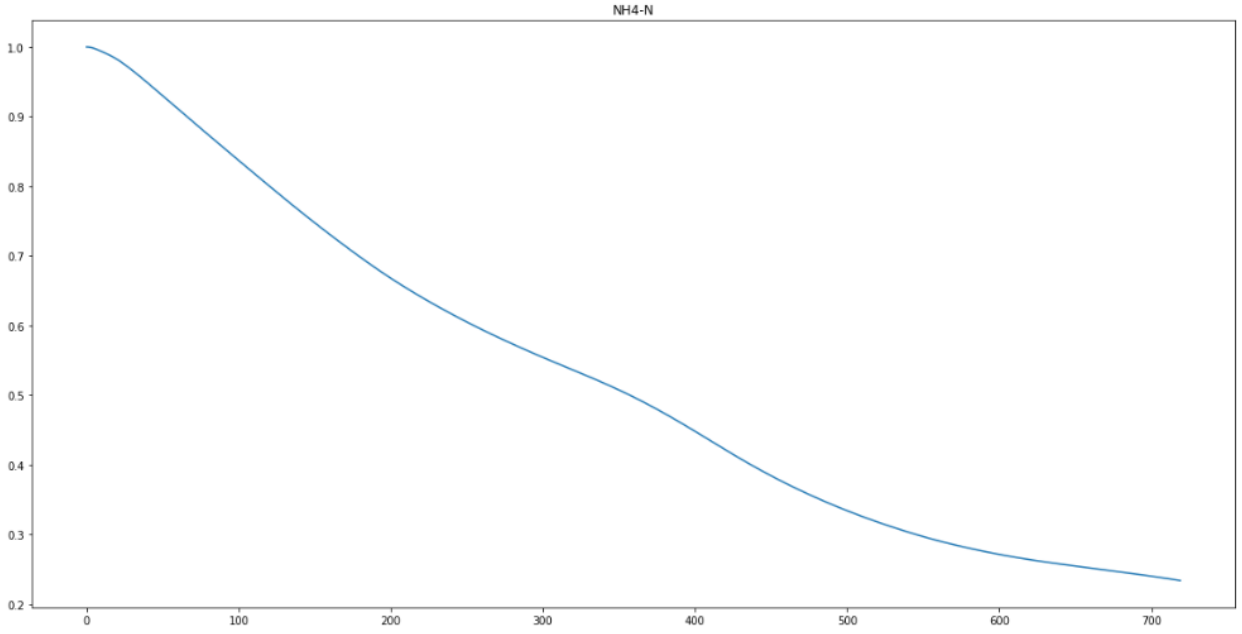

Figure 6-29. NH_4-N autocorrelation – Generic anomaly detection

pH autocorrelation plot (see Figure 6-30) shows how correlated the pH is between the past lags. The correlation is so big that the decrease is not hard.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

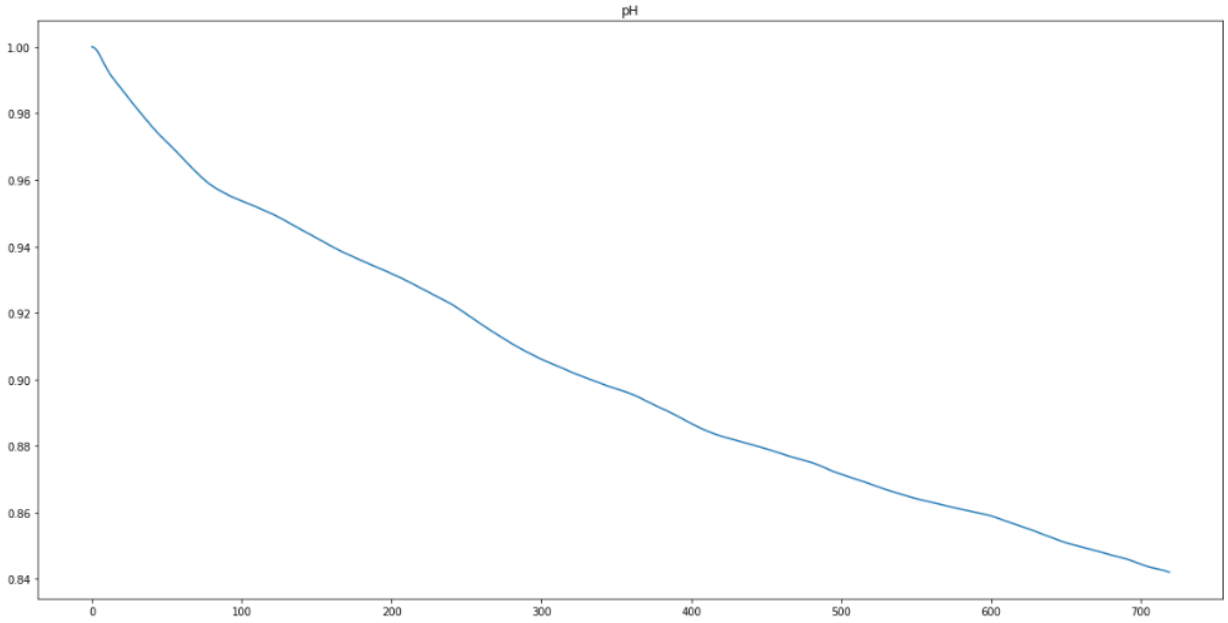

Figure 6-30. pH autocorrelation – Generic anomaly detection

After knowing, which is the autocorrelation for all the variables, the correlation between them is analysed in Figure 6-31.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

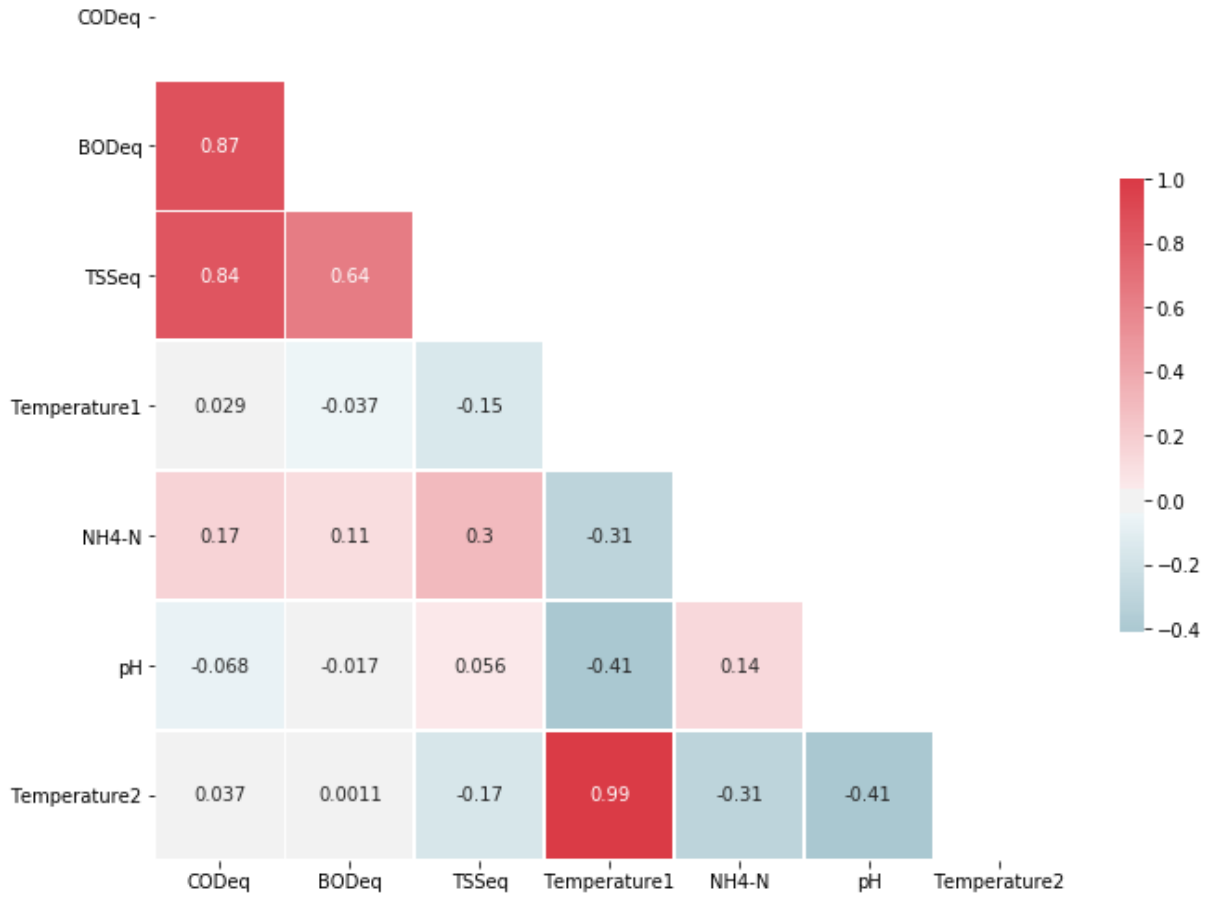

Figure 6-31. Correlation matrix – Generic anomaly detection

The matrix shows a high correlation between the COD-BOD-TSS while there is a low correlation with the other values, except the correlation between the temperature and NH_4-N and temperature and pH.

To end with the first data exploration, the time series were decomposed, and the vital parts of the time series were evaluated. The Figure 6-32 shows how once seasonality is removed, the trend is high in some subsets of the time series. Additionally, the residual subplot shows parts in the time series with high noise.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

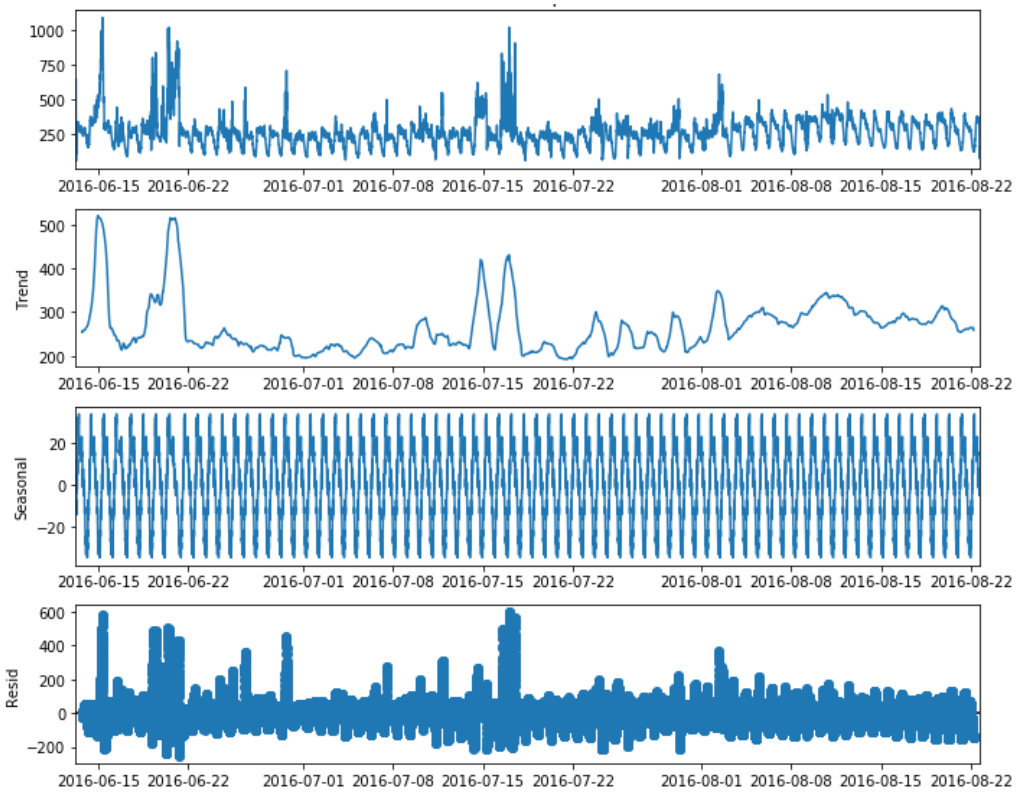

Figure 6-32. Decomposition of COD – Generic anomaly detection

Finally, before starting the data preparation and design of the data model, a univariate distribution was done. The Figure 6-33 shows 4 different distributions, one contains normal behaviour and the other three are anomaly behaviour. It can be seen how two of the anomalies have a bigger window, but another one is near the normal behaviour. Despite of similar pattern, statistical distribution measurements, like Q1, median, Q3, differs totally.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

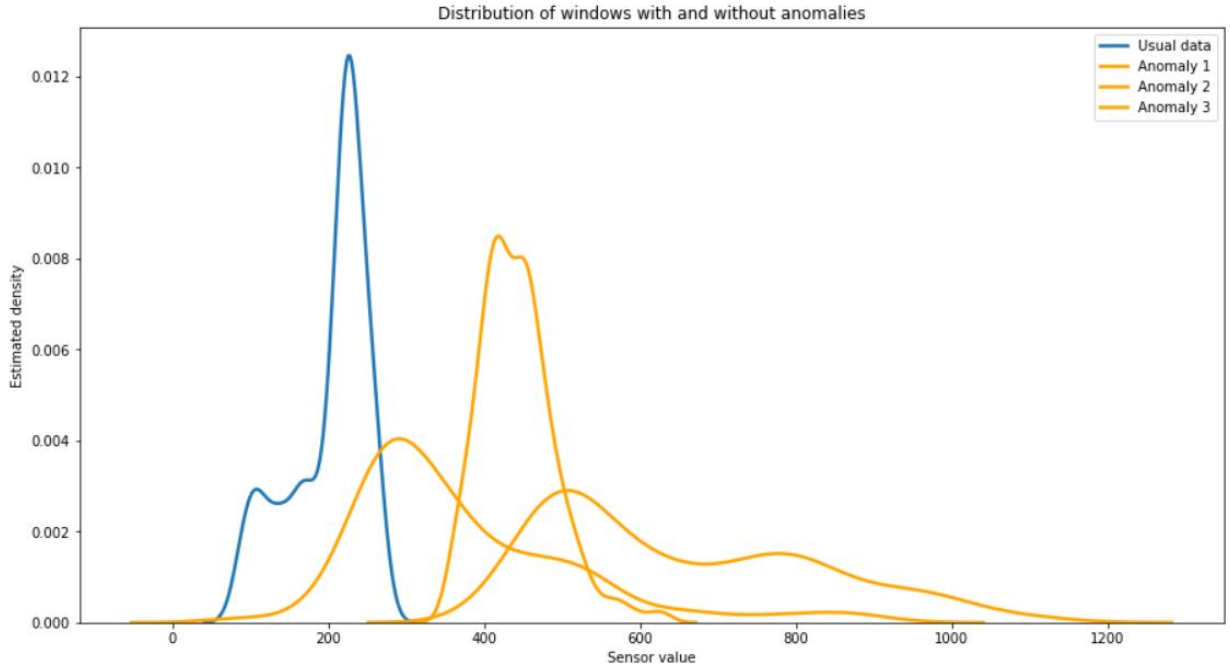

Figure 6-33. Distribution comparison between anomalies and normal behaviour – Generic anomaly detection

6.3.1.3. DATA PREPARATION

As specified before, the data-modelling should work on the larger number of sensors possible and in real-time. The metrics used to differentiate anomalous and normal behaviour are extracted from a sample window of the total time-series, which size is decided based on the past analysis. Each register will contain the different metrics, and each one will contain a sample between the register timestamp and the past hours. In Figure 6-33, the anomalous distributions contain higher variance, mean, and quantile values, so ensuring correct estimations is crucial. Table 37 explains the features of the designed data model.

Table 43. Data model used to learn – Generic anomaly detection

Feature	Description	Type
Mean	Mean of the past 30 registers in a certain point of time.	Float
Variance	Variance of the past 300 registers in a certain point of time.	Float
Trend	Trend of the past 100 registers in a certain point of time.	Float
Mean lags, 1 to 9	After calculating the mean feature, each register contains the past 9 “mean feature”.	Float
Variance lags, 1 to 9	After calculating the variance feature, each register contains the past 9 “variance feature”.	Float
Trend lags, 1 to 9	After calculating the trend feature, each register contains the past 9 “trend feature”.	Float

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.3.1.4. MODELLING & EVALUATION

The modelling and evaluation phase consists of testing a batch of algorithms and evaluate them using NAB. During this iteration, the team used the time series cross-validation technique to secure the algorithms being tested have consistency and can generalize well. Out of all the algorithms, Adaboost had the best result with 0.78 NAB (with a 0.5 precision weight). Using a 0.75 precision weight in the NAB metric (which rewards the early prediction of anomalies) and a Grid Search to optimize hyper-parameters, Adaboost achieved 0.75 NAB. The Figure 6-34 shows the different iterations of the grid search.

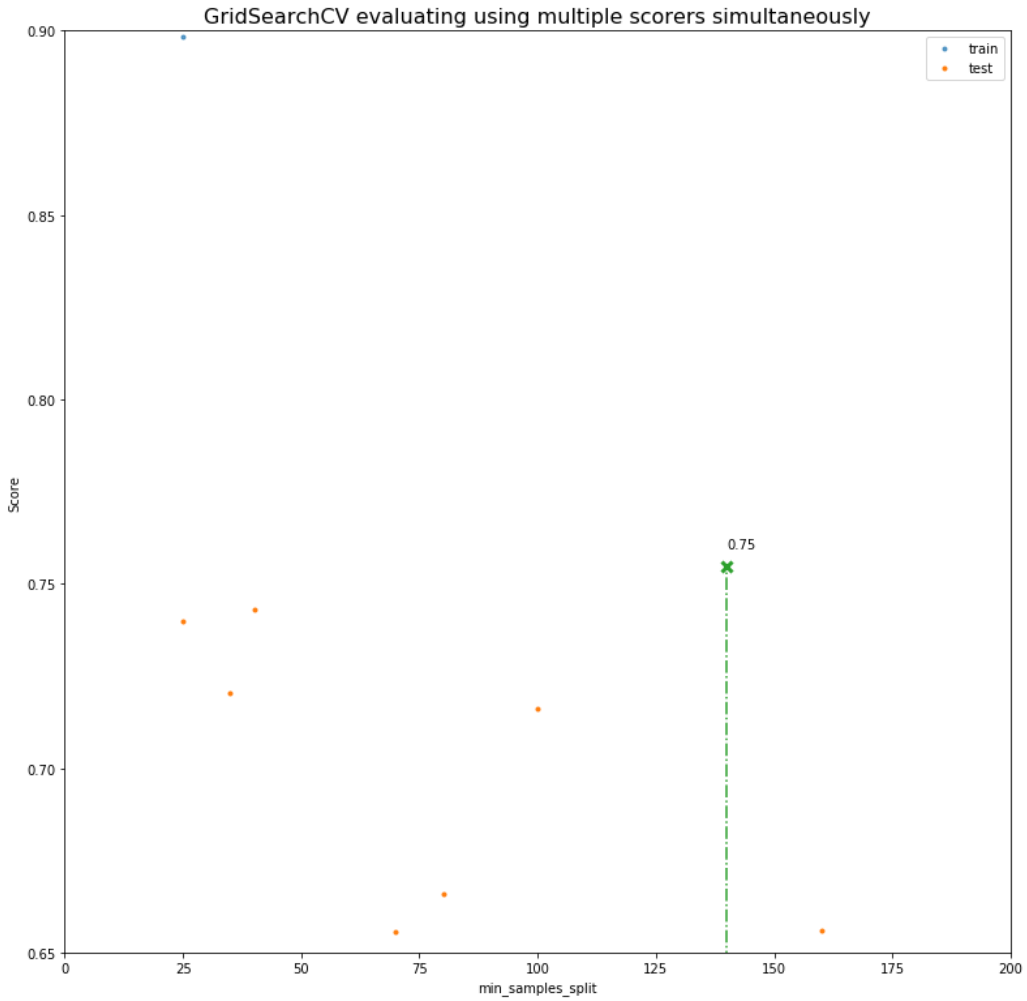

Figure 6-34. Grid search cross validation of the Ada Boost optimization – Generic anomaly detection

To know visually which points are detected by the algorithm, a plot was created showing which points of the time series were predicted as an anomaly, shown in Figure 6-35. To compare, Figure 6-25 contains the labelling of different anomalies, and the ones being shown in Figure 6-35 are the first anomalies in Figure 6-25.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

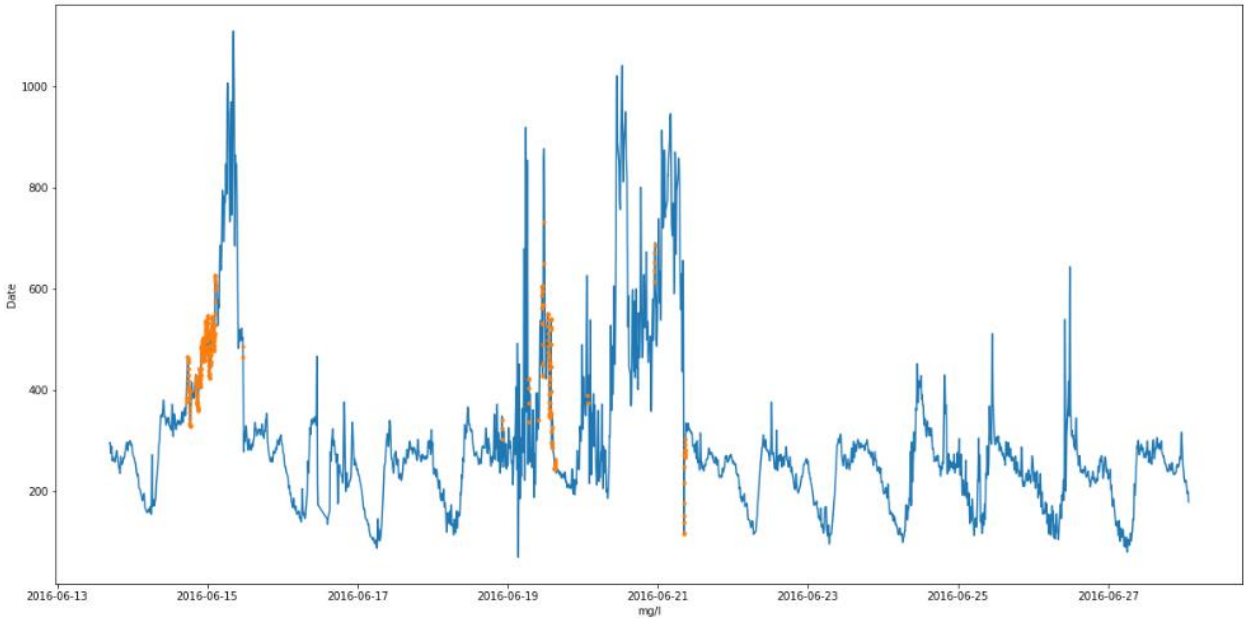

Figure 6-35. Points detected by the algorithm – Generic anomaly detection

Most of the predictions were done early in the anomalies and the false positives are not far from the anomaly or are just after the anomaly correction. The team concludes on having a good model that can predict this type of anomalies.

Finally, to know the importance of each variable, the team plotted the weight of the features of the model and extract conclusions.

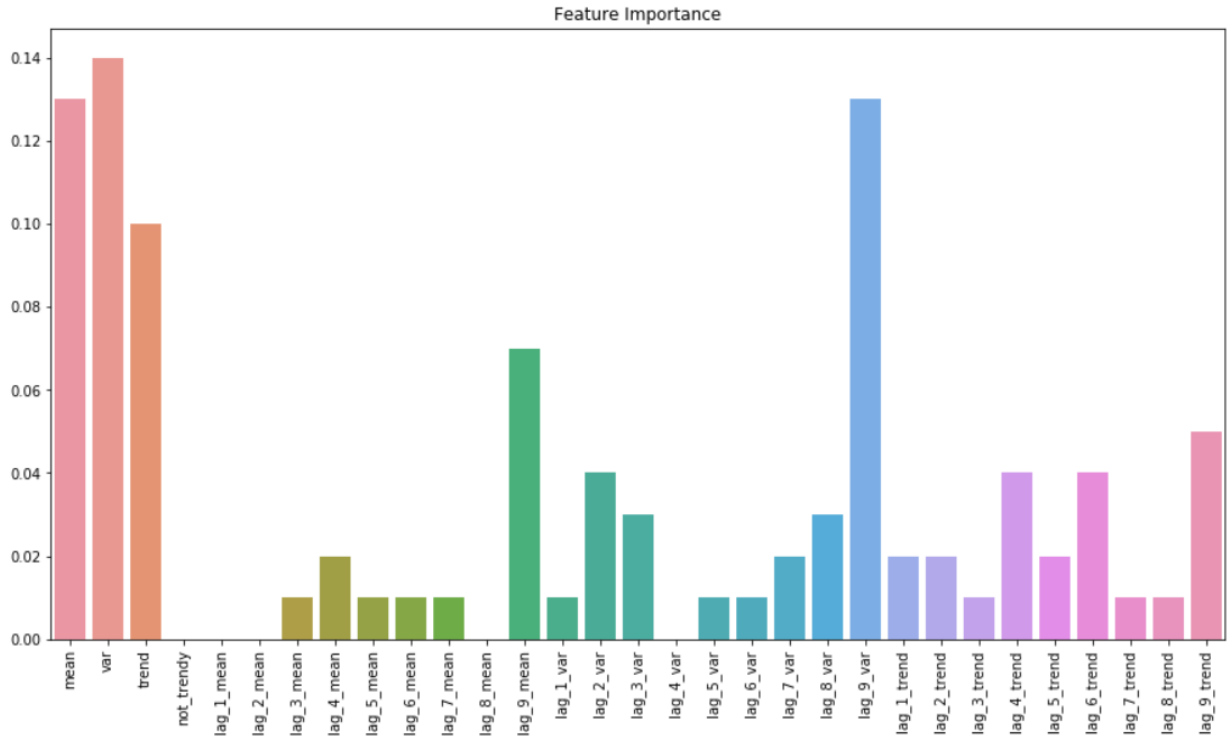

Figure 6-36. Feature importance of the Ada Boost algorithm – Generic anomaly detection

The main thing extracted from the plot is that the bigger lags have more importance than the lowers, but the point at the moment also has a lot of importance. In future iterations, the lag frequency should be increased and try if the models work better.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.3.2. ITERATION 2

6.3.2.1. DATA PREPARATION

The past iteration built a model that predicted high-increase anomalies. While it solves detecting the most critical anomalies, it does not consider high noise or really low deviation problems. The Figure 6-37 shows normal behaviour, the anomalies predicted in the past iteration (in red) and other types of anomalies in yellow (which were not predicted before).

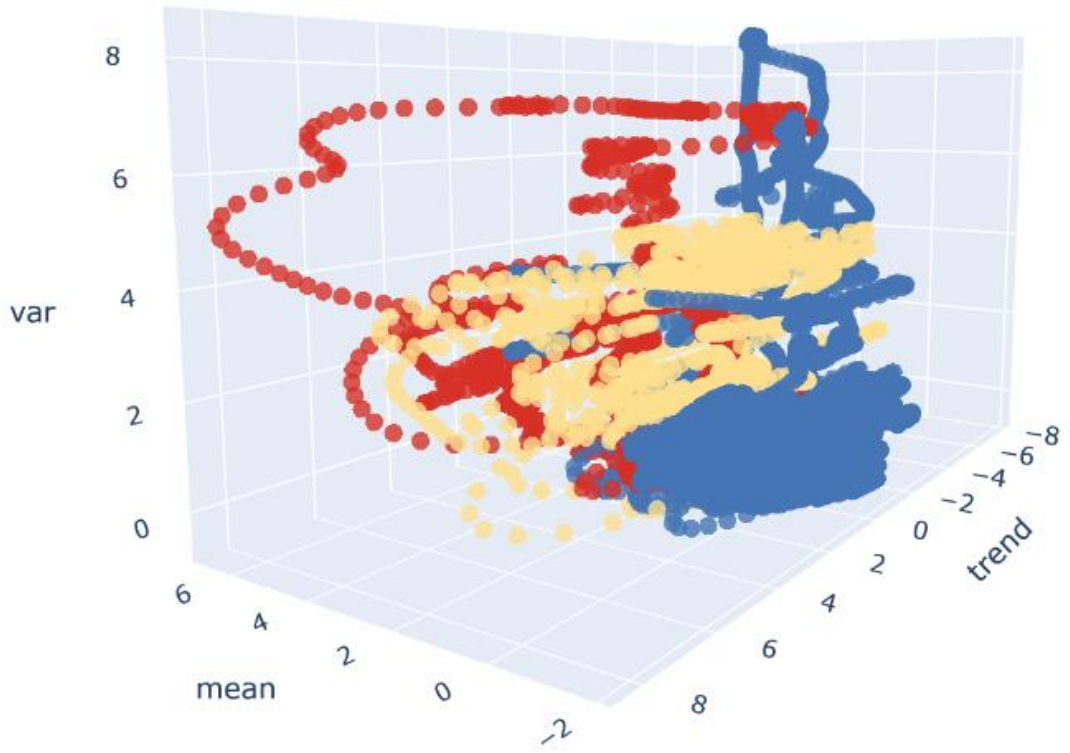

Figure 6-37. Previously detected anomalies, non-detected anomalies and normal behaviour – Generic anomaly detection

The objective in this iteration is to get a model able to detect all the anomalies shown and to detect future anomalies not studied until now.

The anomalous registers in the data are only a subset of the total possible population, meaning that we should build a model that considers any anomaly, even the classes that we do not know. To solve the problem, we need algorithms that learn from normal behaviour and are capable of classifying if the incoming registers differ and therefore are anomalous. *One-class classification* (*OCC*) techniques predict if an instance pertains to a specific class or not, by primarily learning from a training set containing only registers of that class, although some variants use counterexamples to better define the classification boundary.

In this iteration, the data model used is the same as the past iteration, since the model also includes the variance, which is a crucial point to detect noise problems. Nevertheless, a change of paradigm is applied, from supervised to unsupervised learning, which is going to affect to the predictions obtaining different results.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

6.3.2.2. MODELLING & EVALUATION

In this section, *One-class Support Vector Machines* (*O-SVM*), *Local Outlier Factor* (*LOF*) and *Isolation Forest* (*IF*) will be prepared and compared. These algorithms are unsupervised, they learn from normal data containing some anomalous observations called "contamination" and will predict if a register is normal or not.

The dataset used contains 50% of normal data, 14% of anomalous data and 36% of normal data with a little bit of noise. As Local Outlier Factor only needs normal behaviour to learn, the data has been split in 60% of normal data to train and 40% of the remaining to evaluate. *One-class SVM* and *Isolation Forest* need contamination in the dataset to learn and establish a better margin, so they have been trained using 70% of the total data and evaluated using the remaining 30%.

As explained in the previous section, the *Numenta Score* and the *Precision Score* are adopted to evaluate the predictions of the different models. Below, Table 44 is introduced to show the results of the evaluation of models with different parameters, using the *Numenta Score* with a 0.5 false positives rate parameter and the *Precision Score*.

Table 44. Results of the O-SVM, IF and LOF evaluation – Generic anomaly detection

Algorithm	Contamination	neighbours	Numenta Score	Precision Score
O-SVM	0.05	n/a	0.79	0.61
O-SVM	0.04	n/a	0.80	0.63
IF	0.03	n/a	0.79	0.62
IF	0.05	n/a	0.78	0.60
IF	0.04	n/a	0.79	0.62
IF	0.03	n/a	0.76	0.61
LOF	0	150	0.78	0.57
LOF	0	80	0.77	0.56

Different contamination values have been assessed for each algorithm (for *Local Outlier Factor* the contamination needed to be 0, but a different number of neighbours have been tried). The best combination has been the use of *O-SVM* with 0.04 contamination, obtaining a *Numenta Score* of 0.8 and a *Precision Score* of 0.63. *IF* shows good results with a 0.04 contamination, and finally, *LOF* shows the worst values of all.

Figure 9-14 shows the predictions done by the model based on *O-SVM*. As it can be seen, all the anomalies previously labelled are correctly predicted, the zones with more noise are also detected, but there are some points predicted as anomalous when their behaviour is good. This last sort of predictions are the false positives and the reason of having a low precision score in each algorithm.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

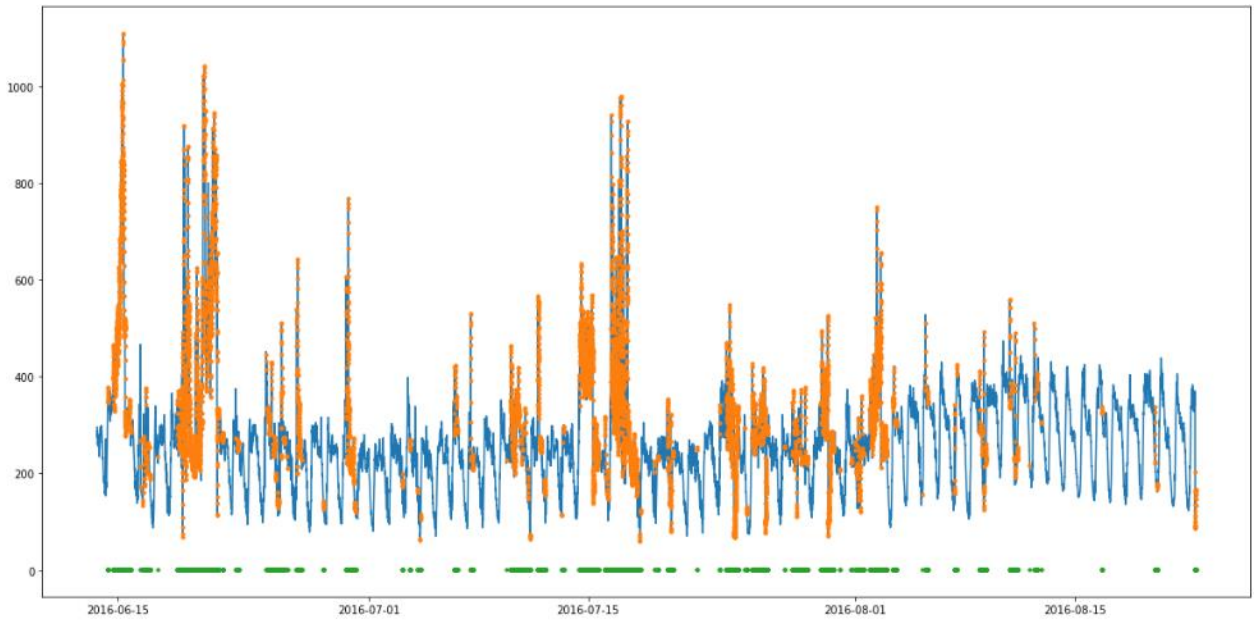

Figure 6-14. Points detected by the algorithms – Generic anomaly detection

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

7. CONCLUSIONS AND FUTURE WORK

The work done during this first year include all sorts of Machine Learning techniques, such as sensor simulation, outlier detection, spatial predictions, data quality evaluation, drift detection and anomaly detection.

The use case for the city of Amersfoort introduces the idea of an early warning system to detect flash flood using a short-time prediction on a 2-hour horizon value and a heat risk prediction. The studied case for flash flood makes an introduction to future implementation of the Machine Learning model using historical prevision data, and the heat risk prediction introduces the first iteration on the data model used and the training/evaluation of the model created.

In this use case, the temperature time series are validated automatically by a novel Machine Learning algorithm to work with the data model containing some null values. The algorithm Histogram-based Gradient Boosting Classifier shows some initial results on the temperature validation, but it will be improved in the future. Also, it is planned to improve the data model, adding spatial features, and improving the quality of the registers.

The Barcelona use case introduces the idea of predicting sediment in the sewer grid using spatial prediction, considering not only physical properties of the section but also properties of the nearby sewer sections and sediments to predict the objective sediment. The number of resources that could be saved with a good model would improve economically and socially the city council and the citizens, but the problem is difficult to solve because the number of registers is low, and the models are not able to learn the relations between the features. More data is expected in the future to obtain more registers and improve the models. Additionally, new strategies will be faced, like the prediction of future sediment level in a section using the trend of the past values, with the aim of improving current results.

A solution to predict anomalies in the water from construction sites is introduced for the Gothenburg use case, training unsupervised learning algorithms with water quality data. The algorithm Isolation Forest presented good results, but only one month of normality was used and in the final data model only two anomalies were available to evaluate the model. Despite this data problem, the model results are good enough to plan future iterations were the quality of the registers will be improved and more data will be added to improve the model.

Additionally, two generic models are introduced which can be used for solving urban water problems, the univariate detection of anomalous data on quality sensors and incremental drift detection using univariate data.

The unsupervised anomaly detection solution for water quality sensors can detect different type of abnormalities in the data with a good score, with a counterpart of detecting some false positives. One-class Support Vector Machine is the most reliable algorithm between some state-of-the-art anomaly detection algorithms tested, being an opening step to continue studying novelty detection algorithms in the future. In this study, there is a lack of deep learning algorithms, which have not been tested because the volume of data gathered is not significant. In the future, more data will be added, and the team will experiment with deep learning algorithms such as Deep Belief Networks or hybrid solutions between auto-encoders and O-SVM. Another improvement will be the addition and testing of more sensors and different types of anomalies to improve the validity of the models.

The drift detection model can predict dangerous drift in real-time on ammonium and turbidity sensors. The study compares a batch of classification algorithms, from linear predictions to ensembles and neural networks, using the Numenta Anomaly Benchmark, which is a novel benchmark to evaluate real-time anomaly detection models. The empirical results highlight the feedforward neural network as the best model, obtaining high NAB and precision. Knowing the good results provided by a simple architecture, one of the future actions is to study different neural networks. As explained, these models were tested using data from two different sensors, so it's a must to add new sensors in the future and secure the generalization of the models.

Finally, it is important to note that a new version of this document will be presented on M36 (D2.5), enhancing current accuracy of data-driven models, and adding new ones.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

8. REFERENCES

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat., 46*(3), 175-185.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Chapman and Hall/CRC.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). LOF: identifying density-based local outliers. *ACM sigmod record, 29*(2), 93-104.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Das, S., Fern, A., Dietterich, T., Emmott, A., & Wong, W. (2016). Anomaly Detection Meta-Analysis Benchmarks.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit., 74*, 406-421.

Gray, J. (19 de 02 de 2020). *Sustainable Build*. Obtenido de http://www.sustainablebuild.co.uk/

Ho, T. (1995). Random Decission Forests. *Proceedings of 3rd international conference on document analysis and recognition, 1*, 278-282.

Kriegel, H., & Zimek, A. (2008). Angle-based outlier detection in high-dimensional data. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 442-452.

Lavin, A., & Ahmad, S. (2015). Evaluating Real-Time Anomaly Detection Algorithms -- The Numenta Anomaly Benchmark. *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 38-44.

Liu, F., Ting, K., & Zhou, Z. (2008). Isolation forest. *Eighth IEEE International Conference on Data Mining*, 413-422.

Rosenblatt, F. (1957). The perceptron, a perceiving and recognizing automation. *Cornell Aeronautical Laboratory*.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., & Platt, J. (2000). Support vector method for novelty detection. *Advances in neural information processing systems*, 582-588.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29-39.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ANNEX 1 – DATASCIENCE AND MACHINE LEARNING CONCEPTS

Autocorrelation

An autocorrelation plot has in the y-axis the correlation value (usually the Pearson correlation) and the lag in the x-axis. A lag means a past point, so if a point is gathered each minute, a lag of 1 means a point one minute ago, and a lag of 10 means the points 10 minutes ago.

Cross Correlation

Cross-correlation measures the similarity between a vector x and shifted (lagged) copies of a vector y as a function of the lag.

Supervised Algorithms

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output, that is, Y = f(X). The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Below, more common supervised algorithms are introduced:

The **Support Vector Machine (SVM)** (Cortes & Vapnik, 1995) algorithm focuses on finding a hyperplane that divides the n-dimensional space defined by input data into two regions, maximizing its distance or margin (see Figure 8-1). Depending on the chosen kernel, the method allows both linear and non-linear classification, thanks to the mapping of entries to a larger space, where the separation hyperplane can be found in a simpler way.

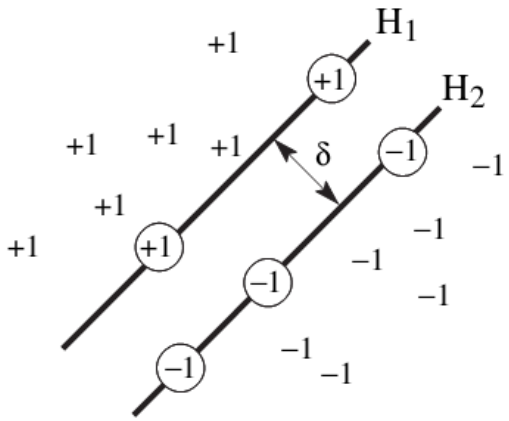

Figure 8-1 Rationale behind the Support Vector Machine

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The **K-Nearest Neighbours (KNN)** (Altman, 1992) is based on the premise that the prediction or classification of an unknown instance can be accomplished through the relationship with known instances, weighted by a metric or distance. Typically, the Euclidean distance is used as a measure of similarity, but other distances can be implemented and used to better adjust the operation of the algorithm and the type of data. Figure 8-2 illustrates the rationale of the KNN algorithm: it finds the k neighbours closer to the new sample to determine its class.

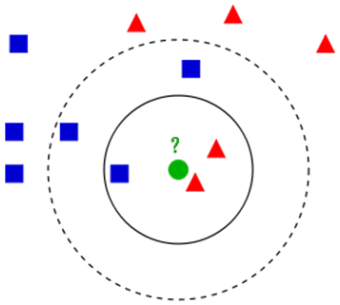

Figure 8-2. Rationale behind the K-Nearest Neighbours

The **Decision Tree (DT)** (Breiman, Friedman, Olshen, & Stone, 1984) is a popular tool in machine learning that that makes divisions in the data set ensuring the maximum number of data in the same category or tag within each division. In the example of Figure 8-3, during the training of the decision tree 4 divisions, also called "leaves", have been created. Against new data from x1 and x2, the DT would be able to determine the class following the reasoning shown in Figure 8-4.

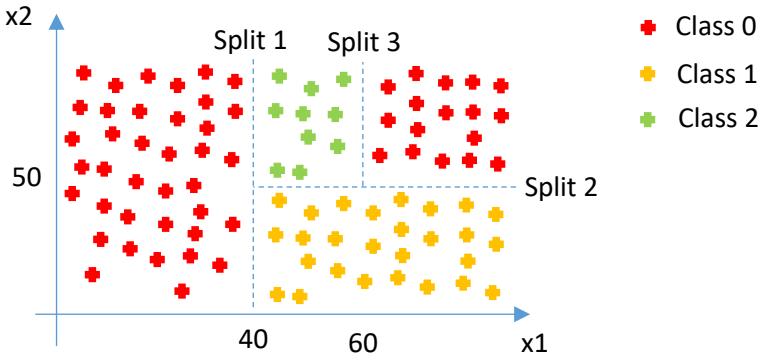

Figure 8-3. Divisions (leaves) created by the Decision Tree

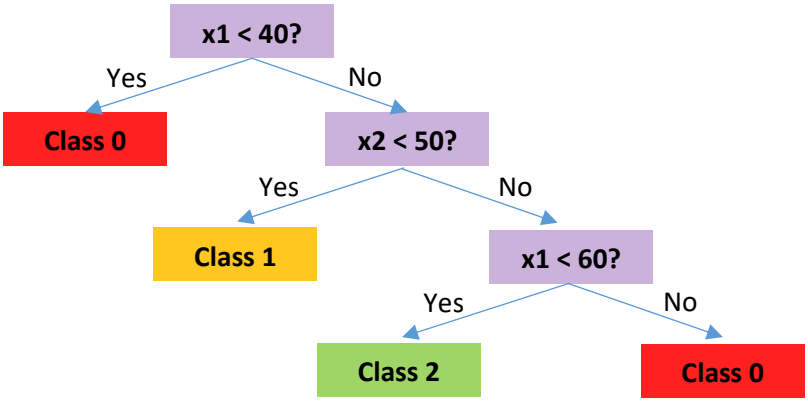

Figure 8-4. Decision Tree created for the example shown in Figure 8-3

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

The **Random Forest Tree (RFT)** (Ho, 1995) is an ensemble algorithm which is built on a multitude of decision trees during the training of the model (see Figure 8-5). The principle in which the "ensembles" are based is the following: a set of "bad" predictors can together become a good predictor. In the case of the Random Forest Tree, "bad predictors" are in fact decision trees, while the good predictor is the set of random trees. Each tree of the set makes a prediction and the most voted by the set of trees is the winning prediction. The RFT can also provide additional information, such as the identification of the most relevant variables in the process.

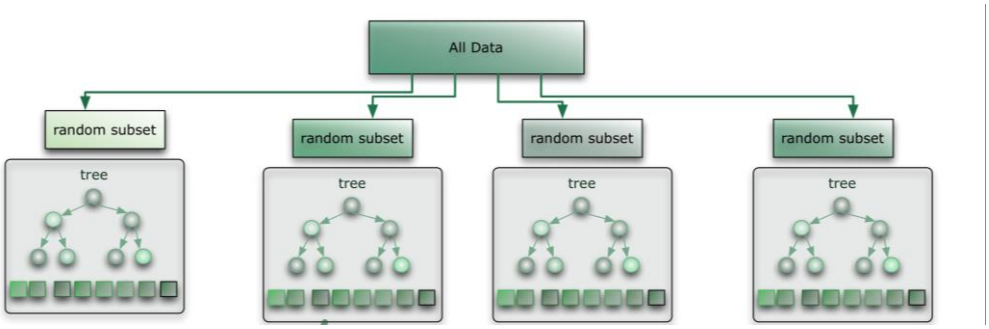

Figure 8-5. Rationale behind the Random Forest Tree

The **Artificial Neural Network (ANN)** (Rosenblatt, 1957) is one of the main tools used in machine learning, which intends to replicate the human brain learning process. Neural networks consist of input and output layers, as well as hidden layers that transform the inputs into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize. During the training of the ANN, the backpropagation technique allows the ANN to adjust its parameters in order to improve the predictive performance of the model.

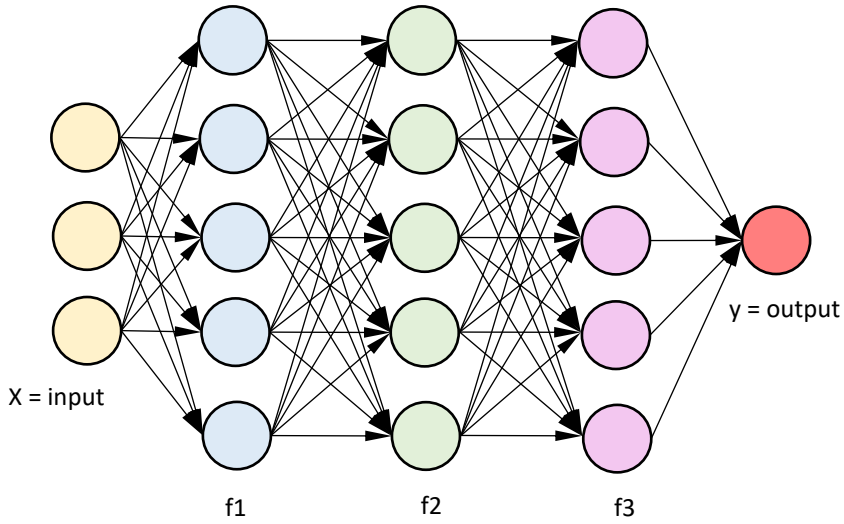

Figure 8-6. Artificial Neural Network example

Unsupervised Algorithms

Unsupervised learning is where you only have input data (X) and no corresponding output variables. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular algorithms of unsupervised learning focused on Novelty Detection, where training data is not polluted by outliers and we are interested in detecting whether a new observation is an outlier, are:

One-class SVM (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 2000)is a domain-based method which relies on the construction of a boundary separating the nominal data from the rest of the input space by applying the support vector machine algorithm to one-class problems. The method computes a separating hyperplane by maximizing the margin between the input data and the origin in the high-dimensional space. The algorithm allows a percentage of data points to fall outside the boundary in order to prevent over-fitting from happening. This percentage acts as a regularization parameter.

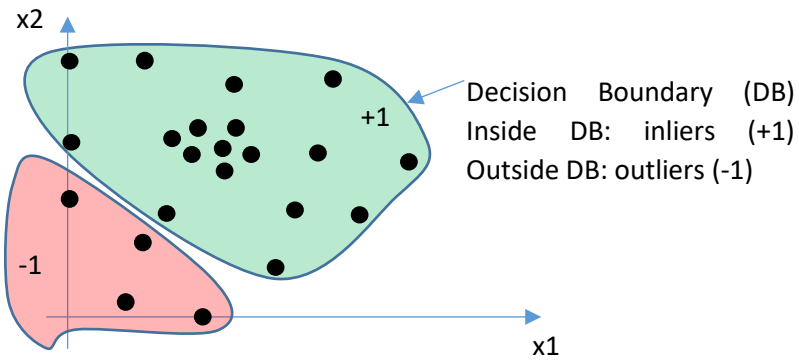

Figure 8-7 One-Class SVM Classifier

Local outlier factor (LOF) (Breunig, Kriegel, Ng, & Sander, 2000) is a well-known distance based approach that studies the neighbourhood of each data point to identify outliers. For a given data point, this algorithm computes its degree of being an outlier based on the Euclidean distance between the data point and its closest neighbour. A recent study (Das, Fern, Dietterich, Emmott, & Wong, 2016) shows that LOF outperforms Angle-Based Outlier Detection (Kriegel & Zimek, 2008) and One-class SVM when applied on real-world datasets for outlier detection.

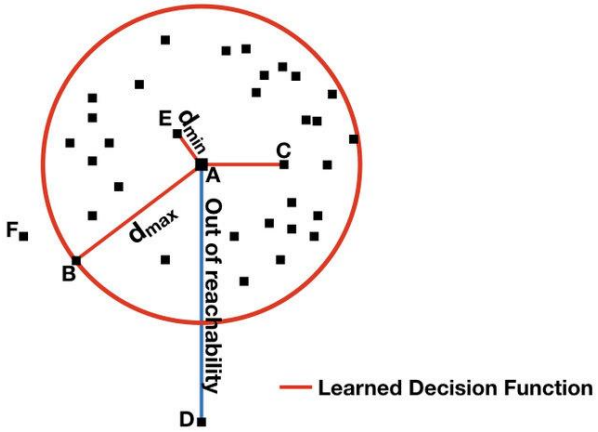

Figure 8-8. Local Outlier Factor: each point is compared with its local neighbours instead of the global

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Isolation forest (Liu, Ting, & Zhou, 2008) is a method that focuses on isolating anomalies instead of profiling normal points (see Figure 8-9). It uses random forests to compute an isolation score for each data point. Recursive random splits are performed on attribute values, hence generating trees able to isolate any data point from the rest of the data. (Domingues, Filippone, Michiardi, & Zouaoui, 2018) showed that this was the most performing outlier detection method for the real-world datasets of their study.

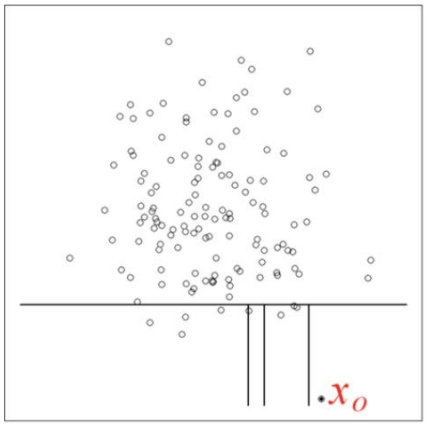

Figure 8-9. Identifying outliers with Isolation Forest

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ANNEX 2 – SCORING METRICS

Regression Scoring Metrics

MAE

Mean absolute error: The arithmetic average of the absolute difference between the predictions and the real values.

MSE

Mean squared error: The arithmetic average of the squared difference between the predictions and the real values.

R^2

Coefficient of determination (R2): The proportion of the variance in the objective variable that is predictable from the predictive variables (input vector).

Classification Scoring Metrics

Precision

Precision, also named Positive Prediction Value (PPV), measures how well our algorithm identified only instances of one class, for example anomalies.

Precision = TP / (TP + FP), where TP = True Positive and FP = False Positive.

Recall

Recall, also named True Positive Rate (TPR), measures how well our algorithm identified all instances of one class, for example anomalies.

Recall = TP / (TP + FN), where TP = True Positive and FN = False Negative

Accuracy

Accuracy measures the ratio of number of correct predictions to the total number of input samples.

Accuracy = (TP + TN) / (TP + TN + FP + FN), where *TP = True Positive*

TN = True Negative

FP = False Positive

FN = False Negative

Numenta Anomaly Benchmark

The Numenta Anomaly Benchmark (Lavin & Ahmad, 2015) uses a set of weights, a scaled sigmoidal scoring function, and an evaluation of false positives and false negatives to score a set of anomaly prediction having into account a set of requirements:

i. Detects all anomalies present in the streaming data.
ii. Detects anomalies as soon as possible, ideally before the anomaly becomes visible to a human.
iii. Trigger no false alarms (no false positives).
iv. The algorithm can be used in real-time environments.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Code 1 presents the body structure of the scoring function. The algorithm needs three inputs: a weight to punish the prediction of false positives, an ordered list of the true values and another ordered list of the predicted values. Step 1 evaluates each point predicted by the algorithm using code 2 and stores the score in the variable SA. Step 2 calculates a perfect score, using the true values as predicted values, and Step 3 calculates a null score using the true values and a list without anomalies predicted. After obtaining the three scores, the precision and SNAB scores are computed in steps 4 and 5. Finally, the SNAB and precision scores get multiplied by the punishing weight and added up.

Code 1 Scorer structure

```
Input: Two sets of real points and predicted points of the same length, true and pred;
a balance of true and false positives X between [0, 1]
Output: Score between [0, 1]
1: SA <- Evaluate each pred point and sum the scores
2: SAperfect <- Calculate a perfect score
3: SAnull <- Calculate a score without anomalies predicted
4: SNAB <- (SA - SAnull) / (SAperfect - SAnull)
5: precision <- TP/(TP+FP)
6: return X * SNAB + X * precision
```

Code 2 shows the procedures to calculate a score given two lists. The first one is used as the correct data and the second as the data to be evaluated. The procedure is divided into two big steps; the identification of anomaly ranges (steps 1 to 10) and the point evaluation of the test set (steps 11 to 21).

Step 1 defines a set of anomaly ranges that will be used to assert each point of the test set. Steps 3 to 4 compute and append the start position of the ranges while steps 6 to 8 append the end position. Steps 11 and 12 define the set of computed scores for each anomalous point and punishing values for undetected ranges. Step 15 appends anomalous point scores computed by Code 3 to the set of tanh scores, while step 18 computes the punishing values. Finally, the sum of the positive scores is done and the punishing score is applied.

Code 2 Score calculator

```
Input: Two sets of points of the same length, one as a correct set and the other as a
set to evaluate
Output: Score value
 1: Default initialization of anomaly_positions
 2: for i = 0 to n_correct_values do
 3:   if i is the start of an anomaly range do
 4:     start_position <- i
 5:   end if
 6:   if i is the end of an anomaly range do
 7:     end_position <- i
 8:     anomaly_positions <- append [start_position, end_position]
 9:   end if
10: end for
11: Initialize tanh_scores
12: Initialize fd
13: for i = 0 to n_evaluate_values do
14:   if i is in anomaly_positions do
15:     tanh_scores <- append point score of i
16:   end if
17:   if anomaly_position has not been predicted do
18:     fd <- append false negative punishment
19:   end if
20: end for
21: return sum(tanh_scores) - sum(fd)
```

To evaluate each anomalous point, the Numenta Anomaly Benchmark uses a scaled sigmoidal scoring function, where the predictions made sooner are rewarded positively. In code 3, an alteration of the function is shown, where the weight of the false positives has no impact on the evaluation, using a tanh function to estimate the point score. The team decided to evaluate the false positives using the precision metric to punish the false alarms a bit more.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Code 3 Tanh

```
Input: Relative position in the anomaly range y, between 0 and 1
Output: Score between (-1, 1)
1: return 2 * (1 / (1 + e^(5*y))) - 1
```


This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ANNEX 3 – CROSS-VALIDATION BASED ON TIME SERIES SPLIT

Time Series Split is a variation of k-fold, returning first *k* folds as train set and *(k+1)* th fold as test set. It is important to remark that unlike standard cross-validation methods, successive training sets are supersets of those that come before them.

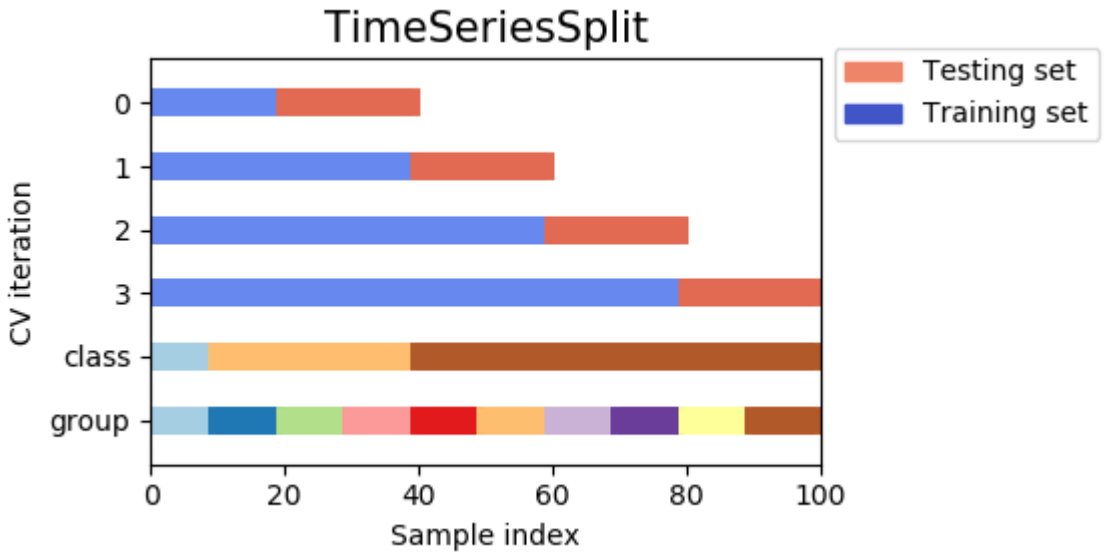

Figure 8-10. Example of Time-Series Split technique (source: scikit-learn.org)

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

ANNEX 4 – STOCKTAKING

A final Annex of stocktaking was included in all Deliverables of SCOREwater produced after the first half-year of the project. It provides an easy follow-up of how the work leading up to the Deliverable has addressed and contributed to four important project aspects:

1. Strategic Objectives
2. Project KPI
3. Ethical aspects
4. Risk management

STRATEGIC OBJECTIVES

Table 45 lists those strategic objectives of SCOREwater that are relevant for this Deliverable and gives a brief explanation on the specific contribution of this Deliverable.

Table 45. Stocktaking on Deliverable's contribution to reaching the SCOREwater strategic objectives.

Project goal	Contribution by this Deliverable
SO3 Enable the monetization of water cycle data and create new markets and business opportunities for development and innovation of new products and services.	By providing a set of data-driven models which can be considered new products or services
SO 4 Demonstrate benefits of smart water management for increased water-system resilience against climate change and urbanisation	By demonstrating relevant outcomes related to the data-driven models' performance

PROJECT KPI

Table 46 lists the project KPI that are relevant for this Deliverable and gives a brief explanation on the specific contribution of this Deliverable.

Table 46. Stocktaking on Deliverable's contribution to SCOREwater project KPI's.

Project KPI	Contribution by this deliverable
KPI 3: Number of innovation tools that illustrate the use and potential of the SCOREwater	Multiples smart services were designed throughout the D2.4
KPI 4: Reduce the pollutant load from construction work in Gothenburg	The data-driven model improves the management of pollution events though a preventive notification.

ETHICAL ASPECTS

Table 47 lists the project's Ethical aspects and gives a brief explanation on the specific treatment in the work leading up to this Deliverable. Ethical aspects are not relevant for all Deliverables. Table 47 indicates "N/A" for aspects that are irrelevant for this Deliverable.

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

Table 47. Stocktaking on Deliverable's treatment of Ethical aspects.

Ethical aspect	Treatment in the work on this Deliverable
Justification of ethics data used in project	N/A
Procedures and criteria for identifying research participants	N/A
Informed consent procedures	N/A
Informed consent procedure in case of legal guardians	N/A
Filing of ethics committee's opinions/approval	N/A
Technical and organizational measures taken to safeguard data subjects' rights and freedoms	N/A
Implemented security measures to prevent unauthorized access to ethics data	N/A
Describe anonymization techniques	N/A
Interaction with the SCOREwater Ethics Advisor	N/A

RISK MANAGEMENT

Table 48 lists the risks, from the project's risk log, that have been identified as relevant for the work on this Deliverable and gives a brief explanation on the specific treatment in the work leading up to this Deliverable.

Table 48. Stocktaking on Deliverable's treatment of Risks.

Associated risk	Treatment in the work on this Deliverable
Low commitment of the partners to the project plan and deadlines	All partners have been active and given input at time
Lack of consensus on scientific or technological approach	Consensus have been assured by Skype discussions
Data from Cases are sparse and are not enough to apply all methods and tools	Plan Skype discussion to improve of gathering data
Outputs generated by the smart algorithms not as useful as expected	Fine-tune of algorithms hyperparameters Creation of new features Integration of new data-sources

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751

scoreWATER

This project hase recieved funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no 820751