ANOMALY DETECTION ON LOW-COST AIR QUALITY SENSORS USING CNN PREDICTION

CHRISTINE WILSON

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

732345

COMMITTEE

dr. Samaneh Khoshrou
dr. Nevena Rankovic

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2023

Word count: 7205

ACKNOWLEDGMENTS

I would like to thank dr. Samaneh Khoshrou for her supervision and advice during the process of this thesis.

ANOMALY DETECTION ON LOW-COST AIR QUALITY SENSORS USING CNN PREDICTION

CHRISTINE WILSON

Abstract

The monitoring of air quality is of vital importance. With new local emissions emerging, from wood burning for heating, local measurements are needed. Due to the cost of high-quality sensors, ones of lesser quality need to be used to create a high-density measurement network. The problem is to ensure the data quality of these sensors. Calibration works but has drawbacks in the form of money or impossibilities. A proposed other technique to ensure the proper use of these low-cost sensors is proposed in this thesis. The research question: "How well can CNN prediction anomaly detection on pm2.5 low-cost sensor time series data be used instead of calibration that takes environmental factors into account to find true high emissions?" is researched. This technique has already been used in other time-series anomaly detection, but not yet in air quality measurements. Other environmental features have been successfully modeled with deep learning techniques. From the literature it has become clear that a CNN model works well with time-series data. The dataset used to look at this technique will be the Samen Meten pm2.5 dataset, consisting of measurements done in different citizen science projects. Calibrated and uncalibrated data is compared. With the calibrated data an autoregressive model is used to detect anomalous points. The uncalibrated data uses a CNN model and anomaly detection. The results is a F1-score of 0.16 which indicates that the model as proposed in this paper cannot be used as an alternative way to factor out environmental influences form the time series.

1 DATA SOURCE/CODE/ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis. The

author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is publicly available. [1] [2]

2 INTRODUCTION

The importance of air quality monitoring is of vital importance in a densely populated country like the Netherlands. The RIVM monitors the air quality levels through high quality, accurate sensors around the country. The air quality is evaluated based on yearly averages from these sensors. Luchtmeetnet, which reports data collected by the RIVM, reports that the yearly average pm2.5 level between 2021 and 2023 was 9.78 ug/m3 per hour. The European council has declared that the average yearly level of 25 ug/m3 per hour is the level under which the air quality is safe. Seeing this average over the last few years it looks like there are no problems with the the air quality.

However, in recent years, the Netherlands has seen a rise in wood burning as residential heating. This results in very localized, short spikes of emissions, with the small particles that are released being the main health concern. Research done by (Favez, Cachier, Sciare, Sarda-Estève, & Martinon, 2009) found that around 20% of aerosols in Paris in the winter came from residential wood burning. The level of pm2.5 can be as high as 1000 ug/m3 as found by (Hellén, Hakola, Haaparanta, Pietarila, & Kauhaniemi, 2008). They also found that these very high and very local emissions were due to residential wood burning. These high spikes in pm2.5 concentrations are also concerning regarding health. Therefore, more insight is needed into these spikes, as the government needs to take into account these increasing unhealthy sources of emission. The few sensors that the RIVM uses cannot give information about emissions that are this local and short term. These high quality sensors would also be too expensive to create the high-density network needed. Therefore, a different system of measurement is needed to be able to measure the scope of these new emissions.

A possible solution would be to use low-cost pm2.5 sensors which are cheap and easy to install. The RIVM has started the Samen Meten program to combine these low-cost sensors from different projects into a national network. As of December 2022 6404 pm2.5 low-cost sensors were registered, with more being added as they are installed. The main problem is that these sensors are very sensitive to environmental factors like humidity, and therefore do not comply with the European standards

[1] Code used in thesis: https://github.com/ChristineWilson99/master-thesis-DS-S

[2] Datasets: https://drive.google.com/file/d/136LjoKUqWGuvpenA1hokKNXJb6UJbqFY/view?usp=sharing https://drive.google.com/file/d/1Gz9EvCoomLKk3HeotSFlzwk2liyDdAla/view?usp=sharing

of accuracy (Bartonova et al., 2019). Therefore, the RIVM is doing research on how to use these sensors to identify spikes in local air quality. The RIVM has been trying out calibration methods to take out the variation in measurements due to environmental factors. The main assumption used with these calibrations is that the measurements locally do not differ greatly. Groups of sensors are compared against high quality sensors, also taking into account the distance to a high-quality sensor. With this method of calibration, the environmental factors are taken into account and sudden high data points can be identified as anomalies in the time series measurements, as done in research by (Chen et al., 2017).

However, this calibration method has multiple flaws. Firstly, to properly calibrate the sensors, high-quality sensors are needed nearby. This is not always the case. A solution to this would be to calibrate the sensors close to a high-quality sensor for a few weeks and deploy them after calibration. (Bartonova et al., 2019) makes the point that due to these environmental factors the calibration does not stay accurate and should be redone periodically. This would be too time consuming and expensive to implement on a large scale. Secondly, the assumption that measurements are locally similar holds less now that citizens are burning more in their homes and creating very local and short high emissions. Taking a reference station a few kilometers away would be a completely different measurement and would not represent the local concentration of pm2.5. Therefore, a different method is needed to take into account the environmental factors that influence these cheaper sensors, but that does not rely on secondary sensors.

A different method to incorporate the environmental factors that influence these sensors should be researched. Research has been done into different methods to incorporate these environmental factors. Different papers have found that deep learning model show promising result, with CNN being the best. (Ali, Glass, Parr, Potgieter, & Alam, 2020; Okafor, Alghorani, & Delaney, 2020; Veiga, Ljunggren, Bach, & Akselsen, 2021) These techniques have been used on data sets that are neater than the Samen Meten data set. Using CNN on this data set will explore the uses of this technique on a more realistic data set. This will show the broader application of the technique for new air quality projects.

To explore the possibilities of using low-cost pm2.5 sensors to measure local air quality and emissions the following research question will be explored in this paper:

> *How well can CNN prediction anomaly detection on pm2.5 low-cost sensor time series data be used instead of calibration that takes environmental factors into account to find true high emissions?*

To test this research question, anomaly detection will be done on calibrated data and uncalibrated data that has had CNN prediction used to possibly incorporate the environmental factors. The anomaly detection on the calibrated data will be used a s the baseline. In Section 3 the theoretical background to this research question will be further elaborated.

3 RELATED WORK

In this section the related works to the topic will be explored. The section will look at the uses of anomaly detection in relation to time series, how high emissions and anomalies relate, the environmental factors that influence the low-cost time series data, data quality improvements done on low-cost pm2.5 sensors. Lastly, this section will look at the best model to use for supervised anomaly detection on air quality time series data.

Finding high levels of measurements from sensor measurements is more commonly known in the literature as anomaly detection in time series data. (Aggarwal, 2017) describes that anomaly or outlier detection in time-series looks at the break in temporal continuity of a certain data point. This means that the data point breaks away from the trend that the time series follows. Time series have a strong correlation across time and a sudden deviation from this trend is considered an anomaly. In the case of pm2.5 data we are looking for abrupt changes is the time series, an anomaly, those are the sudden high emissions from wood burning. The temporal continuity of the air quality time series follows environmental influences. The wood burning is an human action that is not in line with the gradual fluctuations of these time series. We can find these anomalies by looking at the context, the measurements preceding the higher data point and the normal trend of the time-series. A good method for this is to use autoregressive models (AR), which takes the preceding values and an error to define a certain data point. (Aggarwal, 2017) The classification of the data point as an anomaly or not is based on the difference from the previous data points looked at with a certain threshold.

With the detected anomalies we can gain insight into high emission point by looking for anomalies. (Hellén et al., 2008) found high concentration peaks in the average hourly measurements they used and was able to link these to residential wood burning, that was their origin. They found in their research that peaks could go up to 1000 ug/m3 and that the average was only 8 ug/m3. This large difference in levels of pm2.5 make it so that anomaly detection can be used to find these high emissions. Using anomaly detection to find high emission points was done by (Chen et al., 2017) who used a statistical model on a low-cost pm2.5 sensor network data set to find anomalies, which were taken as the high emission points.

However, from the literature it also becomes clear that all time-series from low-cost pm2.5 sensors suffer from the same strong influences of environmental factors which make the time-series not reliable. (Okafor et al., 2020) High levels of pm2.5 could be due to humidity as much as actual emissions. Much research has been done to improve the calibration and integrate the environmental factors into the calibration of pm2.5 sensors with both machine learning and deep learning techniques. (Ali et al., 2020; Okafor et al., 2020; Veiga et al., 2021) In these papers, environmental factors were identified with different models (feature selection, multiple linear regression) and calibration, using various deep learning approaches, was used to remove the influence of the environmental factors from the time-series. However, these techniques all relied on environmental factors being measured. no research has been done with air quality data and incorporating environmental factors without external measurements.

The above-mentioned research into data quality improvement using calibration was not done with the specific goal of this project in mind. Because the goal for this project is to be able to use the low-cost pm2.5 sensors to detect local high emissions, environmental factors can also be taken into account when identifying anomalies. (Ali et al., 2020) identifies humidity as being the largest correction needed for pm2.5 sensors. With a high humidity there are more water droplets in the air, that the sensors identify as particles in the air, making the measurement higher than the true value of fine particles in the air. This results in the measured time series being made up of a measurement of air particles, an added value to the measurement for humidity, and a smaller extra variance for other environmental factors. These dependencies can be measured separately as with the calibration method (Watne et al., 2021), but can also be modeled.

In the field of air quality no research has been done yet on using deep learning techniques to take environmental factors into account without the use of additional measurements, as mentioned above. However in other fields it has. An example is (Jin et al., 2021) who looked at the possibilities of using deep learning (Long Short-term Memory) to predict environmental factors. They concluded that with deep learning techniques the fluctuation of environmental factors could be accurately modeled. This worked better than machine learning models as the features were identified and learned by the model itself. As environmental factors fluctuate in predictable patterns, deep learning can learn their fluctuations. A technique to take environmental (or other variance) into account when doing anomaly detection on time series is to use supervised prediction to learn the temporal continuity of the time-series and find the abnormal data points. (Aggarwal, 2017)

The latest research into this technique has therefore been with the use of deep learning to model the time series and predict the following data point of a sequence. The predicted value is then compared to the actual measurement and, depending on the distance between these, the measurement is classified as either an anomaly or not. (Munir, Siddiqui, Dengel, & Ahmed, 2018) and (Gao et al., 2020), both use a CNN as the deep learning algorithm. Their choice of a CNN model comes from the benefits that the CNN gives. CNN has the ability to train on smaller data sets and has great generalization capabilities. In the overview paper of (Jácome-Galarza, Realpe-Robalino, Paillacho-Corredores, & Benavides-Maldonado, 2022), who compares different models for time series prediction, CNN also comes out as a very good deep learning model for this application. The difference between the papers is that (Munir et al., 2018) uses prediction to do the anomaly detection while (Gao et al., 2020) uses decomposition. The DeepAnt of (Munir et al., 2018) is slightly better with F1-scores of 0.87, compared to the 0.693 that (Gao et al., 2020) manages. These papers do not try their models on air-quality data. Which leaves a gap to be explored in this paper.

To conclude this section, in the scientific work done so far anomaly detection has not yet been used to find anomalies in pm2.5 data. The combination of anomaly detection, incorporating environmental factors into the model and low-cost pm2.5 sensors has also not been explored. This method will be used in this thesis and will be explained in more details in the following section.

4 METHOD

This section will cover the methods used in the thesis. It will start with a general overview of the used methods and will continue with explaining the different used methods in detail. The methods covered will be the data selection and preprocessing done on the data set, as well as details about the data set itself. Further methods detailed will be the autoregressive model, the anomaly detection and the CNN model. This section will end with the evaluation metrics used.

To test the research question mentioned in Section 2 two data sets will be needed. A data set with uncalibrated data, the raw measurements, and a second data set where those measurements have been calibrated to take out the influence of the environmental factors altering the measurements. To take the environmental factors into account without using calibration, a CNN model will be used on the uncalibrated data. From this model the predicted values will be used to classify the actual measurements as either anomalous or expected. To compare the usefulness of this technique

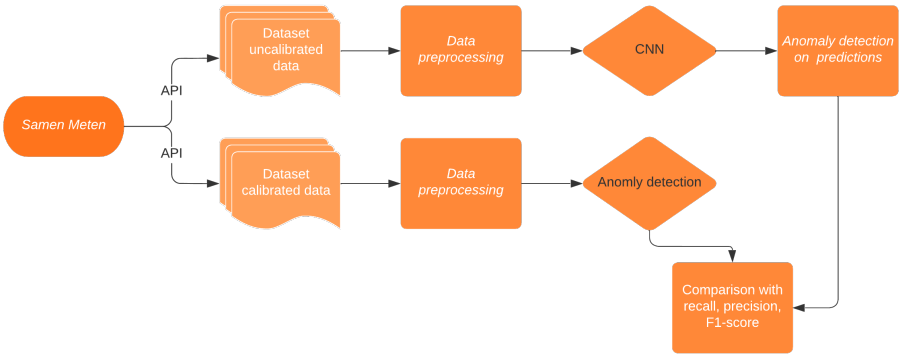

Figure 1: Flowchart data process.

compared to the calibration method, anomaly detection will also be done on the calibrated data set. This will create the baseline. On the calibrated data set a statistical prediction model, autoregression, will be used to create predicted values. By comparing these predicted values to the calibrated measurements the anomalous points will be classified. These sets of detected anomalies are compared to examine the capabilities of the CNN model as a substitute for calibration. Figure 1 shows the workflow of the thesis. In the Subsections below the mentioned methods are explained in further detail.

4.1 *Data set*

To explore this research question, the data set of pm2.5 sensors from the Samen Meten project has been used. The data set consists of 6404 low-cost sensors that measure fine particles in the air that are smaller than 2.5 micrometer. Measurements are recorded on an hourly basis for each sensor. For most sensors this raw data is then calibrated to create a second calibrated measurement. The calibration is done using reference measurements from high-quality air quality sensors that are relatively close by. Since not every low-quality sensor is near a high-quality sensor, calibration can not always be done. Sometimes the reference sensors also stop working or are decommissioned, which creates large gaps in the calibrated data available. Some inconsistencies in the calibration are also present, as not all calibrated measurements are different from the uncalibrated measurements. This would not be strange if not for some sensors, all the calibrated data is the same as the uncalibrated data. A second point to make is that the documentation on the data set does not divulge the exact calibration done on the measurements.

Metadata that is recorded with the measurements (in ug/m3) and calibrated measurement (ug/m3) includes the sensor ID, the date and time, the location, type of sensor and observed property. This data set was collected via API. When using the API, the filter used was the observed property, which was set to pm2.5. In the interest of time, as the API was quite slow, the data set is limited to the 1700 most recent sensors. For this thesis the only data that will be used and looked at in the models will be the measurements, the sensor and datetime features.

4.2 *Data Selection*

After collecting the data sets via API a few preprocessing steps were used to get usable data sets for the models. The first steps were to select all the sensors in the data sets that had enough measurements to be used in the models and to filter for the sensors that had both uncalibrated and calibrated data. The data sets had to be collected separately via API and since not all sensors have calibrated data this reduced the number of usable sensors. Since the API collected the most recent 1700 sensors, not all sensors had long time series. Another factor was that not all sensors recorded measurements from the first instance in the database completely until the use of the API. Some sensors have been disconnected before the use of the API or for other reasons were not recording measurements anymore. This has made the lengths of the time series differ greatly. Only sensor that had more than 672 measurements (a month of measurements) have been included in the final data set. With a month of measurements the sequence is large enough to be used in both models. These filters resulted in two data sets with 860 sensors that have both uncalibrated and calibrated measurements and have enough measurements to include in the models.

4.3 *Exploratory Data Analysis*

Table 1 shows the statistics of the resulting datasets. The discrepancy in the number of measurements, with the calibrated dataset having more, is due to the timing of the use of the API. The API was first used to download the uncalibrated measurements, three weeks later the API was used to collect the calibrated data. This resulted in an extra three weeks of calibrated measurements. When comparing the anomalies detected between the two models, the extra three weeks of measurements will fall away as only measurements with anomalies detection in both uncalibrated and calibrated data will be evaluated. The next statistic shows the amount of missing values in the two data sets. For the uncalibrated measurements

Table 1: Statistics data sets calibrated and uncalibrated measurements

	Calibrated	**Uncalibrated**
Sensors	860	860
Count total	4559054	4342708
Count NaN	1664644	79
Mean	33.51	32.27
Median	5.81	4.54
Minimum	-0.91	-0.91
Maximum	65535.0	65535.0

these are few, but the calibrated data set has many. In the preprocessing these missing values will be dealt with.

The last four statistics describe the distribution of the measurements. The minimum and maximum values describe a large range, with both the minimum and maximum being measuring errors, as both are not realistic values. A negative amount of particlet possible. 60000 ug/m3 is not a true value. (Hellén et al., 2008) recorded a highest value of 1000 ug/m3. When looking at the mean and median we can see that the mean is a lot higher than the median. For both data sets the difference is about 28 ug/m3. This difference is higher than the what the yearly average maximum should be. This tells us that the distribution is very skewed. Figure 2 shows the distribution of measurements for both data sets. The top two graphs show the cumulative distribution over the whole range. They show clearly that almost all measurements are far below 1000 ug/m3. To show the distribution better the bottom two graphs show the distribution of the measurements under 200 ug/m3. The distributions between the two data sets look similar, apart from the smaller amount of calibrated measurements. This is due to the high number of missing values in this data set. The distributions otherwise seem very similar. The statistics also lie closely together, with the calibrated measurements being slightly higher.

The amount of measurements also differed greatly between sensors. The minimum number of measurements was 1 and the largest amount for the uncalibrated data set was 11861, 17 months worth of measurements. For the calibrated data there were a few even larger time series due to the extra weeks of measurements. In Figure 3 the distribution of the amount of measurements can be seen.

The large disparity in amount of measurements is due to a few factors. The first is that the sensors selected for the data set were the the 1700 sensors that have been added the most recently. Therefore, not all sensors have been connected as long and therefore would have less measurements. Another factor is that some sensors don not stay connected. Some do

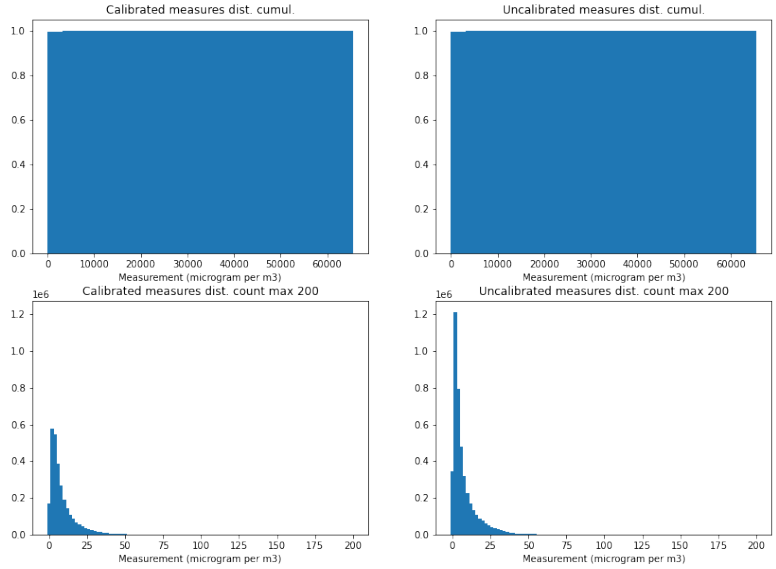

Figure 2: Measurements

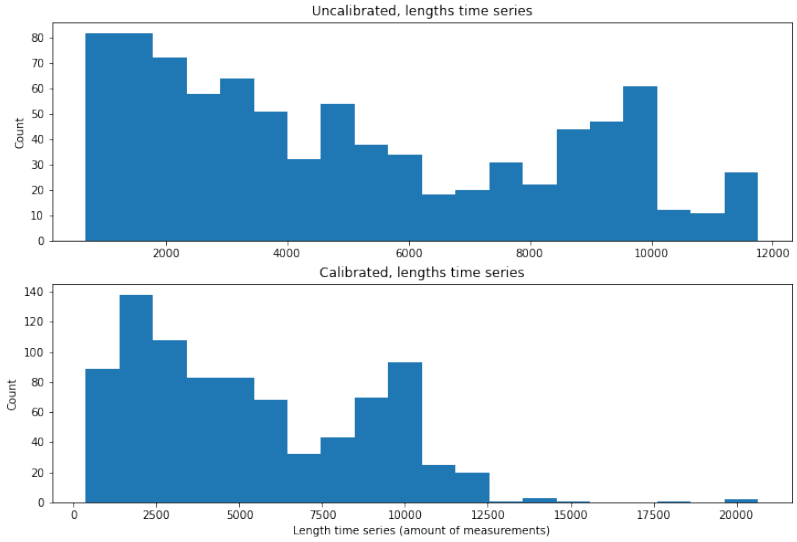

Figure 3: Distribution counts of length time series.

not record measurements past a certain date. Some had been active and measuring for over a year, with the first measurements starting in 2021. Other sensors only had measurements of a few days of weeks.

4.4 *Preprocessing*

With the data sets explored, the data can be processed to be used in the models. The first issue is the missing values. For the uncalibrated data there is not much missing data, only 0,0018% of values are missing. To solve these missing values, the next valid measurement was imputed. This could be done due to the small number of missing values and the fact that we want to detect the anomalies in the time series. When checking the measurements of the calibrated data many more missing values are found. For each sensor between 25 to 40% of all calibrated measurements are missing. Investigating how the missing values are spread out over the time series reveals that missing values occur mostly in large blocks, see Figure 4. For certain periods measurements sometimes cannot or are not calibrated. However, there are also individual missing values, as seen by the incomplete lines in Figure 4. Since the missing values are dispersed all over the time series, they need to be filled. Otherwise the autoregressive model will not work. Taking the rows out is not an option as otherwise the sequences needed for the autoregressive model cannot be created. The method used will be to fill in the missing values with the median of a particular series. As seen from the statistics in Table 1 the mean of the measurements is very skewed due to high outliers. Therefore, the median is a better value to use as a 'normal' value for the distribution. The median imputation will be done for each sensor, using the median derived from that particular time series. After the anomaly detection the measurements that were imputed will not be taken into account in the evaluation.

Further preprocessing was not done. Normalization or scaling would not be appropriate as it would make the scale more similar to each other and help in the training for both models. However, the goal of the models is to detect anomalous points, for which the true scale needs to be kept intact. Normalization or scaling would hinder this goal. Furthermore, resampling was also not necessary, as the hourly rate at which the measurements were done as the hourly rate at which the measurements were done are sufficient to show the increase in pm2.5 concentration due to residential wood burning. (Hellén et al., 2008)

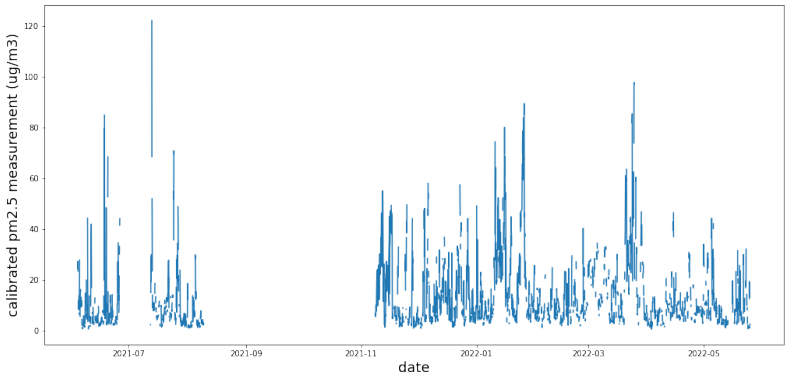

Figure 4: Sensor 4976 time series calibrated measurements.

4.5 *Autoregressive Model*

The anomaly detection of the calibrated data will be done using an autoregressive model. This type of model uses a sequence of the timeseries to statistically predict the next value. This method is based on the assumption that measurements preceding and following each other are correlated. To test this assumption the correlation between the calibrated masurements will be explored. Figure 5 shows a lag plot showing the correlation between the sensor measurements. A clear correlation can be seen, which holds the assumption.

With the relationship between following measurements, the autoregressive model uses Equation 1 to determine the next value following the sequence. The model needs to be trained on the time series of each sensor and can then predict the last part of the time series. In the data selection all sensors with less than 672 (four weeks) measurements were filtered out. Therefore, the last month of observations can be selected as the test data, on which the anomalies can be detected. Because of the choice of a whole month, ten sensors were too short to train and could not be used in the autoregressive model. The final amount of sensors that is used in the autoregressive model is 850.

$$X_t = \sum_{i=1}^{p} a_i \cdot X_{t-i} + c + \epsilon_t \tag{1}$$

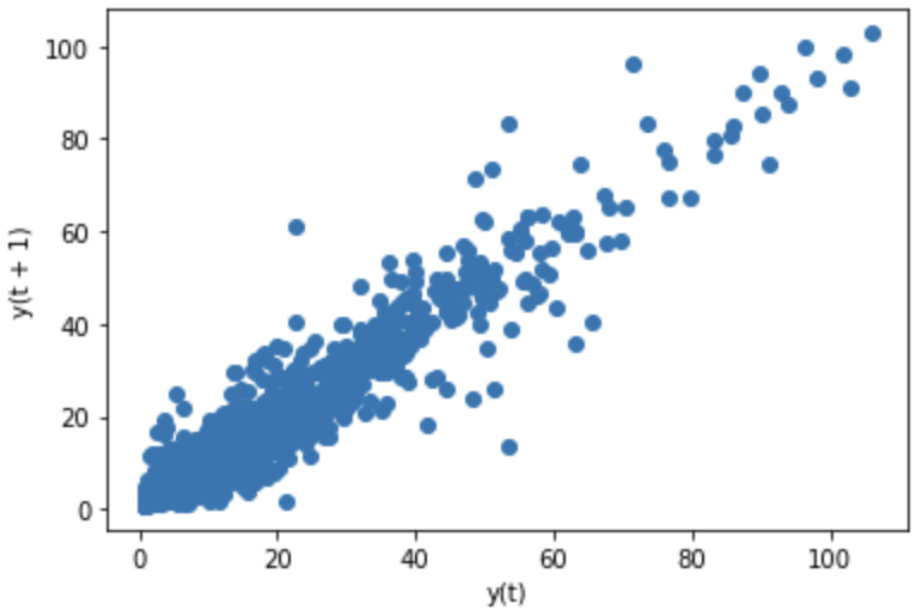

Figure 5: Lag plot of the calibrated data of sensor (ID = 5786).

4.6 *Anomaly Detection Calibrated Data*

From these prediction values the anomalies can be classified. By comparing the difference between the prediction and the calibrated measure and comparing that to a threshold the anomalous points are classified. The threshold for this classification has been set to 15. The choice for this threshold was made by looking at the published RIVM data about pm2.5. With high-quality sensors they measure a yearly average of 9.78 ug/m3 per hour Luchtmeetnet. As mentioned in the introduction this is lower that the 25 ug/m3 that is the maximum that the hourly average should be. The difference between these measures will be the threshold used in this paper, 15 ug/m3. A rise of 15 ug/m3 should indicate a rise from a normal level of pm2.5 to a abnormal, too high amount of pm2.5.

4.7 *CNN Model*

The main model used will be a convolutional neural network (CNN). A CNN has mostly been used in applications with images as it specializes in data that is like a grid. ADD CITATION Therefore, a time-series is also a viable application as sequences can be seen as one-dimensional grids. As mentioned in (Munir et al., 2018) a CNN is also good when training with smaller data sets and is generally good at generalizing the trained model to new data sets. The model structure that will be used in this thesis is

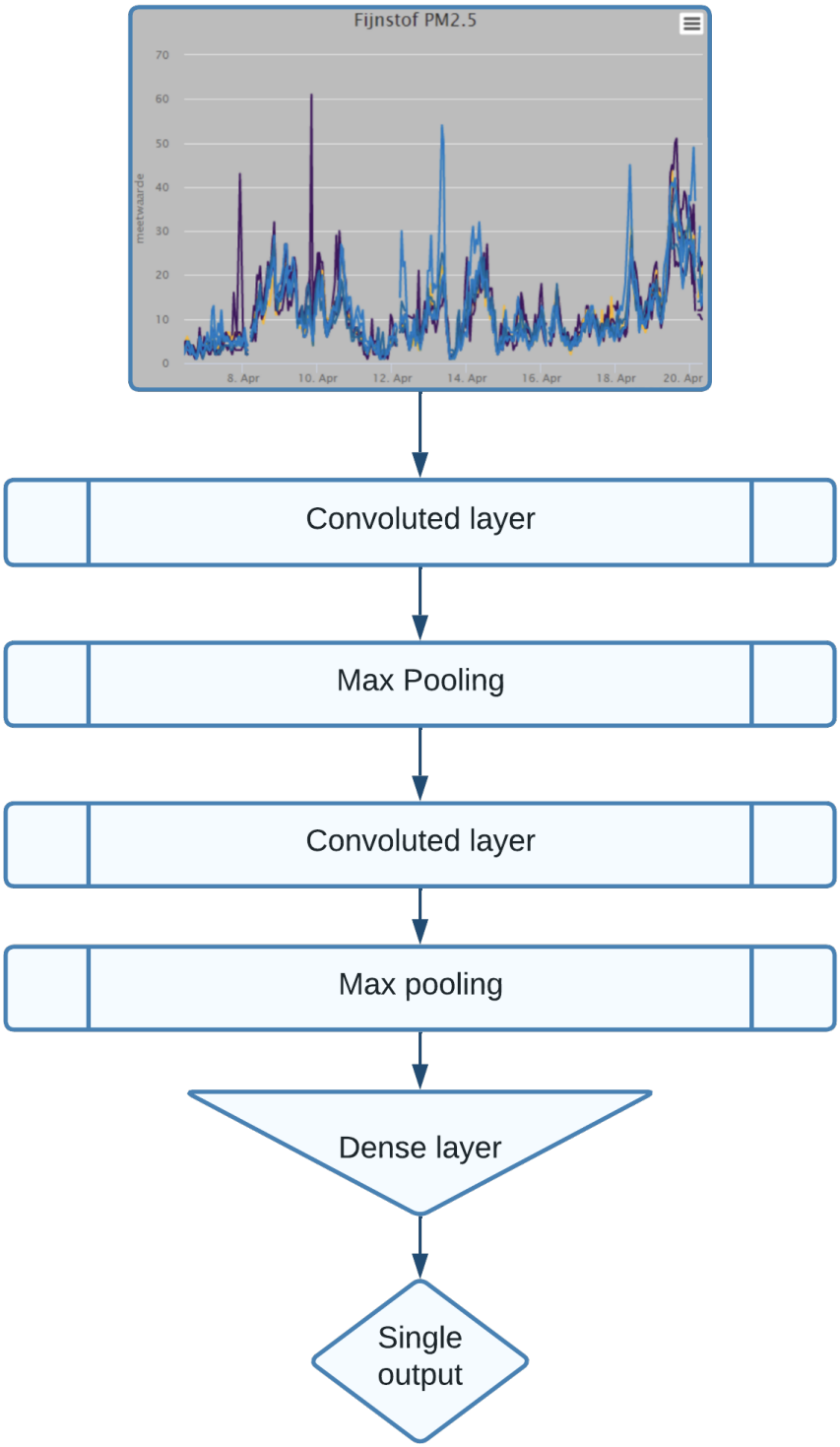

Figure 6: CNN model architecture. Modeled after (Munir et al., 2018). Time series from Samen Meten (CC BY 4.0).

the same architecture as used in (Munir et al., 2018). Figure 6 shows the different layers that are used. The used model consists of two convoluted layers followed by two max pooling layers. Latly, a dense layer creates the prediction output, a single value.

The model works as follows. First, the data is split into a train set and a test set. This is done by setting a date, two months before the last measurement, before which all the measurements will be in the training set and measurements after will be in the test set. From these two sets small windows are assigned, sequences of the time-series. The windows of the training data are selected at random from the collection of sensors. The windows of the test set are assigned in order. The length of the windows will be 24, a full day of measurements. As this will give a fluctuation sequence of a full day, showing the changes in humidity, which is the main influence on the measurements. (Bartonova et al., 2019)

The windows of the training subset will be the input-data for the CNN model. The prediction should be the measurement directly following the sequence, see Equation 2.

$$\{x_1, x_2, x_3, x_4 ... x_{length}\} \rightarrow x_{length+1} \tag{2}$$

The train set is further divided into a train set and a validation set. The first 80% original training set is the actual training set. The other 20% is the validation set.

When the model starts to train it uses the training data windows. In these windows filters of a certain assigned size will look over the sequence for certain features. An activation function will be used to assign if the filter has found the feature or not. The output of this activation function will be used as input for the next layer. In the case of this thesis the ReLu (Rectified Linear Unit) activation function is used. The choice for this activation function is to follow the structure in (Munir et al., 2018). A main reason why it is one of the most used activation functions is its speed.

With the output from the activation function the max pooling layers are passed through. Here the maximum parameters are filtered out and kept. It reduces the number of values for the next layer. The last layer is a fully connected layer, the dense layer, where all features are activated. Here a matrix multiplication is done using a bias offset. The result will be a single output, the predicted next value in the time series.

While training the model, the loss function that will be us is MAE (mean absolute error). This measures how far off the prediction is from the true value. Based on the MAE of the validation loss function the ideal number of epochs will be determined. The optimizer that is used in the model will be SGD (stochastic gradient descent).

4.8 *Anomaly Detection Uncalibrated Data*

From the predicted measures as a result of the CNN model, the anomalous point can be identified. This will be done by taking the difference between the predicted value and the actual measurement and checking to see if it is larger than the threshold. As with the anomaly detection on the calibrated data the threshold will be set at 15 (ug/m3). The same threshold will be used as both methods use the difference between the predicted measure and the actual measurements to determine the anomalous points. Therefore the same threshold needs to be used to make the models comparable.

4.9 *Evaluation*

The anomaly detection via prediction done on the uncalibrated data will be compared to the anomaly detection done on the calibrated measurements. For the anomalies detected on the calibrated data with the autoregressive model, only the points for which the data was not imputed will be used int he evaluation. Since the measurements on which anomaly detection have been done do not completely overlap for the two models, only evaluation will be done on the points which have had anomaly detection done on both the calibrated and uncalibrated measurements.

The anomalies detected on the calibrated measurements will be seen as the true values. Anomalies detected on the calibrated data are the true positive values, the measurements which are not anomalies are true negatives.

$$Precision = \frac{TruePositives}{TruePositives + FalsePostives} \tag{3}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

To evaluate the prediction anomaly detection method against anomaly detection on calibrated data, the overall precision, recall and F1-score will be used. This choice is made because the data set is imbalanced. These metrics will be determined over the whole resulting data set with detected anomalies, the subset of measurements which have had anomaly detection done on both the calibrated and uncalibrated measurements. Most of the data points will be seen as non-anomalous. Recall, precision and F1-score will give insight into the amount of overlap the two methods have. If the recall is high, the prediction method will find many of the same anomalies.

If the precision is high, we will not have identified many other anomalies. The F1- score will tell us how well the prediction method fits with the calibration method. These metrics combined with a confusion matrix will give an insightful answer to the research question.

5 RESULTS

In this section the anomalies from the baseline model, the anomaly detection combined with the autoregressive model will be compared to the anomalies detected using the CNN model with anomaly detection. The results will be shown in the form of a confusion matrix, the evaluation metrics. To gain more insight into the anomalies detected this section will end with a look at the time series with anomalies detected of a particular sensor.

5.1 *CNN prediction*

The CNN model was trained using a window 24 measurements long (a day of measurements), using two features and a filter of (8,8,8). The learning rate of the SGD was kept small, 0.01. First the ideal amount of epochs, based on the MAE loss value of the validation data was researched. In Figure 7 MAE loss of the training and validation data is shown. From this figure we can see that the more epochs have been done, the more the model learns on the training data. However, when evaluating on the validation data, the graph shows that more epochs do not improve the generalization capabilities of the model. The validation shows that after six epochs the model does not improve. The lowest validation loss is at epoch six, but from three onward are in the lower area. Since the prediction are used to detect anomalies in the time series, we do not want to overfit the model, as the outliers will be learned more and more. Therefore, three epochs is the best choice for this model.

5.2 *Anomaly Detection*

In 2 the results from the two methods of anomaly detection can be seen. The autoregressive model detects more anomalies than the method that uses the CNN model, more than twice as much. About 4% of measures from the classified data set have been classified as anomalous data point. Of the unclassified data set only 1,4% is calassified as anomalous data points.

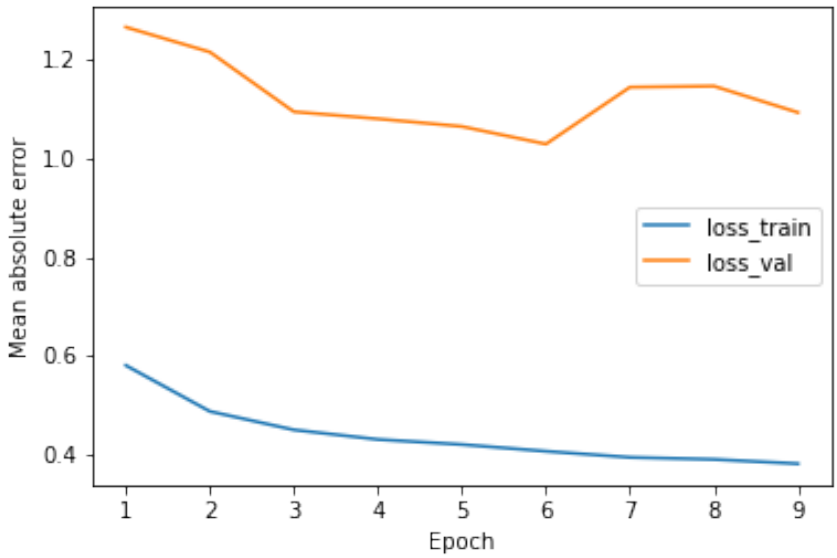

Figure 7: Mean absolute error loss for training and validation CNN

Table 2: Anomalies detected, CNN model, autoregressive model.

	Anomaly	**Not anomaly**
CNN model	3071	206252
Autoregressive model	8130	201193

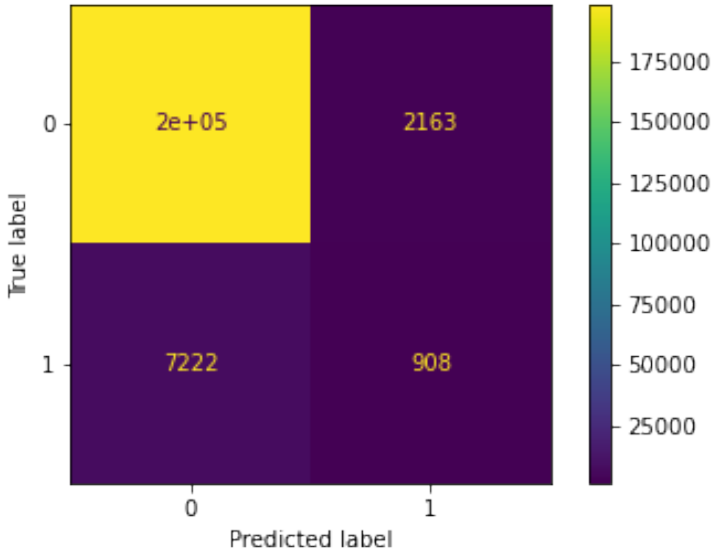

Figure 8: Confusion matrix

To find which anomalies the CNN model method detects compared to the autoregressive model, a confusion matrix of the anomalies is shown in Figure 8. Using the CNN model to detect anomalies gives the correct bias as the data set is imbalanced, detecting little anomalies. However, looking at the matrix we see that there are 2163 False Positive, 7222 False Negative and 908 True Positive classifications. The largest group, the True Negative has a value of 199030 and is compressed in the figure.

5.3 *Evaluation metrics*

With the anomalies from both models the evaluation metrics can be computed. In Table 3 the scores for the precision, recall and F1-score are shown. The precision score is low at 0.30. This means that 30% of the anomalies found with the CNN model were correct, according our baseline of anomalies detected by the autoregressive model. The recall is lower at 11%. When comparing the anomalies found by the CNN model to the autoregressive model, the CNN model only found 11% of the anomalies found by the autoregressive model. The overall F1-score of 0.16 is quite low. This indicates that the fit of anomalies detected with the CNN model on the anomalies detection of the autoregressive model is not good.

Table 3: Evaluation metrics

	Value
Precision	0.30
Recall	0.11
F1-score	0.16

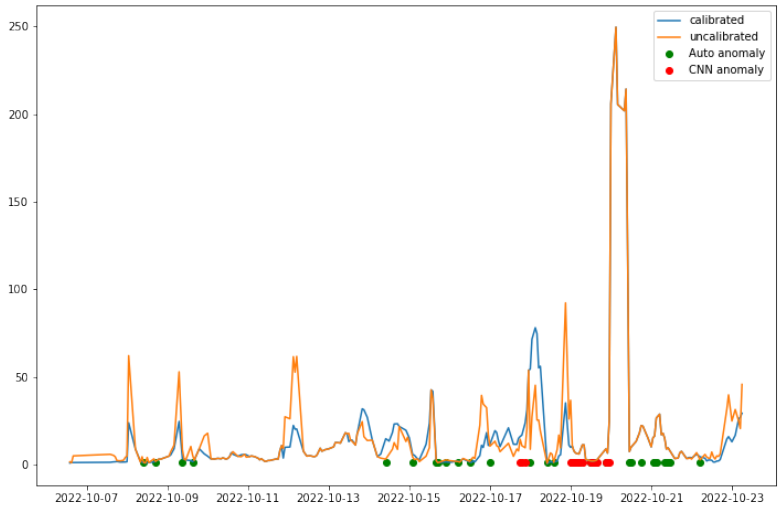

Figure 9: Times series sensor 6551, including anomalies detected.

5.4 *Exploration Anomalies Detected*

To look more deeply into the difference in what the two anomaly detection methods are doing Figure 9 shows the time series with calibrated and uncalibrated data for the sensor 6551. In the figure the anomalies detected by both methods are also shown. Both anomaly detection show anomalies that indicate steep slopes. However, both methods also seem to miss some high local maxima, around the 12th for example. Between the 19th of October and the 20th we can see many anomalies detected by the CNN model method. However, there does not seem to be a significant increase or decrease for the point between the local and global maxima. The same happens for the autoregressive method around the 16th. Many points that seem to be sudden steep increases or decreases are also not seen as anomalous points by either the CNN model or the autoregressive model.

6 DISCUSSION

In this section the results of thesis will be discussed. The main points that this section will go over will be to reiterate the goal of this thesis. Then the results will be interpreted to see if the CNN model can compare to the baseline model. Then different possible improvements on the method are discussed, with suggestions for further research that can be done on this topic.

The goal of this thesis was to find an alternative way to find high emission points in low-cost air quality sensors. The proposed technique was to use a deep learning model, CNN, to incorporate the environmental factors, which influence the measurements, into a prediction and use this to find high emission points. This would help giving insight into the emissions caused by wood burning stoves. These predictions would be compared to anomaly detection done on calibrated data to see if the proposed technique would be a suitable alternative, as calibration already takes the environmental factors into account.

From the results we have seen that the CNN model method does not come close to detecting the same anomalies as the autoregressive method. With a F1-score of 0.16 many different and wrong points were seen as anomalous. Anomalies were also missed. When looking closer at why this could be we see that both methods do not do what they were deemed to be able to do from the literature. We see in Figure 9 that after a local maxima, both methods keep classifying measurements as anomalies, even though there is no increase or decrease in the values. For the autoregressive method (Aggarwal, 2017) states that the breaks in temporal continuity are classified as anomalies. However, here we see that the predictions are influenced too much by the local maxima and keep the predictions high for long after the maxima. Thus, more measurements are seen as anomalies. The same thing seems to be the case for the CNN model. For the CNN model a smaller window could possible improve this.

The size of the data set could be increased. Due to time constraints a smaller number of sensors were included in the data set than were available on Samen Meten. With a larger data set, it could be possible that the model could be trained better and validated and tested on a larger set. This could result in greater generalization capabilities and better anomaly detection when used on the test data. The data set was also limited due to the sensors having measured for no more than 18 months. Because of the way the API is set up the newest added sensors are collected first. Creating a bigger data set with more sensors would therefore also have as a consequence that those sensors would have been active for longer. The seasonality could therefore possibly be better taken into account in the

model. However, this would create problems with running the model on most regular devices. The size of the data set used in this thesis already was almost 7 GB. Increasing the data set would require specialized equipment to run the CNN model.

The low F1-score is not a strange outcome for supervised anomaly detection on time series. Both (Munir et al., 2018) and (Gao et al., 2020) report for their worst-working models F1-scores between 0.2-0.4. However, this still makes the F1-score of 0.16 very low and not indicative of a properly working model. A reason why the CNN model worked worse in this paper compared to the research done by (Munir et al., 2018) could be the data set. The influences of environmental factors on the low-cost pm2.5 sensors could be different per sensor. This has not been revealed by related works but is an interesting topic to research further. The data set also showed some inconsistencies in the calibration of the data. Some sensors had calibrated data that did not differ from the uncalibrated data at any point. Since the calibration method used was not stated clearly in the documentation of the data set, it is possible that the calibration was not up to the standards that was assumed. With these two points in mind, a different data set could be tested to see if that improves the use. The differences between the data sets could then be used to examine why the method used in this paper did not prove sufficient.

This difference per sensor could be explored in a different way as well. In this paper the method was used to train the CNN model on the complete collection of measurements, from all sensors. Improvements could possibly be made by training the CNN model on a single time series. This could better learn the fluctuations of that particular sensor and the unique way the environmental factors influence that sensor. This alternative method could also improve the ability to compare the models. The method used for the autoregressive model was to train the model for each sensor separately. This could have led to a greater discrepancy between the anomaly detection done on the autoregressive model and the anomaly detection done on the CNN model.

This paper shows that for low-cost pm2.5 sensors more research needs to be done. Both from the workings and influences of environmental factors on these sensors. To use these sensors to give more insight into local wood burning emissions a lot higher precision and recall is necessary. However, because of limitations in the research done in this paper, the use of deep learning to extract environmental factors from the time series, cannot be cast aside either. To possibly use these methods to gain more insight into these sudden and local emissions due to residential wood burning a different methodology will be needed. The recommendation based of this paper for a way to test the uses of the CNN model, would be

to create an experimental setup. High quality sensors should be used to find the true anomalous point, the true high emissions. From this labeled data the evaluation of the models could be done better and more insights into what the methods miss or do not learn can be seen. In this paper also see a large discrepancy in the points that were found by the two different methods. The CNN found points anomalous that were not found by the autoregressive model. This is another indication that a better baseline is needed to explore which points are classified as anomalies. With a baseline with true high emission points more insight can be given in the way the anomalies are assigned. With these insights improvements can be made to the model. In this thesis the choice of baseline and impossibilities of the data set made improvements to the model difficult as the autoregressive method was also showing strange classification behaviour.

This paper has given a novel insight into the complications that the use of these low-quality sensors bring and the assumptions that do not hold due to the workings of these sensors. More research will need to be done with a different methodological setup and more research will need to be done regarding the influences of environmental influences on these particular types of low-cost pm2.5 sensors.

7 CONCLUSION

To conclude this thesis the research question will be examined and possible future work and improvements summarized.

For the research question: "How well can CNN prediction anomaly detection on pm2.5 low-cost sensor time series data be used instead of calibration that takes environmental factors into account to find true high emissions?" the conclusion is that anomaly detection done with CNN prediction is not a suitable way to detect high emission concentrations, as done with the method used in this paper. The F1-score of 0.16 makes clear that the anomalies found by using the CNN model do not accurately represent the anomalies detected with the baseline autoregressive model. Furthermore, the assumptions about the environmental influences on the low-cost pm2.5 sensors do not hold up in this paper, which indicates that more research needs to be done on these influences. Because of the uncertainty with using a anomaly detection method as a baseline further research should follow an experimental setup where the true anomalies, high emission points are known. This paper has set up a good pathway to test the methods that have been explored here for the first time. With further research it will be possible to definitively conclude if the proposed CNN method could be a suitable way to incorporate environmental factors into the pm2.5 measurements.

REFERENCES

Aggarwal, C. C. (2017). Time series and multidimensional streaming outlier detection. In *Outlier analysis* (pp. 273–310). Springer.

Ali, S., Glass, T., Parr, B., Potgieter, J., & Alam, F. (2020). Low cost sensor with iot lorawan connectivity and machine learning-based calibration for air pollution monitoring. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Bartonova, A., Castell, N., Colette, A., Schneider, P., Viana, M., Voogt, M., ... others (2019). Low cost sensor systems for air quality assessment. *Possibilities and Challenges.| NILU–Norsk Institutt for Luftforskning. Available online: https://www. nilu. no/pub/1761931/(accessed on 2 March 2021)*.

Chen, L.-J., Ho, Y.-H., Hsieh, H.-H., Huang, S.-T., Lee, H.-C., & Mahajan, S. (2017). Adf: An anomaly detection framework for large-scale pm2. 5 sensing systems. *IEEE Internet of Things Journal*, *5*(2), 559–570.

Favez, O., Cachier, H., Sciare, J., Sarda-Estève, R., & Martinon, L. (2009). Evidence for a significant contribution of wood burning aerosols to pm2. 5 during the winter season in paris, france. *Atmospheric Environment*, *43*(22-23), 3640–3644.

Gao, J., Song, X., Wen, Q., Wang, P., Sun, L., & Xu, H. (2020). Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks. *arXiv preprint arXiv:2002.09545*.

Hellén, H., Hakola, H., Haaparanta, S., Pietarila, H., & Kauhaniemi, M. (2008). Influence of residential wood combustion on local air quality. *Science of the total environment*, *393*(2-3), 283–290.

Jácome-Galarza, L.-R., Realpe-Robalino, M.-A., Paillacho-Corredores, J., & Benavides-Maldonado, J.-L. (2022). Time series in sensor data using state-of-the-art deep learning approaches: A systematic literature review. *Communication, Smart Technologies and Innovation for Society*, 503–514.

Jin, X.-B., Zheng, W.-Z., Kong, J.-L., Wang, X.-Y., Zuo, M., Zhang, Q.-C., & Lin, S. (2021). Deep-learning temporal predictor via bidirectional self-attentive encoder–decoder framework for iot-based environmental sensing in intelligent greenhouse. *Agriculture*, *11*(8), 802.

Munir, M., Siddiqui, S. A., Dengel, A., & Ahmed, S. (2018). Deepant: A deep learning approach for unsupervised anomaly detection in time series. *Ieee Access*, *7*, 1991–2005.

Okafor, N. U., Alghorani, Y., & Delaney, D. T. (2020). Improving data quality of low-cost iot sensors in environmental monitoring networks using data fusion and machine learning approach. *ICT Express*, *6*(3), 220–228.

Veiga, T., Ljunggren, E., Bach, K., & Akselsen, S. (2021). Blind calibration of air quality wireless sensor networks using deep neural networks. In *2021 ieee international conference on omni-layer intelligent systems (coins)* (pp. 1–6).

Watne, Å. K., Linden, J., Willhelmsson, J., Fridén, H., Gustafsson, M., & Castell, N. (2021). Tackling data quality when using low-cost air quality sensors in citizen science projects. *Frontiers in Environmental Science*, 461.