Evaluation of calibration techniques in low-cost air quality sensing

Kasimir Aula

Helsinki June 10, 2019

Master's Thesis

UNIVERSITY OF HELSINKI
Master's Programme in Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Studieprogram — Study Programme
Faculty of Science	Master Programme in Computer Science

Tekijä — Författare — Author
Kasimir Aula

Työn nimi — Arbetets titel — Title

Evaluation of calibration techniques in low-cost air quality sensing

Ohjaajat — Handledare — Supervisors
Eemil Lagerspetz, Petteri Nurmi, Sasu Tarkoma

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	June 10, 2019	48 pages + 1 appendix

Tiivistelmä — Referat — Abstract

Air pollution is considered to be one of the biggest environmental risks to health, causing symptoms from headache to lung diseases, cardiovascular diseases and cancer. To improve awareness of pollutants, air quality needs to be measured more densely. Low-cost air quality sensors offer one solution to increase the number of air quality monitors. However, they suffer from low accuracy of measurements compared to professional-grade monitoring stations.

This thesis applies machine learning techniques to calibrate the values of a low-cost air quality sensor against a reference monitoring station. The calibrated values are then compared to a reference station's values to compute error after calibration. In the past, the evaluation phase has been carried out very lightly. A novel method of selecting data is presented in this thesis to ensure diverse conditions in training and evaluation data, that would yield a more realistic impression about the capabilities of a calibration model.

To better understand the level of performance, selected calibration models were trained with data corresponding to different levels of air pollution and meteorological conditions. Regarding pollution level, using homogeneous training and evaluation data, the error of a calibration model was found to be even 85% lower than when using diverse training and evaluation pollution environment. Also, using diverse meteorological training data instead of more homogeneous data was shown to reduce the size of the error and provide stability on the behavior of calibration models.

ACM Computing Classification System (CCS):
Modeling and Simulation → Model development and analysis → Model verification and validation
Machine Learning → Learning Paradigms → Supervised Learning → Supervised Learning by Regression
Hardware → Communication hardware, interfaces and storage → Sensor applications and deployments

Avainsanat — Nyckelord — Keywords
air quality, low-cost sensors, sensor calibration, model validation

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Thesis for the Algorithms study track

Contents

1 Introduction

Air pollution poses one of the biggest environmental risks to health according to a report by the World Health Organization (WHO) [38]. The health issues poor air quality can cause range from headache [46], lung and cardiovascular diseases [4, 20] to cancer [40]. However, air pollution has even been linked to infant mortality [32], autism [24] and decreased cognitive performance [51]. Indoor and outdoor air pollution is estimated to have caused the death of around 7 million people in 2012 [45].

Air pollutants in general are substances in the air that can have harmful effects on humans and the environment. The most harmful ambient air pollutants are gases such as NO_2, O_3 and SO_2 and particulate matter (PM). Gaseous pollutants are either directly or indirectly emitted from combustion [37], whereas particulate matter consists of respirable particles such as sulphate, nitrates, ammonia, black carbon and mineral dust [3]. PM is also likely to contain micro plastics and their concentration is expected to increase in the next decades [39]. PM is often discussed with respect to the size of the diameter of the particle (in micrometers). For example $PM_{2.5}$ and PM_{10} refer to particles that are respectively no greater than 2.5 and 10 micrometers in size.

Indications of Air Quality

For improving monitoring on the amounts of air pollution, air quality standards and air quality indexes have been developed. This section goes through their purposes, and the differences between the two.

Air Quality Standards

To help protect public health against polluted air on a global scale, WHO has released guidelines (Table 1) for the amounts of ambient air pollution that may cause health issues [37]. Similar *air quality standards* have been set by every country to support public health, but they are not bound to the WHO guidelines. Air quality standards specify threshold values for each pollutant that should not be exceeded, usually separately for short and long-time period. As an example the air quality standards of China[1] and Finland are shown in Table 1. All thresholds except NO_2

[1]The air quality standards in China are divided into two classes. This is class 2 corresponding to standards in urban and industrial areas. Class 1 applies to special areas, such as national parks.

Table 1: Guidelines of exposures for different air pollutants by the World Health Organization [37] and the air quality standards of Finland [6] and China [36].

Air quality standards for each pollutant (µg/m^3)		WHO	FIN	CHN
$PM_{2.5}$	annual mean	10	25	35
	24-h mean	25	NA	75
PM_{10}	annual mean	20	40	70
	24-h mean	50	50	150
Ozone (O_3)	8-h mean	100	100	160
Nitrogen dioxide (NO_2)	annual mean	40	40	40
	1-h mean	200	200	200
Sulphur dioxide (SO_2)	24-h mean	20	125	150
	1-h mean	NA	350	500
	10-min mean	500	NA	NA

differ between the two countries, and many of them are greater than the guidelines set by WHO. This implies that falling below the national threshold values does not necessarily guarantee an equally healthy environment to live in.

Air Quality Indexes

Compared to the threshold value of an air quality standard, a more fine grained way to indicate the amount of pollutants is to use an *air quality index*, which in Finland is set by the Helsinki Region Environmental Services Authority and the National Institute for Health and Welfare (shown in Table 2). Air quality indexes are values indicating air quality level by giving descriptive name to pollutant values that can be used to easily classify the air quality on a discrete scale, such as from *good* to *bad*. Similarly to air quality standards, air quality indexes vary between countries. For example, the PM component of the air quality indexes of Finland and China can be seen in Table 3. The differences in air quality of these two countries are extreme. In case of $PM_{2.5}$, *lightly polluted* in China corresponds to *very poor* in Finland, and the highest category of China goes beyond the highest values in the Finnish index. Air quality index is far from being well defined, as it might be defined individually on a national level, by a company, or even by a single study [10], but the idea is universal — to have different quality levels, such as from good to bad, where each level and pollutant have a defined concentration range. Air quality level is usually defined by the maximum concentration for a single pollutant yielding the worst level from

Table 2: Index values of different compounds (µg/m^3) in Finnish Air Quality Index by the Helsinki Region Environmental Services Authority and the National Institute for Health and Welfare [28].

Index classification	SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
Good	0-20	0-40	0-20	0-10	0-60	0-4000
Satisfactory	20-80	40-70	20-50	10-25	60-100	4000-8000
Fair	80-250	70-150	50-100	25-50	100-140	8000-20000
Poor	250-300	150-200	100-200	50-75	140-180	20000-30000
Very poor	>300	>200	>200	>75	>180	>30000

Table 3: Comparison of PM values (µg/m^3) from the air quality indexes of Finland [28] and China [17].

Pollutant	Finland		China	
$PM_{2.5}$	Good	0-10	Excellent	0-35
	Satisfactory	10-25	Good	35-75
	Fair	25-50	Lightly Polluted	75-115
	Poor	50-75	Unhealthy	115-150
	Very Poor	> 75	Very Unhealthy	150-250
	–	–	Severely Polluted	250-500
PM_{10}	Good	0-20	Excellent	0-50
	Satisfactory	20-50	Good	50-150
	Fair	50-100	Lightly Polluted	150-250
	Poor	100-200	Unhealthy	250-350
	Very Poor	> 200	Very Unhealthy	350-420
	–	–	Severely Polluted	420-600

the index. Providing such information about air quality from different locations can help reduce exposure to polluted air, since exposures of an hour or even less can be harmful, as shown in Table 1.

Current Air Quality Monitoring Regulations

To provide a standard level of surveillance for air quality, the European Union (EU) has set a minimum number of monitoring stations in Europe to protect people, and vegetation of the member countries [3]. The Air Quality Directive 2008/50/EC imposed in 2008, states that the minimum number of monitoring stations in cities is 1 station/ 1M people [3, appendix V]. For some pollutants (SO_2, NO_2, $PM_{2.5}$ and PM_{10}) the number of stations outside cities is set to 1 station/20,000 km^2 [3, ap-

pendix V], where as for some (O_3) the requirement is lower, 1 station/ 100,000 km^2 [3, appendix IX]. The minimum amount of measuring stations set by the EU results in sparse monitoring of air quality. The problem is that sparse air quality monitoring is not capable of capturing hotspots of polluted air that can exist in a very small area [13, 41, 44]. This may result in high differences in pollution exposure of inhabitants [14], which is alarming since health impacts of short term exposures of highly polluted air are still unknown [30]. For example, $PM_{2.5}$ has exhibited a large spatial variability on a scale of tens of meters [44], which encourages measuring air quality spatially more densely than what is currently required.

Air Quality Monitors

Air quality is monitored due to various reasons, which is why accuracy and price between air quality monitors vary. For example there are scientific interests of improving understanding of air pollutants, and commodity devices used by consumers. Unfortunately the trade-off between accuracy and price seems to follow the air quality monitors.

Most expensive ones are professional-grade air quality monitoring stations with the price over 1M euros. An example of such a station can be seen in Figure 1 (top-left). On top of being expensive to acquire, they also require constant calibration and maintenance, that alone causes additional costs that can rise to levels of hundreds of thousands of euros [16, 27]. However, their advantage is that they offer reliable information about the surrounding environmental and pollution conditions. Stations of this level are often used for scientific and governmental purposes, but there are also some companies that offer weather services based on these stations, for example Foreca[2].

Monetary costs are not the only problem with professional-grade air quality monitoring stations, but also the fact that they fail to provide information on places where air quality might vary suddenly for example due to infrastructural reasons. Smaller devices have been developed to monitor air quality from locations where professional-grade stations have not existed. They respond to the challenge of measuring air quality spatially more densely. An example of such a device is the Vaisala AQT420 (in Figure 1 top-right). It measures surrounding meteorological conditions (temperature, humidity and pressure), pollution gases (NO_2, O_3, SO_2 and CO)

[2]https://corporate.foreca.com

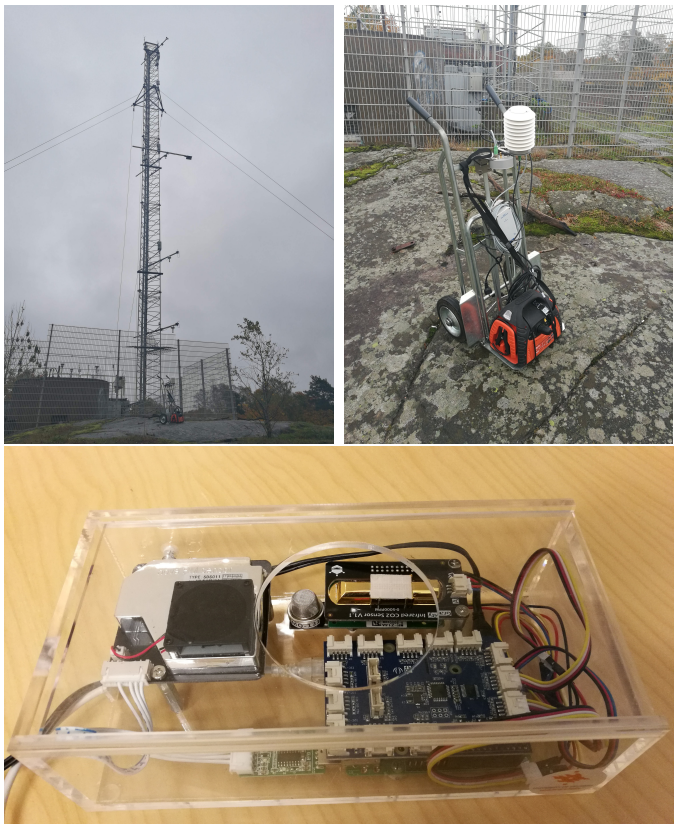


Figure 1: Top-left: an accurate SMEAR tower monitoring station located in Kumpula, Helsinki. Top-right: a Vaisala AQT420 mid-range air quality monitor attached to an external power source. Bottom: the low-cost sensor used in this research.

and particulate matter ($PM_{2.5}$ and PM_{10}) in the air. Although being significantly more affordable than the air quality monitoring stations, these semi-accurate monitors usually cost some thousands of euros (the Vaisala AQT420 costs approximately 5,000 euros). This makes them more an industrial product than a commodity device, limiting their availability for the masses and denser spatial air quality monitoring. In addition, sensors of this price point already suffer from decreasing accuracy of measurements [7, 31].

With decreasing price, there are air quality sensors designed for individual consumers. These sensors use low-cost components, which enable low production cost and selling price. Monitoring stations of this grade usually cost hundreds of euros making them available for a large consumer base. The benefit of these sensors is the high spatial frequency of measuring air quality they enable. They can be fit to

places previously unreachable to air quality monitoring stations, for example among pedestrians to measure daily exposures to air pollution. Unfortunately, low-cost air quality monitors are sensitive to environmental conditions [34] and pollutant cross-sensitivities [12] making their readings unreliable. Therefore special attention must be paid to the data, obtained from these devices.

Towards Denser Air Quality Monitoring

As current technology enables building reliable and accurate air quality monitoring stations, in practice, accurate monitoring stations are very expensive and not portable. This is why they are not suitable for dense air quality monitoring, and so cannot be used for determining air pollution hotspots. Low-cost sensors provide a solution to the economical and portability limitation that accurate air quality monitoring stations face, and they can be used for denser deployments required for detecting pollution hotspots. Unfortunately, low-cost sensors are sensitive to weather conditions [34], and their accuracy is worse than that of air quality monitoring stations [7]. Big measurement errors caused by decreased sensor accuracy report incorrect air quality conditions. Relying on inaccurate measurements can expose people to poor air quality causing several types of health issues [4, 20, 40, 46].

To acquire more accurate measurements, either low-cost sensors need technological advancement or more expensive sensors have to be used. One option for improving current level low-cost sensors is to calibrate them periodically in a laboratory by a calibration authority service[3] to ensure higher level of accuracy. However, the periodic use of calibration authority services becomes unfeasible when using a large number of low-cost sensors such as what is required to measure air quality densely. To perform on-site calibration to low-cost sensors, machine learning methods can be applied to learn the required data corrections. This thesis focuses on the application of machine learning methods to improve the accuracy of low-cost sensors.

Low-cost air quality sensor calibration has been studied extensively with successful applications of techniques such as Artificial Neural Network [9, 10, 43, 48] and Random Forest [8, 53], where the calibration models have shown to improve the accuracy of low-cost sensor measurements. This work will apply these techniques that have shown to be successful in the past with an intention to reduce the error between measurements of a low-cost air quality sensor and measurements of a

[3]https://www.mikes.fi/en/calibrations

professional-grade air quality monitoring station.

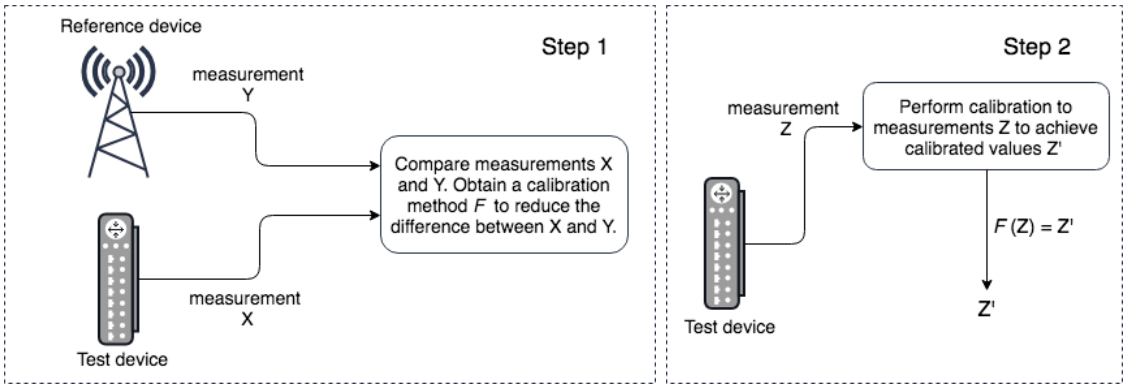
Research Questions

In most works [8, 10, 15, 43, 53], reference concentration values are assumed to be available when determining a calibration model. After a calibration model has been defined, low-cost sensors are assumed to provide reliable measurements. If the sensor cannot provide reliable measurements without comparison with a reference instrument, it has little utility. Therefore, the calibration model must be trusted to perform as required. To gain confidence on the performance, the calibration model must be adequately validated. In past studies of low-cost air quality sensor calibration, evaluation phase has been carried out lightly without considering vastly varying conditions in pollution levels nor in meteorological conditions.

In this thesis state-of-the-art calibration techniques are applied to low-cost sensor measurements, and the results are evaluated considering multiple changes in the surroundings i.e., pollution level and meteorological changes. Through extensive evaluation this thesis answers the following research questions.

RQ1 How well do current state-of-the-art calibration techniques generalize?

RQ2 How can calibration data be selected in a way that better generalizes?



2 Calibration of Sensors

Calibration in measurement technology refers to comparison of measurements between a test device and a reference device [1]. Determining the difference between the two devices allows adjusting test device's measurements to be closer to those of the reference device. Information about the differences can be used to create a predictive model that infers adjustments also for unseen measurements. Such a model is called a *calibration model* in this thesis.

Obtaining a calibration model requires example data, where the information about possible adjustments can be extracted. This is similar to what is called *training* in machine learning, where data is used to create as descriptive mapping as possible between input and output pairs.

Machine Learning

Machine learning can be characterized broadly as any computer program that improves its performance at some task through experience [35]. The process of machine learning, which humans consider as learning, is actually an algorithm minimizing an objective function by changing certain parameters. The information on how the parameters should be updated is extracted from observations.

When training data consists of inputs and corresponding target values, the machine learning task is known as a supervised learning problem. The goal of supervised learning is to learn a model that, given an input, is able to produce an estimation close to the actual value. When the target is a categorical or a discrete value, then the task is called a classification problem. When the target values are continuous, the task is called a regression problem. A classic supervised learning classification

task is digit recognition where the input vector represents values of each pixel in an image, and the target value corresponds to the digit shown in the image. An example of a supervised regression task is predicting the income of a person based on level of education, job title, living place and marital status.

The other main type of machine learning is called unsupervised learning. In unsupervised learning there are no corresponding target values for the input vectors. The goal then is to learn something essential to the data, such as discovering clusters that are formed by similar data points, or essential properties in the data that allow projecting it to small number of dimensions for better visualisation. An example of clustering task would be a task to find different customer groups based on their shopping behaviour. If shopping behaviour is observed through several parameters, for example time of purchase, total bill, number of items purchased, categories of purchased items and payment method, then visualizing data points w.r.t. all dimensions in two dimensions would correspond to projecting high-dimensional data into lower dimensions.

Important part of supervised machine learning algorithms is a part called model fitting or training. In training, a model is given part of the data available, often called training data, and the goal of the algorithm is to minimize its internal error by optimizing parameters. The training data is fed as an input to the model and the output values with initialized parameters are computed. The model estimates its internal error by computing the difference between its current outputs and the target values. By updating model parameters, internal error is minimized. This is essentially the phase that allows the model to improve its performance on a given task, in other words to learn.

In terms of machine learning, calibrating a low-cost sensor's measurements with a professional-grade monitoring station is a supervised learning regression task. The low-cost sensor's readings are treated as model inputs and the professional-grade monitoring station's readings are considered as targets. Several machine learning models were implemented to investigate their performance in context of this work.

Linear Regression

Linear regression is one of the simplest supervised learning approaches, which assumes that there is a linear relationship between two variables X and Y; i.e. Y is dependent on X and only the relationship needs to be defined. In the simplest case,

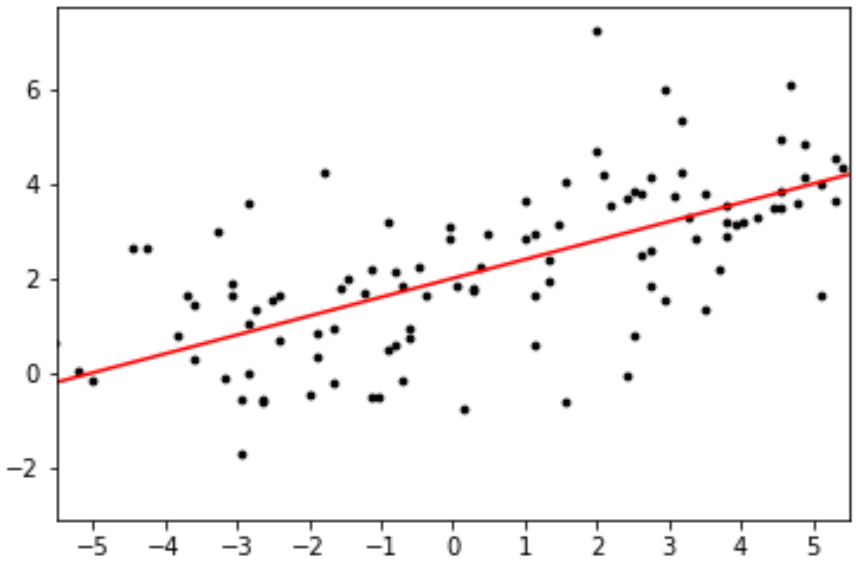


Figure 2: A simple linear regression with coefficients $w_0 = 2, w_1 = 0.5$.

where X and Y are one-dimensional, linear regression is defined as

$$y =: Y \approx WX = w_0 + w_1 x,$$

where the coefficients W define the slope and the intercept of the regression. The value of y approximates the true value as closely as possible, but the predicted output values do not match corresponding target values exactly. More precisely, the model outputs a prediction $\hat{y}$, that is defined by $\hat{y} = w_0 + w_1 x$. The difference between a prediction $\hat{y}$ and a target value y is called error, or residual, and noted often with e.

Before linear regression model can be used for prediction, the coefficients W must be determined. To do this, the model is first trained with data consisting of n input-output pairs $(x_0, y_0), \ldots, (x_n, y_n)$ with the objective of finding coefficients W that best describe the data. The best description is a line that minimizes error e for each data point. There is no way for all data to be on the regression line defined by the input data and the coefficients, so the best description is a line that minimizes the sum of all the errors such as in Figure 2. Since the direction of the error plays no role in the error, sum of squared errors (SSE) is normally selected as the error

function:

$$SSE(W) = \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

for n data points. This approach is called (ordinary) least-squares principle [29].

Linear regression looks the relationship between two variables, however often there are more variables that the target variable is dependent on. A multiple linear regression is a slightly modified version, where instead of looking for a one-to-one relationship, multiple variables are used to output an estimation. The definition of the multiple linear regression is similar to the one dimensional case, only instead X now has j dimensions, and the definition of Y is

$$Y \approx WX = w_0 + w_1 x_0 + \cdots + w_{j+1} x_j.$$

The advantage of a multiple linear regression versus the basic version is to include many features to explain the target variable and to benefit from more complex dependencies.

Decision Trees and Random Forest

A decision tree is a tree-like model, where the idea is to create a model that predicts the value of a target variable based on several input variables. The goal of a decision tree is to split data according to comparisons it makes. As the model is a tree, it has interior nodes that each represent a comparison that is made based on a feature of the data. After a single comparison, the data flows to the next level, where another comparison follows. After going through all the interior nodes, the leaves of a decision tree represent values to which the data points are mapped. Figure 3 shows, how a decision tree might be constructed in the problem of calibrating the values of a low-cost air quality sensor. A decision tree can be used for both classification and regression tasks. If the problem type is classification, the leaves represent classes. In regression tasks, the leaves represent real values.

Decision trees are prone to overfitting training data, resulting in poor generalization. One method to improve the generalizability is simplification of a decision tree, called *pruning* [29]. The goal of tree pruning is to reduce the number of interior nodes by minimizing the validation error through cross-validation. After pruning, decision trees have fewer interior nodes, meaning fewer comparisons, but the ones that remain are the essential ones for separating the data.

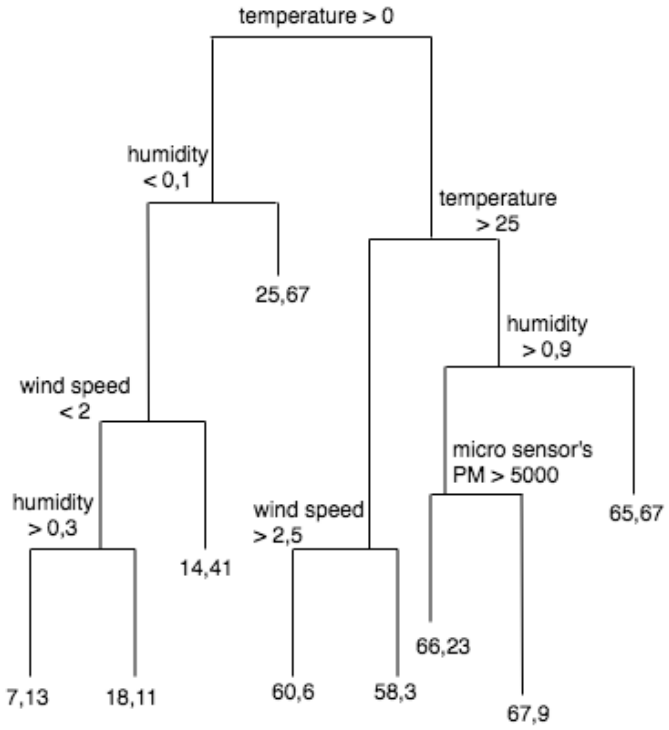


Figure 3: An example of how a decision tree might look like with the data used in this work.

Random Forest is an ensemble method that uses many low-accuracy decision trees to build a strong predictive model. The individual decision trees usually consist of small number of interior nodes. The strength of Random Forest lies in combining these decision trees. Each decision tree is created using randomly selected features. Given the randomly selected features, the decision tree tries to split the data according to those features to improve generality and accuracy.

Artificial Neural Network

Artificial neural network (ANN) is a computational method originating from biology, where it was inspired by how learning happens in brains [21]. The term network comes from having multiple functions combined to a single model so that they form a layered network structure. The core idea is to learn a mapping from input values to corresponding output values through a large number of simple modules called neurons. A network is said to be fully connected if there is a connection from each neuron to all neurons in the next layer. A representation of a simple fully connected

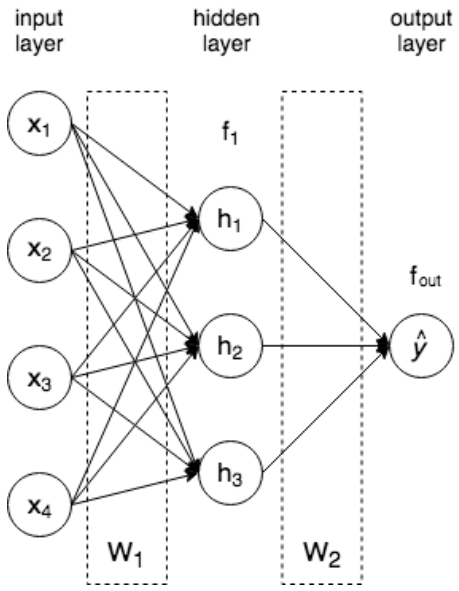


Figure 4: A simple ANN with an input layer, one hidden layer and an output layer. The input in the image consists of four features and the output is a single value.

neural network can be seen in Figure 4.

In a neural network model the input values flow through the network until they reach the final output layer. More specifically, for each data point, the input values form the first layer of the network, called the input layer. Each layer is assigned weights that are used to multiply the values of the neurons. Weights can be thought of as indicators telling how much should we scale the value of each neuron. After multiplying the input values with the weights, the products are passed through an activation function. The outcomes of the activation function form a hidden layer that can be thought as new input values for the next layer. The procedure can be repeated many times, and the number of layers in the architecture defines the depth of the network. If there are many hidden layers the network is said to be *deep*.

In case of an architecture such as in Figure 4 the input values are first multiplied by weights W_1, and passed on to activation function f_1. The outputs of the activation function f_1 form the second layer of neurons in the network. This layer is the first hidden layer. Next, the values of the first hidden layer are multiplied by weights W_2 and passed through activation function f_2. After the second hidden layer, the values are multiplied by weights and passed through activation function f_{out}. This produces a single output, noted $\hat{y}$, that is the model's prediction. So in case of a

structure like in Figure 4, an artificial neural network f^* is defined as

$$f^* :== f_{out}\left(W_2 h\right) = f_{out}\left(W_2 f_1\left(W_1 X\right)\right) = \hat{y} \quad .$$

Since the procedure is done for each data point that has a corresponding target value, the magnitude of the error can be computed with a chosen error function. The required corrections to the weights are computed using an algorithm called *backpropagation* [21, p. 200]. The weights are updated to reduce the error between all the predictions and the targets. The weighs can be thought to transmit the importance of each neuron in order to produce a certain output. This iterative process is continued until a determined condition fulfills, such as the number of iterations is reached, or the error no longer reduces.

One strength of neural netowrks lies in the activation functions. On top of linear functions, they can be chosen to be non-linear, such as the Rectified linear unit (RELU), sigmoid function or hyperbolic tangent function, which enable to create non-linear mapping from input values to output values. This is something a multiple linear regression for example, is not capable of doing. In addition, neural networks are universal approximators [26], so they can asymptotically approximate any target function as the number of neurons goes to infinity.

Neural networks often perform well since the model can be extremely flexible. By increasing the number of hidden layers and neurons in each hidden layer, it is possible to create a near-perfect depiction of training data. There are however issues related to the generalizability of such models, since a model that perfectly fits training data most likely performs much worse on unseen data due to overfitting. There are different techniques that can be used in the network's training phase to improve generalizability, such as early stopping and dropout.

ANN use also has downsides, such as the amount of data needed for the training. To learn how to adjust the weights properly, the network needs to see lots of data. Sometimes there might not be sufficient data for the ANN, resulting in an overly simplified model. In comparison, a linear regression can be used with a fraction of the data usually needed to train an ANN. Another weakness of ANN is the long training time of the models. Predictions can be computed rapidly, but if the model needs periodical retraining, then the computational cost might be intolerable.

Machine Learning Based Calibration

Both, calibration and machine learning models, use information about historical data to adjust new observations. Usually having a coherent dataset helps in determining the mapping between inputs and outputs. This means invalid data, for example caused by measurement errors, needs to be preprocessed to reduce its effect when determining the mapping. Preprocessing of data is often called *data cleaning*.

Pre-Calibration Data Cleaning

Studies rarely consider the first phase of cleaning sensor's output values, instead data that has already undergone transformations such as outlier removal, is sometimes referred as the raw data. Making data cleaning more transparent is vital for reproducibility of results since data cleaning incorporates important procedures before data can reasonably be considered as input data. To approach calibration of a low-cost sensor with reference sensor, the data must be first carefully examined, for example by plotting, to notice any obvious abnormalities that need to be cleaned to have a consistent dataset.

Values from both sensors must be considered reasonable. For example negative PM values are not reasonable given the units of the measurements are µg/m^3 and pieces/l in this thesis. However, negative values might carry important information about the characteristics of the data; mainly about the true increasing or decreasing of concentration where the scaling is simply off.

Missing values are another common problem, since they appear frequently due to abnormal value removal and various sensing and communication issues. The most robust way to handle missing values is to drop the point containing missing values in some feature. This way data contains only values that are actually measured by the sensor. However, this might be problematic, if the measurements contain many features, and many times one of the features is missing. This approach may easily result in significantly lower amount of data available. It might also bias the dataset to only contain certain type of data, for example if one of the sensors stops working in low temperature.

Interpolation is an alternative method for handling missing values. A linear interpolation can be applied, if the missing periods are short. Problems occur when missing value periods become long, since replacing long missing periods could cause high errors. In case longer missing periods are encountered, data from that period

should be discarded. Another application of interpolation is to match the time frequencies of the two sensors to have a correct number of input-output pairs with same timestamp of measurement.

3 Evaluation of Calibration Techniques

Low-cost sensors enable measuring air quality spatially more densely from an environment, which can increase the awareness of polluted areas. Low-cost sensors are however more sensitive to surrounding environmental conditions [34], which is why they might produce irrational values. Since the accuracy of low-cost sensors can be improved by calibration, many studies have investigated air quality sensor calibration [8, 10, 15, 47]. To rely on calibrated measurements, model verification must be done thoroughly in varying meteorological conditions and levels of pollution concentration that are known to affect sensor functionality, and also required by the regulatory air quality standards [2].

Sensor Calibration Techniques

The calibration of gases measured with low-cost sensors has been extensively researched [15, 23, 43, 47, 48, 53]. This is due to the fact that gases strongly react to different meteorological conditions, and strong domain knowledge can be applied. In previous studies Artificial Neural Networks (ANN) are found to give the best calibration results [15, 43, 48]. The reason why a linear regression is not able to perform well is that the relationship between a low-cost sensor and a reference station is often complex and non-linear. Using multiple features is common and one motivator for using an ANN. Another motivating reason for using an ANN [10, 48] is the ability to use a non-linear relationship between the low-cost sensors and the reference measurements, which can provide a better explanation of the complex relationship. There are also some studies focusing on calibration of particulate matter (PM) measurements from low-cost sensors [9, 10, 18, 19]. Many of them use some version of ANN possibly combined with another technique [9, 10, 19].

Although ANN seems to be the most commonly applied technique in low-cost sensor calibration, other machine learning techniques have been used too. Most prominently Random Forest (RF) regression has been applied in calibration due to its ability to take into account the cross-sensitivities between different pollutants [8, 11, 53].

As ANN and RF have shown to yield the best performance level in past studies on calibration of low-cost air quality sensors, they are considered as state-of-the-art calibration techniques in this thesis. To improve a low-cost sensor's accuracy their performance will be tested in this Thesis.

Evaluation of Calibration Models

The quality of training data plays an important role in machine learning tasks, since the parameters of the model are optimized only w.r.t. seen instances (Section 2). Optimally the calibration model would be fed a vast amount of continuous air quality data covering different weather types that the model would learn to recognize. However, the problem of this approach is that training data would contain uneven distribution of occurring weather phenomena and pollution concentrations. Certain events that strongly affect particulate matter (PM), might occur very rarely, whereas others with little or no effect on PM might be more frequent. For example, the model might learn dependencies between PM values and variables through spurious correlation. Alternatively, model having mainly seen low-level air pollution concentrations, might struggle predicting high concentration values. In other words, the confidence over the performance of a calibration model comes down to the model's ability to generalize in different environments.

In previous research, a common way of evaluating performance of a calibration technique is to train the model with a continuous period and compare the calibrated values to reference concentrations over another continuous period of time [9, 10, 15, 43]. Some have considered the effect of humidity by evaluating calibration models in different humidity levels [49], but other factors such as the effect of temperature, air pressure or wind speed have been neglected. Therefore, the effect of diverse environmental conditions is often overlooked. Continuous evaluation is also problematic, since data points close in time are not statistically independent and this has an effect when evaluating the performance of a model [22]. A solid method of selecting training data has been shown to not only improve the performance of the model, but also to increase the robustness and the generalizability of the model [25].

Regulatory standards for air quality monitoring require sufficient correspondence between measurement devices under different environmental conditions [2]. Current evaluations of low-cost air quality monitoring techniques have largely ignored these standards by simply concentrating on consecutive measurement periods without considering the diversity of environmental conditions. To ensure that low-cost calibration techniques can be evaluated fairly and in-line with necessary regulatory frameworks, a novel evaluation approach is proposed in this thesis for low-cost air quality monitoring technology.

To further motivate the need for improved evaluation, Figure 5 shows autocorrelation in the reference station's air quality data w.r.t. different environmental features.

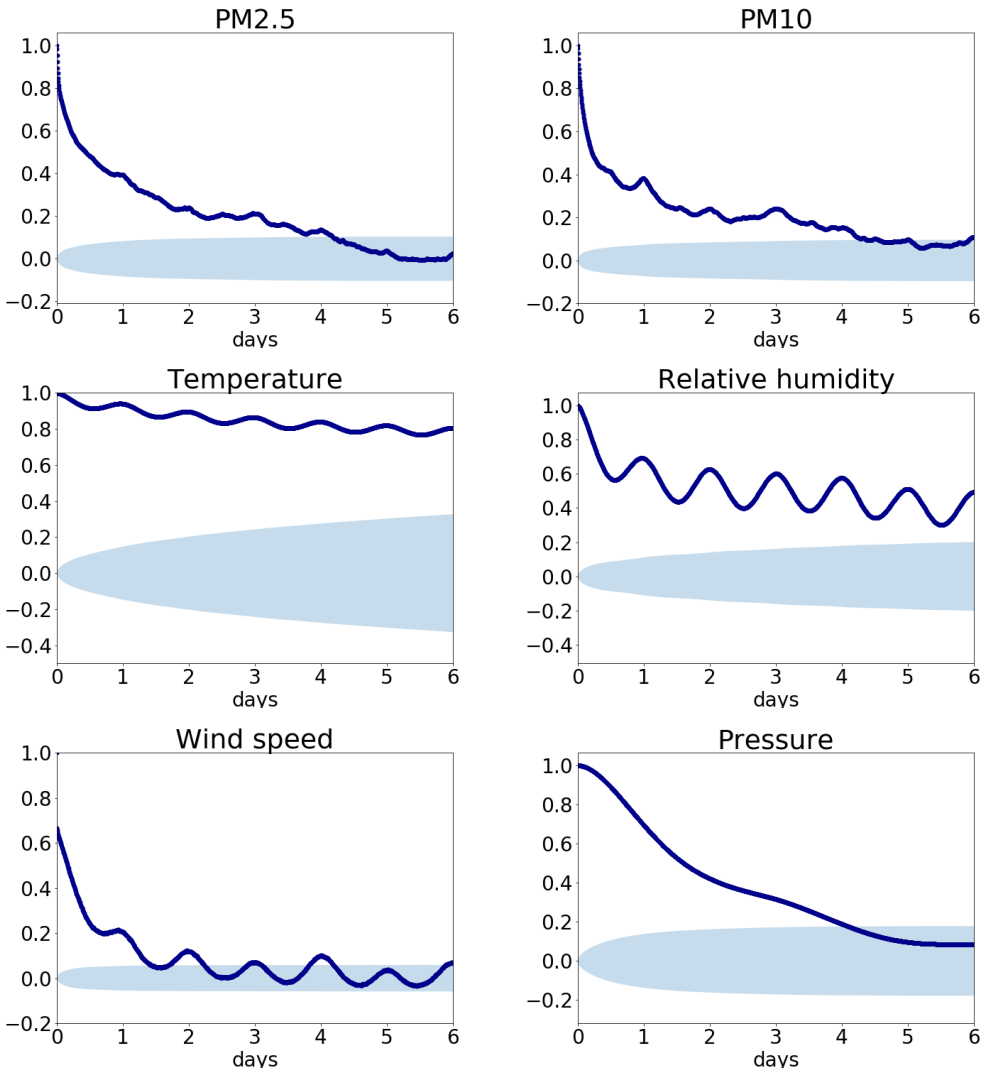


Figure 5: Autocorrelation of $PM_{2.5}$, PM_{10}, temperature, humidity, wind speed and air pressure in the reference data. The y-axis represents correlation, and the x-axis time lag.

It can be observed from the figures that the oscillation pattern, although slightly weaker than in the PM measurements, is similar for all images except for pressure. In other words, knowing the meteorological and pollutant situation at certain timestamp has a clear effect on calibrating consecutive values when the effect of autocorrelation persists. Using a single continuous period that has been split to train and test part gives an overly optimistic impression about the performance of a calibration technique, since the datasets are likely to be very similar in terms of environmental conditions. Therefore, to thoroughly evaluate calibration model's performance another way of selecting training and validation data is needed.

Selecting Diverse Data for Calibration Models

The use of diverse training data has shown to greatly boost the performance of machine learning models in the past [50]. To create a calibration model that generalizes well, good training data is needed, but in practice the available training data is often limited in time. It is possible that the data does not thoroughly depict the dependency between meteorological features and pollutants, since not all weather phenomena are equally likely to occur across the seasons. Rather than selecting arbitrary continuous period of time, where some meteorological phenomena might or might not occur, a novel method of selecting data that considers multiple environmental features to properly cover different pollution levels and weather conditions is introduced in this thesis. Temperature, relative humidity, air pressure, wind speed and air pollutants are selected as features to form various datasets in order to seek their impact in the calibration of a low-cost sensor.

Algorithms 1 and 2 select periods from available data, when the value of a feature is high or low. Using these algorithms the distribution of different meteorological phenomena or pollution levels can be balanced, and the training data covers as diverse conditions as possible. Algorithm 3 combines the two previously mentioned techniques by directly creating a collection that contains both high and low periods.

Algorithm 1: construct-high-dataset

input : $\mathbf{d}, w$, where $\mathbf{d} \ni d_1, \ldots, d_n$ and $w \in \mathbb{N}$

output: indexes

begin

indexes $= \emptyset$

while *True* **do**

$\mathrm{S} = \bigcup_{i=1}^{N} \sum_{n=1}^{w} d_{i+n} K(n)$ where $\mathrm{K} = \begin{cases} 1, & \text{when } n \leq w \\ 0, & \text{otherwise} \end{cases}$

$\mathrm{i} = \arg\max_i(S \setminus \{-\infty\})$

if $i = \emptyset$ **then**

break

end

indexes.push$(d_i, ..., d_{i+w})$

$d_i, \ldots, d_{i+w} = -\infty$

end

end

Algorithm 2: construct-low-dataset

input : $\mathbf{d}, w$, where $\mathbf{d} \ni d_1, \ldots, d_n$ and $w \in \mathbb{N}$

output: indexes

begin

indexes $= \emptyset$

while *True* **do**

$\mathrm{S} = \bigcup_{i=1}^{N} \sum_{n=1}^{w} d_{i+n} K(n)$ where $\mathrm{K} = \begin{cases} 1, & \text{when } n \leq w \\ 0, & \text{otherwise} \end{cases}$

$\mathrm{i} = \arg\min_i(S \setminus \{\infty\})$

if $i = \emptyset$ **then**

break

end

indexes.push$(d_i, ..., d_{i+w})$

$d_i, \ldots, d_{i+w} = \infty$

end

end

To select meaningful data for training and validation, Algorithms 1 and 2 were executed for target pollutant data resulting in high and low concentration datasets. After selecting only the first 10 windows to ensure diversity between the two sets, the remaining data was used to create training sets w.r.t. meteorological factors. The same algorithms were executed again this time individually for each of the meteorological factors. In addition, Algorithm 3 was used to create a diverse dataset from each of the meteorological factors.

Algorithm 3: construct-diverse-dataset

input : $\mathbf{d}, w$, where $\mathbf{d} \ni d_1, \ldots, d_n$ and $w \in \mathbb{N}$

output: indexes

begin

indexes $= \emptyset$

while *True* **do**

$\mathrm{S} = \bigcup_{i=1}^{N} \sum_{n=1}^{w} d_{i+n} K(n)$ where $\mathrm{K}= \begin{cases} 1, & \text{when } n \leq w \\ 0, & \text{otherwise} \end{cases}$

$\mathrm{i} = \arg\max_i(S \setminus \{-\infty, \infty\})$

$d_i, \ldots, d_{i+w} = -\infty$

$\mathrm{S} = \bigcup_{j=1}^{N} \sum_{n=1}^{w} d_{j+n} K(n)$ where $\mathrm{K}= \begin{cases} 1, & \text{when } n \leq w \\ 0, & \text{otherwise} \end{cases}$

$\mathrm{j} = \arg\min_j(S \setminus \{-\infty, \infty\})$

if $i = \emptyset$ *or* $j = \emptyset$ **then**

break

end

indexes.push$(d_i, ..., d_{i+w})$

indexes.push$(d_j, ..., d_{j+w})$

$d_j, \ldots, d_{j+w} = \infty$

end

end

4 Experiment Setup

The data used in this work was collected from beginning of March 2018 to end of December 2018 using two devices. To expose the sensors to same surrounding environmental conditions, a low-cost sensor was installed at similar altitude as close to the reference station as possible. The used location is within 50m from the reference station. As pollutant gas levels of our deployment site are constantly very low, the investigation of pollutants was limited to particulate matter (PM) that showed signs of variability in concentration levels. Also, previous studies have shown low-cost sensors to struggle more with PM than gases [7].

Sensor Deployment

The two devices used for collecting the data are shown in Figure 1. The reference device, Station for Measuring Ecosystem-Atmosphere Relations (SMEAR) (shown in top-left in Figure 1) provides accurate information about many atmospheric features, including both PM_{10} and $PM_{2.5}$. We consider the PM values from SMEAR as target values in the calibration. The other device is a low-cost micro sensor device built by a research group in the University of Helsinki (shown in Figure 1 bottom). It has sensors for temperature, relative humidity, NO_x, CO, CO_2, O_2 and PM. PM is measured by a Shinyei PPD42NS dust sensor[4], which has been lab tested [5] and also used in previous studies [10, 19]. The manufacturer states that it detects particles sizing down to 1µg/m^3.

As can be seen from Figure 1, these two ways of measuring air quality are clearly of different scale. By its size, the smear tower is approximately 50 meters tall, whereas the micro sensor is approximately the size of a small shoe box. The SMEAR sensor costs in excess of 1M euros, and the micro sensor costs around 250 euros.

Calibration Models

In Section 3, past studies of low-cost air quality sensor calibration were found to benefit from using information about meteorological conditions while calibrating air pollutants. For this reason a multi-feature approach is selected for this thesis. In addition to information about pollution level, calibration models will have temperature, relative humidity and wind speed as inputs.

[4]https://www.seeedstudio.com/Grove-Dust-Sensor-PPD42NS.html

A multiple linear regression will act as baseline calibration model to show that a linear model is not capable of capturing the relationship between the two PM sensors. Two other calibration techniques, a Random Forest and an Artificial neural network(ANN) are compared against the baseline. Two different versions of ANN will be compared in this thesis. The other one will be a baseline model consisting of input layer, one hidden layer and an output layer. A similar architecture has been used in low-cost air quality sensor calibration in the past [48]. The hidden layer will have the same number of neurons as the number of input features. The other ANN is the model used in [9]: two hidden layers with 512 neurons on each layer. Both ANN models use ReLU activation function in all layers.

Evaluation Procedure and Baseline

The method of selecting diverse data presented in this thesis (Section 3) is used to create multiple subsets from the available data each one corresponding to extreme conditions of pollution level (PM_{10} and $PM_{2.5}$), temperature, humidity, air pressure or wind speed. These subsets of extreme conditions, i.e. sets of high and low level, were created to highlight model generality in different training and testing environments. In addition, diverse subsets were created by levels of meteorological conditions to improve the robustness of calibration models. Calibration models are then evaluated in different environments and compared to a continuous evaluation used in previous studies [15, 43, 48] to prove that continuous training and testing results in optimistic impression about capabilities of a calibration model.

Data Cleaning

As explained in Section 2, pre-calibration procedures are required for the data before the low-cost sensor can be calibrated with the reference station. In this thesis negative PM values smaller than -1 are replaced with *missing* due to their irrationality (Figure 6). After that linear interpolation is selected to handle missing data as the missing gaps in the data are at most few hours (Figure 7). Also, the frequency of the reference measurements are matched with the frequency of the low-cost sensor, as it requires less interpretation and assumptions on the low-cost sensor's data[5].

[5]The micro sensor collects data approximately every 5 minutes, reference station collects data every minute.

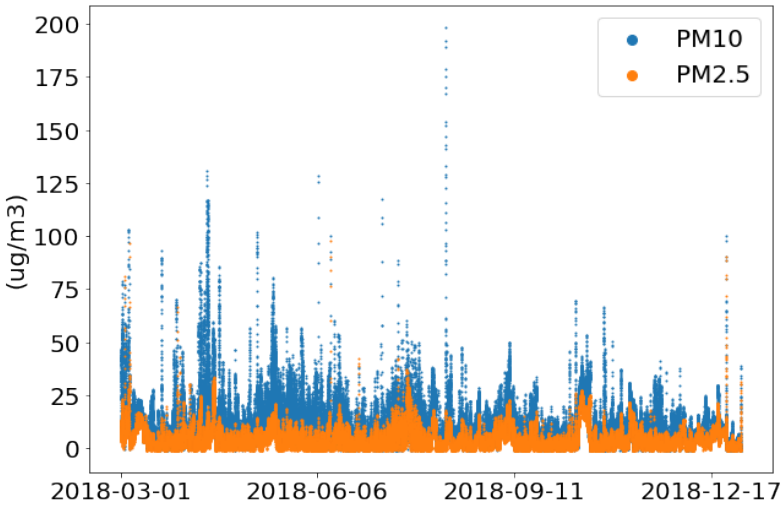


Figure 6: Reference measurements of $PM_{2.5}$ and PM_{10} from 2018. Large negative values have been removed from the data, but missing values still occur in the data.

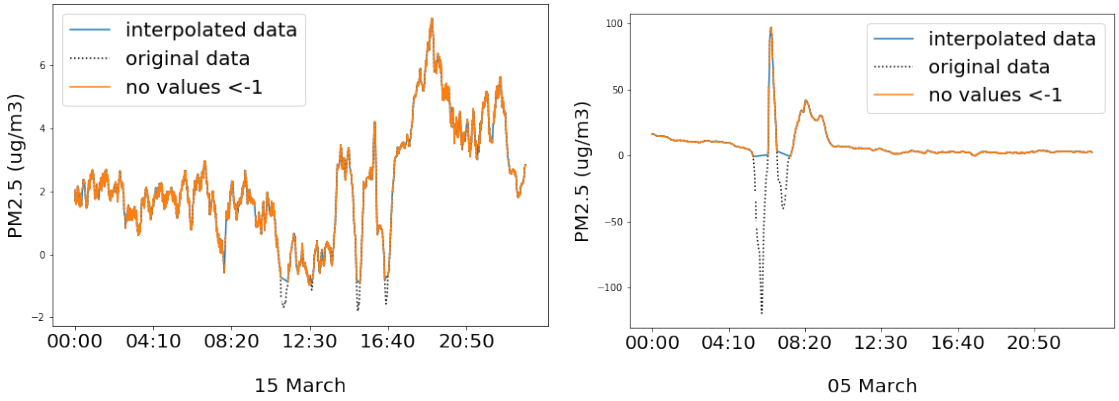


Figure 7: Examples of two types of interpolation scenarios in the SMEAR data. **Left:** moderate interpolation, where original negative values are greater than -2. **Right:** extreme interpolation case, where original negative values are close to -120.

Starting point accuracy

The low-cost sensor uses a different particle detection method from the method used in the reference station, therefore the range of values is different, as can be seen on the left in Figure 8. To visualize the correspondence of the values between the sensors, Min-Max scaling[6] is applied to both values separately. The scaled image is shown on the right in the same figure. The scaled version of the micro sensor's PM measurements shows that the correlation with the reference PM measurements is small or at least not obvious. Indeed the resemblance of the visualized PM data of the two sensors is close to none.

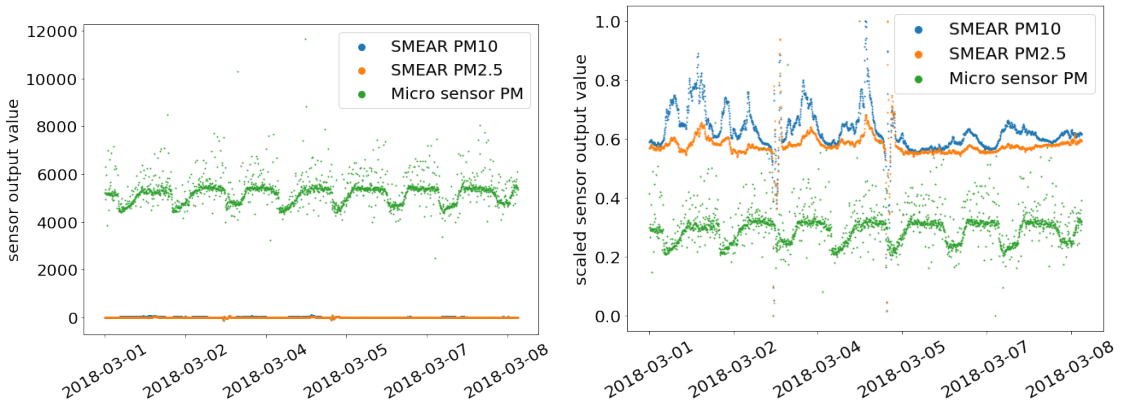


Figure 8: PM concentration is measured in pcs/l by the micro sensor, but in µg/m^3 by the reference station. **Left:** Original scale of values. **Right:** The min-max scaled version of the image on the left.

The starting point accuracy between the sensors can be investigated through correlation. If the sensors react similarly to some measured substance, the correlation between the two should be strong. Correlations between the micro sensor and the reference station are computed for all available clean data, and are shown in Table 4. As can be seen from the table temperature and relative humidity values correlate strongly between the two sensors, which indicates that the micro sensor reacts rather accurately to those and that the measurement locations are comparable. The PM detection sensor from the micro sensor on the other hand is not showing any sign of correlation with neither of the reference station's PM sensors. This indicates that before calibration the micro sensor's capability of detecting PM concentrations is poor.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Table 4: Correlations between the SMEAR (S) reference station and the micro sensor(M) (t=temperature, rh=relative humidity, p=pressure, ws=wind speed).

	t(S)	**rh(S)**	**p(S)**	**ws(S)**	**$PM_{2.5}$(S)**	**PM_{10}(S)**	**t(M)**	**rh(M)**	**PM(M)**
t(S)	1	-0,41	0,01	-0,04	0,03	0,12	0,94	-0,36	-0,01
rh(S)	-0,41	1	-0,19	-0,03	0,09	-0,25	-0,46	0,85	0,04
p(S)	0,01	-0,19	1	-0,18	0,12	0,23	0,01	-0,14	-0,01
ws(S)	-0,04	-0,03	-0,18	1	-0,2	-0,18	-0,04	0,05	-0,07
$PM_{2.5}$(S)	0,03	0,09	0,12	-0,2	1	0,71	0,03	0,04	0,04
PM_{10}(S)	0,12	-0,25	0,23	-0,18	0,71	1	0,14	-0,26	0,01
t(M)	0,94	-0,46	0,01	-0,04	0.03	0.14	1	-0.53	-0.04
rh(M)	-0.36	0.85	-0,14	0,05	0.04	-0.26	-0.53	1	0.03
PM(M)	-0.01	0.04	-0,01	-0,07	0.04	0.01	-0.04	0.03	1

5 Results

The calibration techniques explained in Section 2 and further discussed in Section 3 were applied to the micro sensor's measurements. The predictions produced by the calibration models were then compared against the reference values. The calibration techniques compared were multiple linear regression (MLR), Random Forest regressor (RF), a baseline Artificial Neural Network (ANN (BL)) and an Artificial Neural Network (ANN), as explained in Section 4.

As explained in Section 3, to ensure that low-cost sensor calibration techniques can be evaluated fairly the evaluation must be carried out in diverse conditions and the effect of autocorrelation must be take into consideration i.e. not use a single continuous period for training and evaluation. This is why we trained the calibration models in various environmental conditions and against different levels of pollution to consider the level of generality achieved with specific kind of training and testing data. The data selecting method proposed in this thesis (Section 3), was compared with a more traditional method of selecting data. All model trainings and evaluations were done for the two types of particulate matter ($PM_{2.5}$ and PM_{10}) separately.

Continuous Training and Evaluation Datasets

Evaluation of a calibration model in diverse conditions is compared to an evaluation method, where a model is trained using part of a continuous dataset and the remaining part is used for evaluation of the model as used in previous works [9, 10, 15, 43]. As explained in Section 3, such evaluation might give an overly optimistic impression

Table 5: The errors of continuous training and evaluation periods.

	$PM_{2.5}$		PM_{10}	
error metric	MAE	RMSE	MAE	RMSE
MLR	3.54	4.62	9.01	11.36
RF	3.73	5.05	9.31	12.45
ANN (BL)	4.55	6.59	7.56	10.32
ANN	3.77	5.21	8.44	11.46

about the performance of a calibration technique due to the strong autocorrelation of environmental features (Section 3, Figure 5). Results from such a way of training and evaluating the models are shown in Table 5. The continuous period covering 28,800 data points in 5 minute intervals was chosen arbitrarily, and was not tweaked after selecting it for the first time. The training and testing periods were formed by splitting data into two equal sized sets without shuffling. The datasets in the evaluations contained the same amount of data points (14,400) each, unless other is mentioned.

Impact of pollution concentration

We first consider the performance of calibration techniques when concentration levels of particulate matter (PM) differ greatly. A high concentration dataset was selected with a method depicted in Algorithm 1 (Section 3), and a low concentration dataset with Algorithm 2, both with a window of 5 days (1,440 data points). The outputs of the algorithms are datasets containing several continuous time periods corresponding to a certain selection criteria. As the algorithms work by sorting the maximum number of 5-day sequences, only the first 10 windows were selected to ensure that the high and low sets remained as different as possible and it was confirmed that no overlap occurred between the sets.

To verify the difference between the sets, a Kolmogorov-Smirnov (KS) two sample test was applied for the high and low set. The KS test evaluates whether two underlying one-dimensional probability distributions differ i.e. whether the two datasets come from the same distribution. The difference between the distributions is confirmed by rejecting a null hypothesis, when a p-value falls below a certain threshold. The computed p-value was close to zero verifying the difference of the datasets. The mean and standard deviation of both sets were computed to explore

Table 6: The errors between true concentrations and predictions from models trained with data selected considering the level of PM. Values from Table 5 have been added to the table's rightmost column for comparison purposes.

calibration target – $PM_{2.5}$

train:	low		high		high		low		continuous	
test:	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	**1.38**	1.98	4.95	6.78	**9.28**	9.58	**8.58**	10.62	**3.54**	4.62
RF	**1.5**	2.16	6.31	8.36	**11.68**	12.5	**8.62**	10.71	**3.73**	5.05
ANN (BL)	**1.39**	1.97	5.0	6.83	**10.23**	10.55	**10.21**	12.02	**4.55**	6.59
ANN	**1.4**	2.01	4.7	6.35	**7.59**	8.32	**8.43**	10.51	**3.77**	5.21
Average	**1.42**	2.03	5.24	7.08	**9.69**	10.24	**8.96**	10.97	**3.90**	5.37

calibration target – PM_{10}

train:	low		high		high		low		continuous	
test:	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	**2.76**	3.63	11.19	15.29	**15.36**	16.32	**17.08**	22.42	**9.01**	11.36
RF	**3.31**	4.36	16.29	21.85	**12.01**	14.58	**17.64**	22.91	**9.31**	12.45
ANN (BL)	**2.81**	3.67	12.83	18.41	**15.8**	16.73	**17.08**	22.48	**7.56**	10.32
ANN	**2.99**	4.03	12.32	17.23	5.58	7.57	**16.46**	21.85	**8.44**	11.46
Average	**2.97**	3.92	13.16	18.20	12.19	13.8	**17.07**	22.42	**8.58**	11.40

them more carefully. They are shown in Table 8.

The calibration models were evaluated by training with either high or low concentration data and evaluating with the remaining one. To give an impression about the performance of the model on similar data, a leave-one-out cross-validation was performed by leaving one of the continuous periods out from the training data and then using that left out period for validation. The validation was done multiple times by leaving each of the continuous periods out once. The results in Table 6 verify the expected. The cross validation results prove that too homogeneous data gives optimistic impression about the performance of the model. Using ANN (BL) to calibrate $PM_{2.5}$ a homogeneous training and evaluation environment resulted in 85% lower mean absolute error (MAE) than using distinct training and evaluation environment. In PM_{10} the biggest difference occurred using MLR. Even 83% lower MAE was achieved when using similar training and evaluation conditions.

The continuous evaluation results from Table 5 are shown in the rightmost columns for comparison. Results from the continuous evaluation are closer to the results of

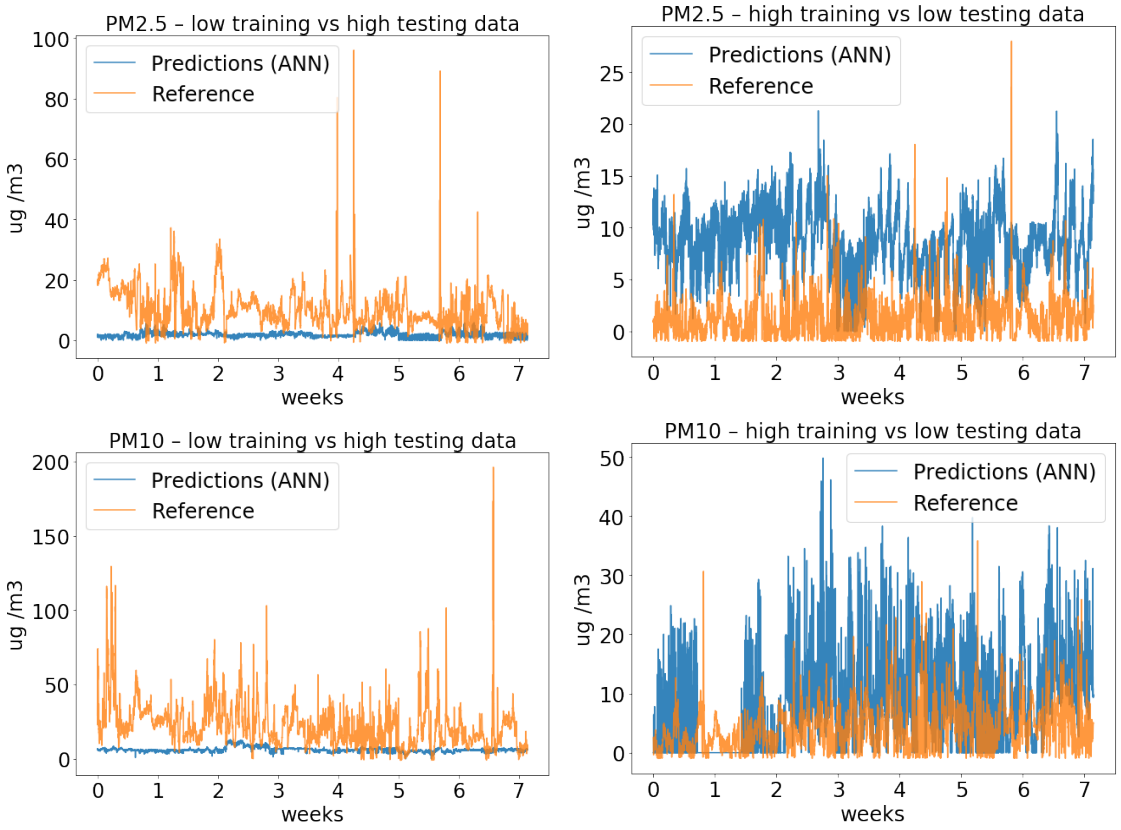


Figure 9: Predictions after training and evaluating with data corresponding to different pollution levels. The plots depict the results shown in Table 6. Only the ANN's predictions are shown to keep the plots interpretable, but other techniques behaved similarly.

the cross validation results, which suggests that the continuous evaluation method does not cover diverse PM concentrations. The models trained with high concentration data tend to overestimate the concentrations when evaluated with low data, whereas the models trained with low concentrations seem to underestimate them when evaluated with high concentration data. This is confirmed by the plots of the predictions in Figure 9. One exception in the experiment is the ANN's performance when trained with high concentration PM_{10} data and evaluated with low concentration data. The predictions are higher, but still quite reasonable, and the error is the smallest among all techniques.

The average error of all techniques in different training and evaluation environments is shown at the bottom. For $PM_{2.5}$, the magnitude of the average errors in homogeneous and continuous evaluation environments are from low to very low in terms

of air pollution concentration. However diverse evaluation yields an error that is almost the size of the annual guideline value set by WHO (shown in Table 1, Section 1). Similar observations can be made for PM_{10}. This highlights the significance of generalizability of a calibration model.

It should be emphasized that, besides few peak values in Figure 9, even the highest concentrations in $PM_{2.5}$ high concentration dataset are rather low as they yield an air quality index of *Satisfactory* to *Fair* on Finnish scale, or *Excellent* on Chinese scale (Table 2, Section 1). The PM_{10} levels in the high dataset on the other hand indicate that higher concentration periods might occur occasionally in the test environment instead of just witnessing individual peak values. The Finnish index ranks the highest levels of the high PM_{10} dataset to *Poor*, the Chinese index to *Good*.

Impact of Meteorological Factors

The next evaluation is similar to the previous one, but emphasizing selecting periods based on different meteorological factors. After selecting the two datasets based on the level of PM concentration, low and high training datasets were selected from the remaining data with the same Algorithms 1 and 2 by looking at different meteorological features with a window of 5 days. For some training sets, the length of the window had to be shortened since there were not enough non-overlapping 5-day periods. The meteorological features were selected to cover temperature, relative humidity, air pressure and wind speed.

The training sets contained roughly the same amount of data as the high and low PM sets. However, for those meteorological variables with shorter window, a slight decrease in data resulted. The smaller sets contained 13,824 data points (48 days) whereas the rest had 14,400 data points (50 days). Note that for $PM_{2.5}$ and PM_{10}, the meteorological training sets are not guaranteed to be the same, since they are selected from the remaining data after forming the high and low sets w.r.t. each PM. Sizes of all datasets are shown in Table 7.

All high and low partitions in Tables 9-12 refer to meteorological conditions. The meteorological high and low datasets are depicted in Table 8 through mean and standard deviation values of each set. Also mean and standard deviation of PM concentration of each set is shown on the right. Although the means of all features vary greatly, the $PM_{2.5}$ concentration stays rather stable through low and high datasets. KS test was used to confirm that all high and low datasets indeed were

Table 7: Number of data points in each data partition. All train, test, low, high and diverse sets were equal by their size w.r.t. partitioning feature.

partition	**target**		
based on	**$PM_{2.5}$**	**PM_{10}**	**partitions**
continuous data	14,400	14,400	train, test
$PM_{2.5}$	14,400	-	low, high
PM_{10}	-	14,400	low, high
temperature	14,400	14,400	low, high, diverse
humidity	14,400	14,400	low, high, diverse
pressure	13,825	13,825	low, high, diverse
wind speed	13,825	14,400	low, high, diverse

different from each other w.r.t. both the feature used for partitioning and the PM. In PM_{10} the differences are more noticeable, but standard deviation remains high. This might be caused by constantly varying PM_{10} concentration levels or by occasional high periods in a generally low pollution level environment.

The models were trained with either the high or low meteorological concentration data and tested with the remaining one. A cross validation was also performed within low-low and high-high sets as in the previous experiment. The results of the experiment are shown in Tables 9-12. The continuous evaluation results from Table 5 are shown in the rightmost columns for comparison purposes.

When selecting training and evaluation data w.r.t. to temperature, the calibration of $PM_{2.5}$ only shows mild effects to different levels of temperature. This indicates that changes in temperature alone do not cause major changes in the functionality of the calibration models as all the errors are close to those of the continuous period.

When calibrating the PM_{10} measurements, all models performed better in low temperature than in continuous period. This again implies overfitting since training and testing in high temperature environment increases MAE by 17% – 92%.

RMSE was chosen as a secondary error metric to depict any odd behaviour of the models. Both MLR and RF achieve lower MAE than the continuous evaluation when trained with low temperature data and evaluated with high temperature data. The RMSE of the same cases are higher to those of the continuous periods. This indicates that the mistakes in the model predictions are relatively large which implies failing to predict a big increase or a decrease in the PM concentration. Almost identical observations can be made from the datasets selected w.r.t pressure (Table 11).

Table 8: The mean values of each dataset selected w.r.t. certain feature.

low datasets

	$PM_{2.5}$				PM_{10}			
feature	$mean_{feat}$	SD_{feat}	$mean_{PM}$	SD_{PM}	$mean_{feat}$	SD_{feat}	$mean_{PM}$	SD_{PM}
PM (µg/m^3)	1.44	2.04	1.44	2.04	4.65	3.76	4.65	3.76
temperature (°C)	-1.36	3.08	3.35	3.05	-1.77	4.32	10.01	7.22
humidity (%)	54.21	16.35	3.91	4.0	56.07	16.14	11.84	9.49
pressure (mbar)	996.58	7.15	3.36	3.51	997.72	7.07	9.23	7.25
wind speed (m/s)	2.26	1.27	3.87	3.11	1.95	1.05	11.36	8.33

high datasets

	$PM_{2.5}$				PM_{10}			
feature	$mean_{feat}$	SD_{feat}	$mean_{PM}$	SD_{PM}	$mean_{feat}$	SD_{feat}	$mean_{PM}$	SD_{PM}
PM (µg/m^3)	8.86	5.28	8.86	5.28	19.17	13.28	19.17	13.28
temperature (°C)	18.82	3.87	4.23	3.39	19.39	3.94	11.96	8.04
humidity (%)	87.91	8.49	3.79	3.19	88.57	9.28	9.04	6.0
pressure (mbar)	1020.4	5.84	3.91	3.18	1019.88	5.8	10.53	7.54
wind speed (m/s)	4.76	2.34	3.08	2.85	5.08	2.24	8.34	6.28

Table 9: Errors between true concentrations and predictions from models trained with data selected considering the level of temperature.

data selected based on temperature, calibration target – $PM_{2.5}$

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	3.41	5.1	3.67	4.65	3.47	5.4	4.86	6.35	3.54	4.62
RF	3.77	5.54	3.98	5.1	3.3	5.38	3.97	5.15	3.73	5.05
ANN (BL)	3.5	5.24	3.53	4.54	3.5	5.09	3.59	4.81	4.55	6.59
ANN	3.61	5.4	3.79	4.96	5.02	7.06	4.15	5.44	3.77	5.21

data selected based on temperature, calibration target – PM_{10}

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	6.19	9.01	8.55	11.95	6.08	9.18	**8.79**	**12.97**	**9.01**	**11.36**
RF	6.75	10.54	9.13	13.06	6.51	9.95	**8.85**	**12.7**	**9.31**	**12.45**
ANN (BL)	6.94	10.05	8.19	11.74	9.06	11.47	8.35	12.84	7.56	10.32
ANN	**5.4**	8.63	**10.35**	14.07	10.77	14.78	8.83	12.65	**8.44**	11.46

Table 10: Errors between true concentrations and predictions from models trained with data selected considering the level of humidity.

data selected based on humidity, calibration target – $PM_{2.5}$

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	3.03	4.11	5.01	6.47	**12.05**	13.23	4.48	6.34	**3.54**	4.62
RF	3.09	4.37	4.78	6.59	6.48	8.07	5.91	7.71	3.73	5.05
ANN (BL)	4.67	6.23	5.97	8.31	3.23	4.31	4.25	6.18	4.55	6.59
ANN	2.87	4.05	5.71	7.51	4.54	5.51	4.25	5.95	3.77	5.21

data selected based on humidity, calibration target – PM_{10}

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	9.64	14.75	6.48	8.63	**15.85**	19.64	7.97	9.96	**9.01**	11.36
RF	10.29	16.38	**6.46**	9.08	12.4	17.51	**31.14**	37.79	**9.31**	12.45
ANN (BL)	**16.27**	21.56	5.66	7.76	9.04	15.05	9.25	12.57	**7.56**	10.32
ANN	8.65	13.64	8.12	10.43	10.42	15.44	10.85	13.48	8.44	11.46

Table 8 shows that mean and standard deviation of $PM_{2.5}$ values in the humidity based high and low datasets vary little. As the two are similar, no major difference in performance can be observed for $PM_{2.5}$ in Table 10 except with the MLR when trained with high humidity data and evaluated with the low humidity dataset.

Looking at Table 8, the mean and standard deviation values of PM_{10} show more differences between the high and the low humidity dataset. The PM_{10} values in Table 10 show varying performance of the models between datasets except the ANN for which performance is reasonably stable.

Training and evaluating with low and high periods of wind speed shows signs of the known negative correlation between PM and wind speed [52]. For example when calibrating PM_{10}, ANN trained in high and evaluated in low wind speed had over 50% higher error compared to the continuous evaluation, and in contrast even 35% lower error when trained in low and evaluated in high wind speed. This kind of model behaviour emphasizes the meaningfulness of verifying the performance under different meteorological conditions.

Table 11: Errors between true concentrations and predictions from models trained with data selected considering the level of pressure.

data selected based on pressure, calibration target – $PM_{2.5}$

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	2.71	3.56	3.76	5.28	3.63	4.19	3.05	4.87	3.54	4.62
RF	2.99	4.0	3.56	5.32	4.4	6.48	3.55	5.31	3.73	5.05
ANN (BL)	3.59	4.87	3.57	5.17	3.59	4.87	3.14	4.92	4.55	6.59
ANN	3.56	4.77	3.79	4.89	6.02	8.27	4.73	6.14	3.77	5.21

data selected based on pressure, calibration target – PM_{10}

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	4.44	6.39	**8.0**	**12.39**	7.16	8.6	**7.94**	**13.44**	**9.01**	**11.36**
RF	4.94	6.93	**8.84**	**14.09**	7.64	10.59	**8.46**	**13.71**	**9.31**	**12.45**
ANN (BL)	7.38	9.65	**7.48**	**12.31**	7.38	9.65	8.53	13.99	7.56	10.32
ANN	4.75	6.7	7.24	11.5	9.58	12.6	**8.07**	**13.37**	**8.44**	**11.46**

Table 12: Errors between true concentrations and predictions from models trained with data selected considering the level of wind speed.

data selected based on wind speed, calibration target – $PM_{2.5}$

train: wind speed	low		high		high		low		continuous	
test: wind speed	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	4.3	5.83	2.36	3.16	4.21	6.3	2.69	3.44	3.54	4.62
RF	4.64	6.44	2.8	3.77	4.24	6.3	6.29	7.18	3.73	5.05
ANN (BL)	6.27	8.22	2.36	3.22	4.4	6.51	3.89	4.41	4.55	6.59
ANN	5.48	7.17	2.44	3.36	5.35	7.42	5.87	7.18	3.77	5.21

data selected based on wind speed, calibration target – PM_{10}

train partition	low		high		high		low		continuous	
test partition	low		high		low		high		continuous	
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	**12.59**	17.87	**4.1**	5.34	11.08	17.8	6.18	7.57	**9.01**	11.36
RF	**15.22**	22.21	**4.46**	5.76	11.31	17.95	7.4	10.47	**9.31**	12.45
ANN (BL)	**17.92**	23.85	**4.0**	5.29	11.22	18.07	6.0	7.77	**7.56**	10.32
ANN	**16.08**	24.43	**4.45**	6.16	12.9	19.6	5.48	7.81	**8.44**	11.46

Impact of Meteorological Factors at Different Levels of PM

The previous experiment considered the effects of meteorological factors in calibration, which are correlated with pollution concentration (Table 4). Testing a calibration model in high wind speeds means testing it in lower pollution concentration than what occurs in low wind speeds (Table 8). Therefore, this section focuses directly on performance in different levels of pollution concentrations. The calibration models trained with data selected based on meteorological conditions were next evaluated with the PM high and low sets that were used earlier in the examination. In addition to the low and high meteorological datasets, a diverse training set was created for each feature using Algorithm 3 (Section 3). The motivation of using diverse datasets was justified in Section 3 where it was noticed to have improved the generalizability and boosted the performance of calibration models. The diverse dataset contained partly the same data as the low and high sets and it was the same size as them as shown in Table 7. A Kolmogorov-Smirnov test confirmed that all the diverse sets were different from the high and low sets.

Summary of Results

The robustness of different calibration models was compared by fixing the test feature to either PM type, while the training feature was iterated over temperature, pressure, humidity and wind speed. The detailed results of mean absolute errors (MAE) and root mean squares errors (RMSE), when training the models with different low, diverse and high sets can be found from Appendix 1. A summary of the tables is shown in Tables 13 and 14. Table 13 shows average performance of calibration models whereas Table 14 highlights the effects of using certain kind of training data.

In case of $PM_{2.5}$, looking at Table 13 the lowest mean of errors is achieved by RF both in terms of MAE and RMSE and it has lowest MAE and RMSE 38% of the time. MLR is close to RF in performance, but it never has the lowest error in either error metric, however sometimes it has the highest error (MAE 25%, RMSE 13% of the time). ANN (BL) has the highest mean MAE, being 11% higher than the MAE of RF. However, the behaviour of the model seems unstable since it yields the lowest error 25% of the time in MAE and 50% of the time in RMSE, but it also has the highest error 50% of the time in both MAE and RMSE. The unstable behaviour highlights both the models poor generality and the meaning of good quality training

Table 13: Summary of calibration techniques from Tables 15-18

$PM_{2.5}$						
calibration model	mean		#lowest		#highest	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	9.46	12.04	0	0	2	1
RF	9.21	11.96	3	3	1	2
ANN (BL)	10.3	12.86	2	4	4	4
ANN	10.19	13.14	3	1	1	1

PM_{10}						
calibration model	mean		#lowest		#highest	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	17.79	24.37	2	3	1	1
RF	18.24	25.21	1	1	2	2
ANN (BL)	20.85	27.26	2	1	4	3
ANN	19.65	26.4	3	3	1	2

data. ANN has slightly lower MAE and slightly higher RMSE than ANN (BL), but its behaviour is more stable: lowest MAE 38% of the time and the highest MAE or RMSE only 13% of the time.

Using the same summary(Table 13) for PM_{10}, MLR has the lowest mean MAE and RMSE. It has lowest MAE 25% and RMSE 38% of the time, and highest MAE or RMSE in 13% of the cases. RF has the second lowest mean MAE and RMSE, but it has lowest MAE and RMSE only in 13% and highest in 25% of the experiment cases. ANN (BL) shows signs of similar behaviour than with $PM_{2.5}$, having lowest error in 25% (MAE) and 13% (RMSE), and highest error in 50% (MAE) and 38% (RMSE) of the cases. ANN (BL) also has the highest mean MAE and RMSE values, the MAE being 17% and RMSE 12% higher than the lowest ones from MLR. The other ANN has lowest error in 38% of the cases in both error metrics and the highest in 13% (MAE) and 25% (RMSE) of the experiment cases.

Generally the levels of $PM_{2.5}$ are quite low, due to which the RF may capture most efficiently the small changes through non-linear decision chains. In PM_{10} the slightly higher concentrations might be easier to spot and the best fit is found using MLR or ANN. Comparing over both PM and all the techniques, ANN achieves the lowest MAE most often.

Table 14: Summary of training data partitions from Tables 15-18

$PM_{2.5}$						
data partition	mean		#lowest		#highest	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
low	9.93	12.65	5	4	4	5
diverse	9.5	12.22	3	2	0	1
high	10.01	12.71	0	2	4	4

PM_{10}						
data partition	mean		#lowest		#highest	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
low	18.75	25.57	3	4	2	4
diverse	17.97	24.87	1	2	1	1
high	18.48	25.18	4	2	5	3

In addition to comparing calibration techniques, the performance of training datasets selected using certain partitioning method (Algorithms 1-3, Section 3) were compared. Similar summary to the comparison of calibration techniques is shown in Table 14. Both in $PM_{2.5}$ and PM_{10} using low and high data partition resulted in a higher MAE and RMSE than when using a diverse partition. Using diverse training data in the calibration of $PM_{2.5}$ yielded approximately 5% lower MAE and 4% lower RMSE. In PM_{10} the errors were approximately 3% (MAE) and 2% (RMSE) lower. The use of diverse training data caused the highest error in case of PM_{10}(MAE and RMSE) only in 13% of the cases. In $PM_{2.5}$, diverse training data never caused the highest MAE and the highest RMSE was caused 13% of the time. The lowest errors in $PM_{2.5}$ were achieved in 38% (MAE) and 25% (RMSE) of the time. In PM_{10} the lowest errors were achieved respectively in 13 % and 25% of the cases.

6 Discussion and Summary

This Thesis has presented a novel method of selecting data (Section 3) to improve both training and evaluation process of low-cost sensor calibration models. Datasets created with this method were shown to give useful information about the generality of calibration models. This section goes through possible extensions and implications of our work, as well as covers its limitations.

Evaluation of Calibration Models

The calibration of low-cost sensor has been studied extensively [10, 15, 23, 43, 48]. A validation process based only on the performance of calibration model over a continuous period of real-life deployment [15, 43, 48] can give an overly optimistic impression, due to similar levels of air pollution, and autocorrelation of both meteorological features and air pollutants. Although the results of past studies are promising, little can be said about the generalizability of such calibration models. To ensure diversity of training and testing data, and to build confidence on the model's capability to handle varying real-life conditions, more emphasis should be put on data selection both in training and validation phase. Throughout the results section the magnitude of errors was found to vary for all calibration models depending on the used training and validation data. This implies that the level of generalization achieved by current state-of-the-art calibration models is not as good as might have been indicated, when evaluation has been based on a single continuous period.

This work has highlighted the importance of generalizability of calibration techniques and proposed a novel method of selecting periods of time based on high and low levels of environmental factors. Through this method of selecting data, calibration models can be guaranteed to have seen different phenomena in the training data. The results presented in Section 5 show that the validation process is easily biased by the homogeneity of the data. This is why model predictions on an arbitrary continuous period do not suffice in depicting the true performance of a calibration model in varying pollution concentrations or atmospheric conditions. The data selection method proposed in this thesis has shown to be useful in the evaluation of a calibration model, since it can be used to determine time periods of diverse environmental conditions. Model's performance can be evaluated in diverse pollution levels and meteorological conditions giving more realistic impression about its capability

to generalize over varying environmental conditions, also required by the regulatory air quality standards [2].

Impact of Results

Results in Table 14 support the suggested improvement of performance and robustness of calibration models, when using diverse training data [25]. Also our multi phased evaluation process has shown how sensitive calibration models are to extreme pollutant levels or meteorological conditions. Compared to a continuous evaluation period, this approach brings out the challenges that calibration models face in real-life deployments. After acknowledging weaker performance in certain environmental conditions, calibration models can be improved by including more diverse training data.

Results obtained in this Thesis indicate that some performance levels reported in previous studies might unintentionally give overly optimistic impression about the current level of state-of-the-art calibration models. Such findings are still valuable in developing new approaches, or as they are if they can be verified to maintain the level of performance under diverse environmental conditions.

The results also question the reliability of current applications that use calibrated low-cost air quality sensors. For example, Cheng et al. [10] introduce several applications utilizing low-cost air quality sensor data. These applications are used to show current level of air quality based in a users location, plan a trip with clean air instead of shortest distance or keep your virtual pet away from high pollution levels [42, p.749]. The models that provide calibrated measurements for these applications need to be re-evaluated in diverse environmental conditions to verify their general accuracy. Otherwise the calibrated accuracy might be even 83 lower when the model is exposed to diverse pollution levels, as shown in Table 6. Since the measurements cannot be confirmed to be reliable in varying environmental conditions, the applications built on calibrated measurements are less useful.

Suggested Future Work

Past studies on calibration could benefit from the observations done in this thesis by investigating the performance of their models under diverse environmental conditions. This would strengthen the obtained results from the perspective of generality,

and help to improve the robustness of their models by using diverse training data.

Examining data selection process more carefully in the future would likely benefit from having access to more significant pollution concentration levels. However, an environment with low pollution concentration is a challenging scenario to model since pollution level changes are subtle which low-cost sensors might struggle detecting. Therefore, it should also be paid attention to as a model should be capable of handling all kinds of pollution levels so that it could be considered truly trustworthy.

Comparing calibration techniques in diverse meteorological and pollution conditions (Tables 15-18, Appendix) showed that lowest error was achieved most often using an ANN. The potential of ANN in calibration however can be developed by using connections that are more complex than fully connected layers e.g., deep learning calibration models have shown to improve accuracy in the past [33]. For example Convolutional Neural Networks and Recurrent Neural Networks can be used to look at more than just a single timestamp. Including information from previous timestamps might enable the model to better learn the temporal effect of certain phenomena without directly copying the autocorrelation pattern.

Instead of attempting to estimate air quality measurements directly, another option is to modify the problem statement into predicting the general level of air quality w.r.t. an air quality index. In such scenario the task is to detect changes in the air quality index instead of exact concentrations. The data could be assigned an indication of the level of air quality at each timestamp as has been done in the past [10, 19]. In terms of machine learning, this turns the regression problem into a classification problem, which enables the use of other techniques. This way the problem could benefit from applying both regression and classification techniques to solve the mapping from inputs to outputs. Regardless whether the problem type is classification or regression, validation should still be conducted more rigorously, as proposed in this work.

Limitations

A limiting factor in this work is the used distance between the reference station and the micro sensor. The distance was less than 50 meters and there were no obstacles between the two, however an optimal positioning of the sensors would be next to each other. However, installing the micro sensor to rooftop ensured similar altitude with the reference station, which would have otherwise been challenging.

Another limitation is the general concentration of particles in the available data. The general PM concentration tends to be rather low making most of the PM data quite similar. The effects of different environmental phenomena might not be as clearly observable. To fully confirm the findings, a similar work should be done with data from an area which has generally higher level of PM concentration.

The effects of using a shorter time window when creating datasets (Table 8) were not fully covered. This should be done in the future to see whether using a shorter time window better balances the distribution of extreme phenomena appearing in training data, and therefore improves generality of a calibration model.

Summary

This thesis studied the calibration of low-cost air quality sensors to improve their accuracy in real-life deployments where environmental conditions constantly vary. As the accuracy of the low-cost sensors is affected by both the air pollution level and meteorological factors, the evaluation of a calibration technique was carried out more thoroughly than what has been done in the past where a single continuous period has been selected. Evaluation on a continuous period does not suffice since a calibration technique is not guaranteed to generalize over diverse environmental conditions, and since daily patterns in data affect model predictions. PM concentrations were found to affect the calibration by giving too optimistic impression about the performance of the model when trained and evaluated in similar conditions. Out of the meteorological factors, high wind speed had the most significant effect, lowering the error up to 78% compared to training with low wind speed periods, but humidity and pressure also affected calibration performance. In summary it is recommended that at least one year of data covering all seasons is used to have the right kind of training data for a calibration model. Results with all available data suggest that selecting diverse training data achieves a better level of performance and robustness in varying environmental conditions and deployment sites.

References

1 BIPM , IEC , IFCC , ILAC , ISO , IUPAC , IUPAP , and OIML . *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*. Joint Committee for Guides in Metrology, JCGM 200, 01 2012.

2 European Environment Agency. Guide to the demonstration of equivalence of ambient air monitoring methods. Publication, European Environment Agency, 2010.

3 European Environment Agency. Air quality in europe - 2018 report. Publication, European Environment Agency, 2018.

4 Zorana J. Andersen, Luise C. Kristiansen, Klaus K. Andersen, Tom S. Olsen, Martin Hvidberg, Steen S. Jensen, Matthias Ketzel, Steffen Loft, Mette Sørensen, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Stroke and long-term exposure to outdoor air pollution from nitrogen dioxide. *Stroke*, 43(2):320–325, 2012.

5 Elena Austin, Igor Novosselov, Edmund Seto, and Michael G. Yost. Laboratory Evaluation of the Shinyei PPD42NS Low-Cost Particulate Matter Sensor. *PLOS ONE*, 10(9):1–17, 09 2015.

6 Finlex Data Bank. Valtioneuvoston asetus ilmanlaadusta, 2017.

7 C Borrego, AM Costa, J Ginja, M Amorim, M Coutinho, K Karatzas, Th Sioumis, N Katsifarakis, K Konstantinidis, S De Vito, et al. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise. *Atmospheric Environment*, 147:246–263, 2016.

8 C. Borrego, J. Ginja, M. Coutinho, C. Ribeiro, K. Karatzas, Th Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, M. Salvato, P. Smith, N. André, P. Gérard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguillón, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise - Part II. *Atmospheric Environment*, 193:127 – 142, 2018.

9 C. Chen, C. Kuo, S. Chen, C. Lin, J. Chue, Y. Hsieh, C. Cheng, C. Wu, and C. Huang. Calibration of low-cost particle sensors by using machine-learning

method. In *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 111–114, Oct 2018.

10 Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. AirCloud: A cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, SenSys '14, pages 251–265, New York, NY, USA, 2014. ACM.

11 José María Cordero, Rafael Borge, and Adolfo Narros. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensors and Actuators B: Chemical*, 267:245 – 254, 2018.

12 E. S. Cross, L. R. Williams, D. K. Lewis, G. R. Magoon, T. B. Onasch, M. L. Kaminsky, D. R. Worsnop, and J. T. Jayne. Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements. *Atmospheric Measurement Techniques*, 10(9):3575–3588, 2017.

13 Ben Croxford, Alan Penn, and Bill Hillier. Spatial distribution of urban pollution: civilizing urban traffic. *Science of The Total Environment*, 189-190:3 – 9, 1996. Highway and Urban Pollution.

14 Evi Dons, Luc Int Panis, Martine Van Poppel, Jan Theunis, Hanny Willems, Rudi Torfs, and Geert Wets. Impact of time – activity patterns on personal exposure to black carbon. *Atmospheric Environment*, 45(21):3594 – 3602, 2011.

15 E. Esposito, S. De Vito, M. Salvato, V. Bright, R.L. Jones, and O. Popoola. Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. *Sensors and Actuators B: Chemical*, 231:701 – 713, 2016.

16 Department for Environment Food & Rural Affairs. A guide for local authorities purchasing air quality monitoring equipment, 3 2006.

17 Fanyu Gao. Evaluation of the Chinese new air quality index (GB3095-2012): based on comparison with the US AQI system and the WHO AQGs. 2013.

18 Meiling Gao, Junji Cao, and Edmund Seto. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China. *Environmental Pollution*, 199:56 – 65, 2015.

19 Y. Gao, W. Dong, K. Guo, X. Liu, Y. Chen, X. Liu, J. Bu, and C. Chen. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, April 2016.

20 Ulrike Gehring, Alet H. Wijga, Michael Brauer, Paul Fischer, Johan C. de Jongste, Marjan Kerkhof, Marieke Oldenwening, Henriette A. Smit, and Bert Brunekreef. Traffic-related air pollution and the development of asthma and allergies during the first 8 years of life. *American Journal of Respiratory and Critical Care Medicine*, 181(6):596–603, 2010. PMID: 19965811.

21 Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

22 Nils Y. Hammerla and Thomas Plötz. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1041–1051, New York, NY, USA, 2015. ACM.

23 David Hasenfratz, Olga Saukh, Silvan Sturzenegger, and Lothar Thiele. Participatory air pollution monitoring using smartphones. In *Proceedings of the 2nd International Workshop on Mobile Sensing*, pages 1–5, Beijing, China, 2012. Academic Press.

24 Volk HE, Lurmann F, Penfold B, Hertz-Picciotto I, and McConnell R. Traffic-related air pollution, particulate matter, and autism. *JAMA Psychiatry*, 70(1):71–77, 2013.

25 Samuli Hemminki, Petteri Nurmi, and Sasu Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, SenSys '13, pages 13:1–13:14, New York, NY, USA, 2013. ACM.

26 Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

27 Finnish Meteorological Institution. Ilmanlaadun seurantojen laatu ja kustannukset, 2015.

28 Finnish Meteorological Institution. Air quality index, 2019.

29 Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

30 Prashant Kumar, Lidia Morawska, Claudio Martani, George Biskos, Marina Neophytou, Silvana Di Sabatino, Margaret Bell, Leslie Norford, and Rex Britter. The rise of low-cost sensing for managing air pollution in cities. *Environment International*, 75:199 – 205, 2015.

31 Alastair Lewis and Peter Edwards. Validate personal air-pollution sensors. *Nature*, 535, 2016.

32 Ian J Litchfield, Jon G Ayres, Jouni J K Jaakkola, and Nuredin I Mohammed. Is ambient air pollution associated with onset of sudden infant death syndrome: a case-crossover study in the uk. *BMJ Open*, 8(4), 2018.

33 Titti Malmivirta, Jonatan Hamberg, Eemil Lagerspetz, Xin Li, Ella Peltonen, Huber Flores, and Petteri Tapio Nurmi. Hot or Not? Robust and Accurate Continuous Thermal Imaging on FLIR cameras. In *2019 IEEE International Conference on Pervasive Computing and Communications*, 12 2018.

34 N. Masson, R. Piedrahita, and M. Hannigan. Approach for quantification of metal oxide type semiconductor gas sensors used for ambient air quality monitoring. *Sensors and Actuators B: Chemical*, 208:339–345, 2015.

35 Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

36 Ministry of Ecology and Environment of the People's Republic of China. Ambient air quality standards, 2012.

37 World Health Organization. Air quality guidelines global update 2005 : particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Technical report, World Health Organization, Copenhagen : Regional Office for Europe, 2006.

38 World Health Organization. Ambient air pollution: a global assessment of exposure and burden of disease. Publication, World Health Organization, 2016.

39 Joana Prata. Airborne microplastics: Consequences to human health? *Environmental pollution (Barking, Essex : 1987)*, 234:115–126, 11 2017.

40 Ole Raaschou-Nielsen, Zorana J. Andersen, Martin Hvidberg, Steen S. Jensen, Matthias Ketzel, Mette Sørensen, Johnni Hansen, Steffen Loft, Kim Overvad, and Anne Tjønneland. Air pollution from traffic and cancer incidence: a Danish cohort study. *Environmental Health*, 10(1):67, 7 2011.

41 Joshua S. Apte, Kyle Messier, Shahzad Gani, Michael Brauer, Thomas Kirchstetter, Melissa Lunden, Julian D. Marshall, Christopher J. Portier, Roel Vermeulen, and Steven P. Hamburg. High-resolution air pollution mapping with Google Street View cars: Exploiting big data. *Environmental Science & Technology*, 51, 06 2017.

42 Houbing Song, Ravi Srinivasan, Tamim Sookoor, and Sabina Jeschke. *Smart Cities: Foundations, Principles, and Applications*. Wiley Publishing, 1st edition, 2017.

43 Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. *Sensors and Actuators B: Chemical*, 238:706 – 715, 2017.

44 Admir Créso Targino, Mark David Gibson, Patricia Krecl, Marcos Vinicius Costa Rodrigues, Maurício Moreira dos Santos, and Marcelo de Paula Corrêa. Hotspots of black carbon and PM2.5 in an urban area and relationships to traffic characteristics. *Environmental Pollution*, 218:475 – 486, 2016.

45 Nada Osseiran Tarik Jasarevic, Glenn Thomas. 7 million premature deaths annually linked to air pollution. News release, World Health Organization, 2014.

46 Claudia Blanco Vidal, Robert E. Dales, and Sabit Cakmak. Air Pollution and Hospitalization for Headache in Chile. *American Journal of Epidemiology*, 170(8):1057–1066, 09 2009.

47 S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750 – 757, 2008.

48 Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sensors and Actuators B: Chemical*, 143(1):182 – 191, 2009.

49 Yang Wang, Jiayu Li, He Jing, Qiang Zhang, Jingkun Jiang, and Pratim Biswas. Laboratory evaluation and calibration of three low-cost particle sensors for particulate matter measurement. *Aerosol Science and Technology*, 49(11):1063–1077, 2015.

50 Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, MobiCom '18, pages 305–320, New York, NY, USA, 2018. ACM.

51 Xin Zhang, Xi Chen, and Xiaobo Zhang. The impact of exposure to air pollution on cognitive performance. *Proceedings of the National Academy of Sciences*, 2018.

52 Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1436–1444, New York, NY, USA, 2013. ACM.

53 N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 11(1):291–313, 2018.

Appendix 1. Impact of Meteorological Factors at Different Levels of PM – Tables

Results from Table 6 are shown in the two rightmost columns for comparison purposes.

Table 15: The errors between true concentrations and predictions from models trained with data selected considering the level of temperature.

	train: temperature						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – low						$PM_{2.5}$ – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	2.34	2.75	2.51	2.92	2.16	2.66	1.38	1.98	9.28	9.58
RF	2.71	3.23	2.05	2.65	2.35	2.93	1.5	2.16	11.68	12.5
ANN (BL)	2.24	2.69	2.5	2.94	5.03	5.39	1.39	1.97	10.23	10.55
ANN	5.28	6.51	2.86	3.3	2.58	3.2	1.4	2.01	7.59	8.32

	train: temperature						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – high						$PM_{2.5}$ – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	6.84	9.04	6.97	9.11	7.57	9.66	4.95	6.78	8.58	10.62
RF	6.73	8.89	7.13	9.3	6.66	8.88	6.31	8.36	8.62	10.71
ANN (BL)	6.65	8.8	6.98	9.18	7.49	9.59	5.0	6.83	10.21	12.02
ANN	5.92	7.84	6.94	9.26	7.0	9.32	4.7	6.35	8.43	10.51

	train: temperature						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – low						PM_{10} – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	4.75	5.48	5.96	6.66	4.27	5.04	2.76	3.63	15.36	16.32
RF	5.6	6.77	5.47	6.98	6.38	7.26	3.31	4.36	12.01	14.58
ANN (BL)	4.15	5.44	5.37	6.27	4.45	5.19	2.81	3.67	15.8	16.73
ANN	4.4	5.16	4.45	5.27	8.16	9.61	2.99	4.03	5.58	7.57

	train: temperature						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – high						PM_{10} – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	13.27	18.99	11.98	17.81	13.14	19.12	11.19	15.29	17.08	22.42
RF	12.61	18.26	12.23	18.09	12.44	18.25	16.29	21.85	17.64	22.91
ANN (BL)	12.1	18.0	22.38	26.86	14.1	20.0	12.83	18.41	17.08	22.48
ANN	15.36	20.92	12.74	18.65	13.84	19.6	12.32	17.23	16.46	21.85

Table 16: The errors between true concentrations and predictions from models trained with data selected considering the level of relative humidity.

	train: humidity						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – low						$PM_{2.5}$ – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	3.23	3.65	2.59	3.03	2.34	2.8	1.38	1.98	9.28	9.58
RF	4.8	6.59	3.03	4.1	2.75	3.64	1.5	2.16	11.68	12.5
ANN (BL)	2.78	3.17	3.18	3.58	2.17	2.64	1.39	1.97	10.23	10.55
ANN	2.74	3.39	2.58	3.2	2.32	3.03	1.4	2.01	7.59	8.32

	train: humidity						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – high						$PM_{2.5}$ – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	6.16	8.3	6.71	8.84	6.86	9.01	4.95	6.78	8.58	10.62
RF	5.84	7.87	6.23	8.52	6.24	8.49	6.31	8.36	8.62	10.71
ANN (BL)	10.22	12.02	6.4	8.61	10.22	12.02	5.0	6.83	10.21	12.02
ANN	6.57	8.64	6.23	8.36	7.83	9.92	4.7	6.35	8.43	10.51

	train: humidity						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – low						PM_{10} – low			
MLR	5.62	6.32	5.39	6.12	5.03	5.74	2.76	3.63	15.36	16.32
RF	7.45	10.29	5.28	6.77	5.32	6.9	3.31	4.36	12.01	14.58
ANN (BL)	3.9	4.84	4.15	5.44	4.83	5.53	2.81	3.67	15.8	16.73
ANN	4.47	5.36	5.72	6.86	5.61	6.53	2.99	4.03	5.58	7.57

	train: humidity						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – high						PM_{10} – high			
MLR	12.05	17.95	12.6	18.65	12.69	18.66	11.19	15.29	17.08	22.42
RF	11.69	17.48	12.8	18.78	12.36	18.04	16.29	21.85	17.64	22.91
ANN (BL)	13.8	19.58	13.1	19.27	14.42	20.23	12.83	18.41	17.08	22.48
ANN	15.64	21.27	12.55	18.6	12.4	18.31	12.32	17.23	16.46	21.85

Table 17: The errors between true concentrations and predictions from models trained with data selected considering the level of pressure.

	train: pressure						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – low						$PM_{2.5}$ – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	2.36	2.76	2.43	2.83	2.97	3.35	1.38	1.98	9.28	9.58
RF	2.27	3.04	2.19	2.76	2.72	3.55	1.5	2.16	11.68	12.5
ANN (BL)	1.98	2.45	2.88	3.26	2.72	3.12	1.39	1.97	10.23	10.55
ANN	1.85	2.53	2.46	3.12	2.06	2.74	1.4	2.01	7.59	8.32

	train: pressure						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – high						$PM_{2.5}$ – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	6.91	9.08	6.99	9.12	6.74	8.86	4.95	6.78	8.58	10.62
RF	7.05	9.34	7.13	9.45	6.54	8.69	6.31	8.36	8.62	10.71
ANN (BL)	7.12	9.29	10.22	12.02	6.51	8.66	5.0	6.83	10.21	12.02
ANN	8.52	10.56	6.45	8.67	7.79	9.83	4.7	6.35	8.43	10.51

	train: pressure						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – low						PM_{10} – low			
MLR	4.79	5.52	5.46	6.21	6.36	7.12	2.76	3.63	15.36	16.32
RF	4.52	5.85	4.82	5.89	4.86	6.27	3.31	4.36	12.01	14.58
ANN (BL)	5.52	6.29	4.15	5.44	6.11	6.87	2.81	3.67	15.8	16.73
ANN	3.86	4.91	4.07	5.15	3.57	4.61	2.99	4.03	5.58	7.57

	train: pressure						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – high						PM_{10} – high			
MLR	12.75	18.68	11.8	17.68	11.47	17.22	11.19	15.29	17.08	22.42
RF	12.87	18.75	12.4	18.35	12.13	17.8	16.29	21.85	17.64	22.91
ANN (BL)	22.38	26.86	11.9	17.83	11.81	17.65	12.83	18.41	17.08	22.48
ANN	13.44	19.39	11.55	17.28	12.96	18.82	12.32	17.23	16.46	21.85

Table 18: The errors between true concentrations and predictions from models trained with data selected considering the level of wind speed.

	train: wind speed						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – low						$PM_{2.5}$ – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	2.28	2.72	2.56	2.96	2.1	2.56	1.38	1.98	9.28	9.58
RF	2.28	2.81	2.28	2.83	2.06	2.74	1.5	2.16	11.68	12.5
ANN (BL)	2.62	3.02	1.58	2.38	1.58	2.38	1.39	1.97	10.23	10.55
ANN	2.32	2.79	2.04	2.57	2.3	2.9	1.4	2.01	7.59	8.32

	train: wind speed						train: $PM_{2.5}$			
training partition	low		diverse		high		low		high	
test partition	$PM_{2.5}$ – high						$PM_{2.5}$ – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	6.8	9.01	6.63	8.85	7.52	9.6	4.95	6.78	8.58	10.62
RF	6.56	8.78	6.97	9.15	7.3	9.52	6.31	8.36	8.62	10.71
ANN (BL)	10.22	12.02	7.25	9.43	10.22	12.02	5.0	6.83	10.21	12.02
ANN	6.66	8.85	7.07	9.26	7.55	9.69	4.7	6.35	8.43	10.51

	train: wind speed						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – low						PM_{10} – low			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	5.92	6.59	5.3	6.04	4.22	4.96	2.76	3.63	15.36	16.32
RF	6.36	8.36	4.66	6.02	3.9	4.89	3.31	4.36	12.01	14.58
ANN (BL)	6.69	7.48	4.15	5.44	4.06	4.82	2.81	3.67	15.8	16.73
ANN	4.95	5.81	5.1	5.95	3.61	4.65	2.99	4.03	5.58	7.57

	train: wind speed						train: PM_{10}			
training partition	low		diverse		high		low		high	
test partition	PM_{10} – high						PM_{10} – high			
error metric	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLR	12.06	17.93	12.17	18.06	13.62	19.54	11.19	15.29	17.08	22.42
RF	11.77	17.55	12.23	18.12	13.88	19.78	16.29	21.85	17.64	22.91
ANN (BL)	12.86	18.74	12.47	18.32	22.38	26.86	12.83	18.41	17.08	22.48
ANN	12.45	18.22	13.19	19.04	11.37	17.05	12.32	17.23	16.46	21.85