Predicting fine particulate matter levels in Finnish buildings

Salvatore della Vecchia

Master’s Thesis

Master of Engineering - Big Data Analytics

2019

MASTER'S THESIS	
Arcada	
Degree Programme:	Master of Enginering in Big Data Analytics
Identification number:	7252
Author:	Salvatore della Vecchia
Title:	Predicting fine particulate matter levels in Finnish buildings
Supervisor (Arcada):	Leonardo Espinosa Leal
Commissioned by:	720 Degrees Oy

Abstract:

Fine particulate matter ($PM_{2.5}$) is considered one of the most harmful air pollutants. While a large proportion of the particles is originating from outdoor sources, people are mostly exposed while indoors. Predicting future trends of $PM_{2.5}$ concentrations could help buildings owners and operators developing better control strategies, and minimizing delays in responding to potential indoor air quality (IAQ) issues. Machine Learning and Deep Learning methods, in particular Long-Short Term Memory Neural Networks (LSTM), have shown good results in predicting sequential data. In this study, $PM_{2.5}$ data from 260 sensors in 119 Finnish buildings were collected during the period 2014/09 - 2019/01. The mean $PM_{2.5}$ concentration observed was 1.01 µg/m$_3$ (SD2.41 µg/m^3). Different methods were compared to predict from one hour up to 8 hours lead times. Three methods were tested for short term predictions (+1 hr): Autoregression, Random forest for regression, and LSTM, while the latter two were tested for long-term predictions (+8 hr). For short term prediction, all methods used a univariate time series (historical hourly indoor $PM_{2.5}$ average). The best prediction was obtained using LSTM with only one lag variable (mean absolute error 0.21 µg/m$_3$, mean squared error 0.85 µg/m^3). For long term prediction, both methods were first tested using a multivariate time series (historical indoor and outdoor $PM_{2.5}$). An additional time series containing the outdoor $PM_{2.5}$ forecasts for the next eight hours was then added to the models, which significantly improved the model accuracy. The lowest Mean Absolute Error (0.49 µg/m^3)and Mean Squared Error (1.83 µg/m^3) were obtained using LSTM with eight lag variables and eight forecasts. In conclusion, long-term predictions are more challenging, but the predictions can be improved by multivariate methods.

Keywords:	Patriculate matter, IAQ, forecasts, multivariate, LSTM, multi-step
Number of pages:	47
Language:	English
Date of acceptance:	03-06-2019

CONTENT

Figures

Tables

1 INTRODUCTION

Indoor air quality (IAQ) is a critical component in people's life. In the modern city life, people spend most of their time in the indoor environment. It has been estimated that in industrialized countries over 90% of the day is spent in indoor spaces (Hope et al. 1998).

Poor indoor air quality can lead to discomfort and illness. Several scientific researches have correlated health problems and complaints with poor indoor air quality (Namieśnik et al. 1992). The problems range from loss in cognitive performance (Allen et al. 2016), headaches, allergic reactions and irritations of the respiratory tract to diseases that can be life-threatening.

In work environments discomfort and illness can lead to lower productivity and absenteeism (Seppänen et al. 2006). Models relating IAQ with health and performance outcomes show the potential benefits from IAQ in terms of cost effectiveness.

Internet of Things (IoT) sensors are nowadays installed in an increasing number of office buildings and schools to continuously monitor thermal comfort and indoor air quality parameters such as temperature, relative humidity, particulate matter, carbon dioxide (CO_2), pressure and volatile organic compounds (VOCs). This continuous monitoring generates a large amount of data that can be analyzed to assess potential risks/benefits related to IAQ over time. Building owners and operators can utilize the information to improve indoor environments, health, and productivity.

Example of typical pollutants affecting the indoor air quality are particulate matter, ozone, carbon monoxide, radon, ozone and volatile organic compounds. The concentration of these pollutants can be predicted in advance using historical data registered by the sensors.

Different methods have already been tested to predict indoor air quality values. The most popular method adopted is the regression model (Allen et al. 2003). For instance, particulate matter and nitrogen dioxide indoor levels were predicted using linear regres-

sion models (Lai et al. 2006). Machine Learning and Deep Learning methods have also been used: Long short-term memory (LSTM)(Hochreiter et al. 1997) and gated recurrent units (GRU)(Cho et al. 2014) have shown good results in predicting time series data, taking into consideration the relationship between the measurements (Ahn et al. 2017). In this study, methods like Autoregression, Random Forest and Long-Short Term Memory Neural Networks are tested to predict short and long-term indoor concentrations of fine particulate matter ($PM_{2.5}$).

A large dataset of measurements is used, with more than a hundred buildings continuously monitored during the last five years in Finland. The dataset is provided by the Finnish IoT company 720 Degrees Oy. The company's mission is to monitor the indoor environment to prevent issues and improve quality in facilities.

The goal of this thesis is to predict in advance indoor air quality anomalies related to $PM_{2.5}$ and promptly inform building owners, operators and occupants. It will help them to take the right actions to improve indoor environments, health and productivity. Checking the status of the filters in the ventilation system, limiting the amount of injected air or keeping windows closed during periods of high outdoor particulate matter levels are some examples of measures that can be taken.

1.1 Background and need

Before the 1970s, the attention for the air quality was mainly focused on the outdoor. After that, indoor air quality started to be monitored mainly in industrial workplaces to prevent occupational diseases and absenteeism (Chu et al. 2003). The situation changed with the introduction of modern buildings, which have a limited amount of fresh air introduced in the indoor spaces. For energy-saving reasons they have tighter sealing and, in many cases, mechanical ventilation and air conditioning.

Building occupants started to be more isolated from the outdoor, spending time in poor indoor environments. This leads to more complaints, more sickness, absenteeism and lower productivity. On a national scale, the number of increased diseases can also impose more costs to the medical system. All these reasons brought to an increasing attention around indoor air quality (Chu et al. 2003).

Even though the topic is quite new, the level of knowledge is rapidly increasing together with the number of studies and researches. The use of modern IoT sensors combined with the new knowledge is already helping facility managers to solve indoor air quality problems. The solution provided by 720 Degrees for instance was adopted in Finland by several companies and is helping to reduce employees' complaints and solving cases of harmful levels of indoor pollution by identifying and removing the sources.

1.1.1 Particulate matter

One of the most dangerous pollutants is particulate matter (PM). Also known as atmospheric particulate matter or particulates, PM is microscopic solid or liquid matter suspended in the Earth's atmosphere. Particulate matter is the particle alone, while the particulate/air mixture may be considered an aerosol. One can visualize PM as tiny pieces 'hanging'/'floating' in the air. Very slowly settling due to gravity. The time it needs for a particle to settle depends on its size and weight. PM also correlates with humidity.

Particulate matter can be a local phenomenon (indoor-office-source) and it propagates in office-space in proportionally less than, for instance, volatile organic compounds. Since particle is a solid or liquid, its propagation requires some external source of propagation, like drifts. The global effect (high PM values on the whole floor of a building) is observed when the source of PM is external.

Subtypes of atmospheric particulate matter include:

- suspended particulate matter (SPM)
- respirable suspended particle (RSP), which are particles with a diameter of 10 μm or less, also known as PM_{10}. Coarse inhalable particles such as those found near roadways and dusty industries (US EPA, 2019).
- fine particles with a diameter of 2.5 μm or less, aka. $PM_{2.5}$, found in smoke and haze, directly emitted from sources such as forest fires, or they can form when gases emitted from power plants, industries and automobiles react in the air. Diesel engines are the main source (Omidvarborna et al. 2015)
- ultra-fine particles, of nanoscale size, less than 0.100 μm in diameter

- doot, a mass of impure carbon particles resulting from the incomplete combustion of hydrocarbons.

1.1.2 Sources

There are different types of sources for particulate matter. **Natural,** when is originating from volcanoes, dust storms, forest and grassland fires, living vegetation, and sea spray. **Human**, when is generated by human activities like burning of fossil fuels in vehicles (US EPA, Particulate Matter, 2019), reaction of gases or droplets in the atmosphere from sources such power plants, various industrial processes, coal combustion for heating homes and supplying energy. **Indoor** sources can be tobacco smoke, cooking (e.g., frying, sautéing, and broiling), burning candles or oil lamps, and operating fireplaces, heaters. Fine particles can be carried long distances from their source: for instance wildfires or volcanic eruptions can raise fine particle concentrations hundreds of km from the event(Coco Liu et al. 2016.

1.1.3 Effects

Exposure to high concentration of particulate matter may cause premature death in people with heart or lung disease, heart attacks, irregular heartbeat, aggravated asthma, decreased lung function, and increased respiratory symptoms, such as irritation of the airways, coughing or difficulty breathing. People with heart or lung diseases, children and older adults are the most likely to be affected by particle pollution exposure (US EPA, Health and Environmental Effects of Particulate Matter, 2019). However, even healthy adults may experience temporary symptoms from exposure to elevated levels of particle pollution.

Exposure to particulate pollution at a work place, can result in general discomfort, absences from work, especially for those with pre-existing heart or lung diseases. The size of particles is directly linked to their potential danger (US EPA, Health and Environmental Effects of Particulate Matter, 2019). Coarse particles (PM_{10}) are of less concern, although they irritate one's nose, eyes and throat. Smaller particles can pass through the throat and nose and enter the lungs. Once inhaled, these particles can affect

the heart and lungs and cause serious health effects. $PM_{2.5}$ and PM_{10} can go into the deepest part (alveolar) of the lung where gas exchange occurs between air and blood. These are the most dangerous particles because the alveolar portion of the lungs has no effective means to eliminate and if the particles are water soluble, they can pass into the bloodstream within minutes. If they are not soluble in water, they remain in the alveolar lung for a long time. Soluble elements may be of PAHs (Polycyclic Aromatic Hydrocarbons) or residues of benzene classified as carcinogenic.

According to a research conduced in 2014 to asses the environmental burden of disease in Europe, "About 3–7% of the annual burden of disease in the participating countries (Belgium, Finland, France, Germany, Italy, and the Netherlands) is associated with the included environmental risk factors. Airborne particulate matter (diameter ≤ 2.5 μm; $PM_{2.5}$) is the leading risk factor associated with 6,000–10,000 DALYs/year and 1 million people" (Hänninen et al. 2014).

1.1.4 Norms

To reduce the exposure of the population to fine particles, the European Union set objectives that must be achieved at national level. They are based on the outdoor average exposure indicator (AEI), determined using a 3-year running annual mean of the $PM_{2.5}$ concentration, (Directive 2008/50/EC of the European Parliament, 2008).

Table 1. European emission standards, PM in the outdoor air according to the EU Directive 2008/50/EC of the European Parliament and of the Council (Directive 2008/50/EC of the European Parliament, 2008)

Description	PM_{10}	$PM_{2.5}$
Max. yearly average [μg/m³]	40	25
Max. daily average (24h) [μg/m³]	50	None
Allowed number of exceedances per year	35	None
Date since the rule:	01.01.2005	01.01.2015

For the indoor air, the World Health Organization guidelines are designed to provide appropriate targets for the air quality management. For particulate matter, they are reported in the table 2.

Table 2. Air quality guidelines and their rationale (WHO Air quality guidelines, 2005)

Description	**PM_{10}**	**$PM_{2.5}$**
Annual mean	20 µg/m³	10 µg/m³
24-hour mean	50 µg/m³	25 µg/m³

1.2 Statement of the problem

The problem that this thesis is addressing is to help building owners, operators and occupants to reduce indoor air quality issues related to fine particulate matter. Predicting indoor air quality anomalies will help them to take actions before problems can appear. This will be done by predicting the absolute values for $PM_{2.5}$ from 1 to 8 hours in advance. The forecast can be compared with the values recommended in the guidelines to predict potential harmful situations.

1.3 Purpose of the study

The aim of the thesis is to predict $PM_{2.5}$ values for different intervals.
When a sensor is deployed in a building, it starts collecting measurements and sending them to the central server. It takes a few days before the data provided can be used to make predictions. The study is therefore divided into two parts.

The first part is focused on short-term predictions (one hour lead $PM_{2.5}$ concentration), using only the indoor historical data. Having little data available (like 40 hours of $PM_{2.5}$ values recorded) is enough to predict only one step ahead.

In the second part instead, the $PM_{2.5}$ values from 2 to 8 hours in the future are predicted. In this case, indoor, outdoor historical data and outdoor forecasts are used in the models. The assumption is that both indoor and outdoor particulate matter are time dependent variables and that each variable depends on its past values and has some dependency on other variables. The dependency between these variables is then used to forecast future indoor $PM_{2.5}$ values. The hypotheses is that using a method that includes also outdoor

forecasts reduces the prediction error. The problem can be seen as a forecasting task with multivariate time series input. This study tests different methods to solve it. Furthermore, since the goal is to predict the concentration of $PM_{2.5}$ at different intervals (+1h, +2h..., +8h), we can define the problem as a multi-step sequence forecast.

1.4 Dataset

The dataset is provided by the Finnish company 720 degrees Oy. It contains fine particulate matter measurements recorded by IoT sensors installed in conference rooms, corridors, office rooms, classrooms of schools and office buildings. The measurements are recorded approximately every 15 seconds by each sensor. For most of the spaces monitored data have continuous measurements for long periods of time, with a few missing data points. The dataset contains also hourly outdoor measurements for particulate matter registered by the Finnish Meteorological Institute (Finnish Meteorological Institute, 2019) and particulate matter forecasts up to 96 hours ahead, provided by Breezometer Ltd (Breezometer, 2019).

1.5 Definitions

Heating, ventilation, and air conditioning (HVAC) is the technology to control the ambient environment in residential, commercial and industrial buildings. It regulates temperature, humidity, air flow and air filtering values.

Indoor air quality (IAQ) is the air within or around buildings and structures and it refers to the health, comfort and well-being of occupants. Monitoring and controlling the levels of pollutants in the indoor spaces can help reducing the risks of indoor health concerns.

IoT: Internet of things is the interconnection between different types of devices that send and receive data (for instance measurements) at different time intervals.

Particulate matter (PM) also referred as atmospheric particulate matter or particulates is the microscopic matter suspended in the Earth's atmosphere. It can be solid or liquid.

Particulate matter is the particle alone. The particulate/air mixture is referred as the aerosol. Particulate matter settles slowly on the ground due to gravity. Small and light particles take longer to get to the ground. Based on the size of the particles, PM is usually divided into two groups:

- the larger particles, the coarse fraction, from 2.5 to 10 µm (PM_{10} - $PM_{2.5}$)
- the fine fraction, with particles with a size up to 2.5 µm

1.6 Limitations

This study focuses only on predicting fine particulate matter values, even though the dataset contains measurements for other factors like temperature, humidity, VOCs, CO_2 and pressure. This decision was taken based on the curiosity to compare different methods using univariate and multivariate time series, with only a univariate output. Furthermore, this project doesn't aim to implement any functionality that will send automatic commands to the any ventilation system, but the method and the procedures described can be used as guidelines. The only aim is to predict the levels of fine particulate matter and eventually send notifications to building operators and occupants.

1.7 Ethical considerations

The details regarding the measurement instruments, the descriptive statistics of the initial dataset and the techniques used to preprocess data will remain classified. The description of the problem and the results obtained will instead be public and discussed at the end of the document.

2 LITERATURE REVIEW

2.1 Previous work

During the past years, multiples attempt to predict air pollutants values have been done, using different techniques, from using simple linear regression to more complicated methods like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) (Ahn et al. 2017).

In the work of Allen et al. (2003), “the use of light scattering data to estimate the contribution of indoor- and outdoor- generated particles to indoor air”, the author showed the results of predicting the concentration of indoor particulate matters using a regression model. The aim of the research was to separate indoor generated particles from ambient-generated components.

In the work of Lai et al. (2006), “Determinants of indoor air concentrations of $PM_{2.5}$, black smoke and NO_2 in six European cities (EXPOLIS study)”, the authors also applied regression models to predict the concentration of $PM_{2.5}$. Measurements of $PM_{2.5}$, black smoke (BS), and nitrogen dioxide (NO_2) were taken for two consecutive days in different locations and different dates and seasons.

Recent studies have shown that indoor pollutants prediction require more complex models than linear regression, because of the existing relationship between different type of data, such as particulate matter and meteorological data. Better results were obtained with the development of Neural Networks to improve the prediction of air pollutants, especially with Recurrent Neural Networks (RNNs), with their ability to learn temporal sequences. However also these methods showed some limitations because the time lag must be defined in advance. Furthermore, RNNs are not suitable because they fail to extract long time dependencies.

Even though ML and DL techniques are not necessarily the best methods when forecasting time series (Makridakis et al. 2018), good results have also been obtained when using long short-term memory neural network (LSTM) and gated recurrent unit (GRU), especially combining different types of measurements in the same model. Those two methods have been used in various applications with great success because of the gate and memory mechanism.

In the work of Ahn et al. (2017) “Indoor Air Quality Analysis Using Deep Learning with Sensor Data”, the authors created a microchip made of several sensors, to measure carbon dioxide, fine dust, temperature, humidity, light intensity and VOCs. The measurements were recorded periodically. The models presented were able to predict the time series data using machine learning, taking advantage of the relationship between factors. The methods tested were long-short term memory (LSTM) and gated recurrent units (GRU). Data for the model was prepared as a three-dimensional tensor: the two-dimensional tensor is created merging the values from each sensor and arranging them as a set of six values. The third dimension is the time step size t. The goal of the project was to predict the measurements when t+1 has elapsed. GRU produced the best results. It was used with two hidden layers of 1270 nodes with sigmoid activation function and ADAM as optimization algorithm. Furthermore, to determine the best observation period for the model, the authors suggest the use of a specific algorithm. A brute force method to define the best time step within a range of possible values can be very expensive in terms of performance. Thus, the algorithm that they present was able to produce better results in terms of learning time and performance. The paper can be used as a starting point to predict the measurements. Its main limitation is that it tries to predict values at time step +1. This project aims to predict the measurements at different time steps.

2.2 State of the art

An interpretable long short-term neural network using multi-variable LSTM for time series with exogenous variables (Guo et al. 2018) can be instead used when, in addition to the target time series values, we want to understand the different importance of each exogenous variables. A predictive model that combines both historical data of the target and exogenous variables is an autoregressive exogenous model (ARX). In the paper, the

authors used additional time series corresponding to exogenous variables. They used the $PM_{2.5}$ time series as a target, and data such as dew point, temperature, pressure, wind direction, speed, hours of snow and hours of rain were used as exogenous time series. LSTM itself fails to understand the importance of each variable. To solve this problem, the authors used each neuron of the recurrent layer to encode information only for a each variable. In this way, they were able to extract the attention value for all the exogenous variables.

A novel long short-term memory neural network extended for air pollutant concentration prediction was presented recently (Li et al. 2017). Historical air pollutant data was combined with auxiliary data such as meteorological data and time stamp data. The LSTME model proposed was able to extract useful features from the data, considering also the spatiotemporal correlations between the stations where the $PM_{2.5}$ data was collected. The model suggested was also able to provide a multi-scale method to predict particulate matter concentration. Short- and long-term predictions were performed in intervals of 1h, 2h, 3h, 4 to 6h, 7 to 12 h and 13 to 24. The accuracy of long-term periods prediction was reduced compared to short-term periods, but the performance of the proposed model can still be considered suitable for long-term prediction tasks.

Recently, short and long-term hourly outdoor pollutants forecasts became available. It is now possible to know what will be the concentration of particulate matter up to 96 hours in the future. In this study the forecasted values are included in the models, toghether with the historical indoor $PM_{2.5}$ measurements and the outdoor $PM_{2.5}$ measurements. Particulate matter concentrations for different time intervals t (t+1, …, t+8) are predicted. The results obtained using different methods, with forecasts or only historical data, are compared.

3 MATERIAL AND METHODS

3.1 Data [classified]

In this paragraph, the three different types of data are described: $PM_{2.5}$ indoor measurements, $PM_{2.5}$ outdoor measurements, $PM_{2.5}$ outdoor forecasts

3.1.1 Indoor measurements [classified]

In this paragraph, the method used to collect data is described: details about the sensors used and the type of measurements are provided.

3.2 Time series forecasting

Time series is a collection of observations of a variable at different times, usually at equal intervals and in chronological order. When we have a series with a single time-dependent variable, we define it as a univariate time series. The indoor $PM_{2.5}$ values measured at hourly intervals are an example of univariate time series.

Table 3. The indoor particulate matter data, univariate time series

building_id	**sensor_id**	**record time**	**value**
building_1	sensor_2	2018-12-18 13:00:00	0.83
building_1	sensor_2	2018-12-18 14:00:00	0.92
building_1	sensor_2	2018-12-18 15:00:00	0.89
building_1	sensor_2	2018-12-18 16:00:00	1.13
building_1	sensor_2	2018-12-18 17:00:00	1.11

If there are two or more time-dependent variables, they depend on their past and they are correlated to each other, we name the series a multivariate time series. The assump-

tion of this study is that outdoor and indoor concentrations of particulate matter are correlated, and the two variables also depend on their previous values.

Table 4. The indoor and outdoor particulate matter data, multivariate time series

sensor_id	**record time**	**indoor value**	**outdoor value**
sensor_id_2	2018-12-18 13:00:00	0.83	11.31
sensor_id_2	2018-12-18 14:00:00	0.92	11.11
sensor_id_2	2018-12-18 15:00:00	0.89	11.51
sensor_id_2	2018-12-18 16:00:00	1.13	10.94
sensor_id_2	2018-12-18 17:00:00	1.11	9.89

3.2.1 Stationarity and non-stationarity of time series

Time series can be stationary or non-stationary. If the mean, variance and covariance don't change with time, the time series can be defined as stationary. It also means that it doesn't show a trend. Vice versa, a non-stationary time series is a series where the properties depend on time. Most statistical models require a series to be stationary in order to get reliable predictions.

3.2.2 Autoregression model

One of the most used method to solve univariate time series forecasting is autoregression. A regression model like linear regression is a method to examine the relationship between two or more variables. For example:

$$y_t = b_o + b_1 \cdot x_1$$

Where y_t is the variable to be predicted, b_0 and b_1 are coefficients found by optimizing the model and x is the input variable. The previous method can be applied to time series, using as input variables the previous observations at different time steps (lag variables). For instance, if we want to predict the value for a variable at a time step in the future (t+1) using the last 2 lags variables, we would have the following regression model:

$$x_{(t+1)} = b_o + b_1 \cdot x_{(t-1)} + b_2 \cdot x_{(t-2)}$$

A regression model of this type is named autoregressive because the output value depends linearly on its previous values.

3.2.3 Autocorrelation

An autoregression model is based on the assumptions that the previous observations of the input variable are correlated. We grave a positive correlation when both variables change up or down together. Vice versa, if they change in opposite directions, this is called negative correlation. In this case we are analyzing the correlation of a variable with its values at previous time steps, so we call it autocorrelation.

When exploring a new dataset, it is important to investigate the correlation between the output variable and the lagged variables. If it appears that there is no correlation, it means that it is going to be very difficult to forecast the time series problem. Instead, if there is a strong correlation, then the autoregression model can put more weight on that variable when modeling.

There are different tests that can help assessing the autocorrelation:

- plotting the observations at the previous time step (t-1) with the observation at the current time step (t) as a scatter plot. If the graph shows a linear relationship, it means that there is a correlation between the variables
- the Pearson correlation coefficient (SPSS Tutorials 2019) is a good indicator to investigate if two variables are correlated. It returns a number between –1 and 1. If the value is negative, the variables have a negative correlation, if the result is positive, there is a positive correlation. The closer the coefficient is to 1 (or –1), the stronger the correlation is. Usually values above 0.5 show that there is a correlation

3.2.4 Random Forest

Random forest is part of the ensemble family. Ensembles combine multiple machine learning models to solve regression and classification problems. Random forest can be

described as a collection of decision trees. Because one of the problems of the decision trees is over-fitting, random forest creates a collection of decision trees, each one slightly different from the others. In this way, the over-fitting problem is reduced by averaging the results obtained with each tree. During the process of generating several decision trees, random forest injects randomness to make sure that each tree is different, but it has still a good predictive power (Andreas et al. 2016).

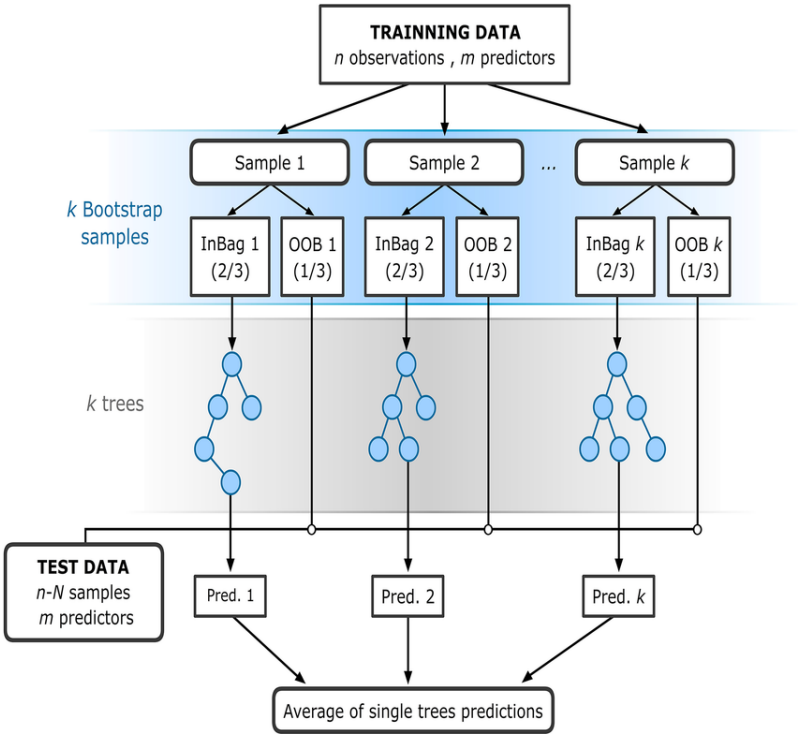


Figure 1. The flowchart of random forest (RF) for regression (Rodriguez-Galiano et al., 2015b)

As a first step, the number of trees to build needs to be defined (it is one parameter when initializing the model). These trees will be different from each other and each tree is built using a bootstrap sample of the data. A bootstrap sample is obtained by randomly replacing the values for a specific number of data-points. The outcome is a new dataset, with the same size as the original dataset but with some repeated or missing data-points.

As a next step, the algorithm builds a decision tree on the new dataset, but it doesn't look for the best test for each node. It randomly selects a subset of features for each node and search for the best set for this subset. The number of subset features is a pa-

rameter that can be tuned in the model. If it is equal to the number of features of the dataset, there will be no randomness in the selection of the features. On the other hand, if we select only a single feature, there will be no choice on which feature to test.

The bootstrap sampling and the random selection of features described ensure that all the trees built during the process are different. For regression, the algorithm makes a prediction for each tree in the forest and uses the average of the results as a final prediction.

Random forest was used in 2012 at the EMC Data Science Global Hackathon (Air Quality Prediction). The winner of the competition generated 390 random forest models (39 target variables x 10 intervals) (EMC Data Science Global Hackathon, 2012).

3.2.5 Long-Short Term Memory Networks (LSTM)

Long-Short Term Memory Networks are a special kind of Recurrent Neural Network (RNN). RNNs are artificial neural networks with loops in them that allow the persistence of the information.

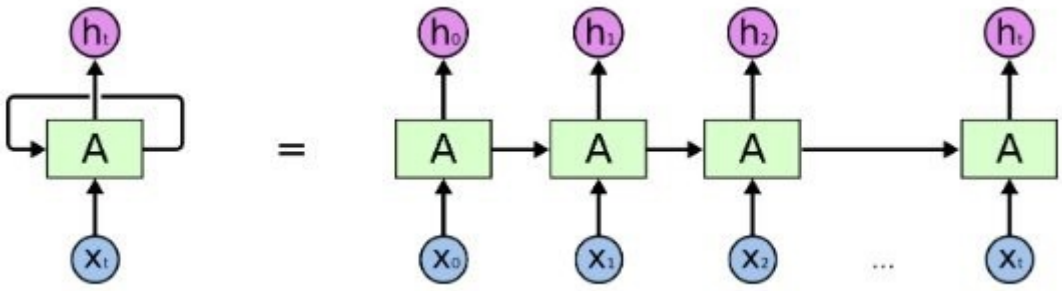


Figure 2. An unrolled neural network (Olah 2015)

The above diagram shows an unrolled neural network. A chunk of neural network A receives an input x_t and returns an output h_t. The information learned is passed to the next step of the network. RNNs can be seen as multiple copies of the same network where each network transfers the information to the successor. The problem with this type of network is that, when the gap between steps grows, they are unable to transfer the information. This means that RNNs work well when we need to predict data based on the re-

cent information, but not in cases of long-term dependency. This problem is solved by LSTM networks. They can learn long-term dependencies, keeping the information between loops for long periods of time. They have an internal state variable that is transferred from one cell to the other and modified by Operation Gates. LSTM networks are capable of deciding how long keep an information, when to discard it and how to connect old memory and new input.

3.2.6 Multi-step Time Series Forecasting

Since the goal of this study is to predict the value of a variable (indoor $PM_{2.5}$) at different intervals in the future, we are facing a multi-step time series forecasting problem. There are different methods to predict a variable at different intervals:

- **Direct Multi-step Forecast Strategy**: with this approach, a model is developed for each of the steps that will be predicted. In this case, we are predicting the values of the output variable for the next 8 hours and we would require 8 different models for each sensor. A similar solution requires high computational power and it is more difficult to maintain. Another disadvantage is that, having separate models doesn't consider the relationship between the predicted values.
- **Recursive Multi-step Forecast Strategy**: with this method, the predicted value at time-step +1 is then used as an observation to predict the output variable a time-step t+2. The limit of this technique is that the prediction error increases each time we use a predicted value instead of an observed value.
- **Direct-Recursive Hybrid Multi-Step Forecast Strategies**: is a technique that combines the two previously described methods. A model is still implemented for each time-step to be predicted, but each model uses as an input the value predicted by the other model.
- **Multiple Output Forecast Strategy**: this method uses a single model which has as output variables the entire sequence to be predicted. In this study, the output variables of a single model would be eight (t+1, t+2...t+8). This approach requires more complex models and more data to train them to avoid over-fitting problems. In this study the multi output forecast strategy is used.

3.3 Framework/procedure

The following framework will be followed as a guide during the research process:

- phase 1 – Data extraction. Raw data is extracted from the database, aggregated by hour, filtered (e.g. by country and factor type) and exported into csv format.
- phase 2 - Data preparation and feature extraction. During this phase, the subset of the dataset extracted is prepared: missing data points are handled and the observations are aggregated. Features are then extracted, data is scaled according to the different method used. The data prepared is passed to the next phase to be processed.
- phase 3 - The dataset prepared is splitted into test and training set. The methods described in the next section are applied on the test dataset and are evaluated using different metrics.

The metrics used to evaluate the goodness of the results are the Mean Absolute Error and the Mean Squared Error. The **Mean Absolute Error** (MAE) is calculated by getting the absolute value of the residuals for each data point:

$$MAE = \frac{\sum_{i-1}^{n} |y_i - x_i|}{n}$$

where n is the number of errors, Σ is the summation symbol, y_i is the actual output value and x_i is predicted output value. The **Mean Squared Error** (MSE) is similar to the MAE but it squares the difference before summing them all:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where n is the number of errors, Σ is the summation symbol, Y_i is the actual output value and Y_i is predicted output value. The best method is selected based on the previous results.

4 DATA EXPLORATION AND PREPARATION[CLASSIFIED]

4.1 Data collection [classified]

In this paragraph, details about the initial indoor dataset are described: amount of data, number of buildings and cities monitored, number of sensors for each building.

Details on how the outdoor measurements have been collected during the time are also provided.

4.2 Data exploration[classified]

4.2.1 Descriptive statistics [classified]

In this paragraph the descriptive statistics are presented: $PM_{2.5}$ concentration observed in Helsinki and in the rest of Finland and the comparison of the indoor and outdoor values by week and month.

4.2.2 Stationarity of a Univariate time series[classified]

4.2.3 Stationarity of a Multivariate Time Series[classified]

In this paragraph the stationarity of the multivariate time series is investigated.

4.2.1 Autocorrelation[classified]

In this paragraph the autocorrelation of the time series is investigated.

4.3 Data preparation [classified]

In this paragraph the methods to prepare data are described.

4.3.1 Resampling [classified]

The resampling method is presented.

4.3.2 Transform the time series into a supervised learning problem [classified]

How the problem is transformed into a supervised learning problem is described in this paragraph.

4.3.3 Dataset division[classified]

How the dataset is splitted into train, validation and test is described here.

5 RESULTS

The process of data collection, dataset processing, feature selection, and dataset division was presented in the previous chapter. In this chapter the methods previously described will be used to predict both short- and long- term indoor $PM_{2.5}$ values. For short-term predictions, Autoregression, Random Forest and Long-Short Term Memory Neural Network are used, with a univariate time series input (indoor concentration). The output is a single $PM_{2.5}$ value at time t+1. Mean Squared Error and Mean Absolute error are used as evaluation metrics. Different numbers of lag variables are tested for the models. The execution time to train models is also tracked and compared between methods.

For long-term predictions, Random Forest and Long-Short Term Memory neural network are compared. Different number of lag variables are tested for the models. Both methods are tested initially using as input only the historical multivariate time series (indoor and outdoor $PM_{2.5}$), then also including as features the outdoor $PM_{2.5}$ forecasts at time t+1, t+2 ... t+8. The hourly MSE and MAE are finally compared to select the best method.

All the experiments were executed on a P2 Instance Type for Amazon EC2 with 4 CPUs, 61 GiB of Memory, 1 GPU NVIDIA GK210 with 12 GiB of memory and 2,496 Parallel Processing Cores. To train the LSTM modesl TensorFlow(+Keras2) with Python3 (CUDA 9.0 and Intel MKL-DNN) was used.

5.1 Short-term particulate matter concentration forecast

Short term predictions (one-hour lead) have been separated from long-term predictions because the model to predict only the next hour requires much less data then predicting eight hours leads. In a real case, after a sensor is deployed in a space, the 1hr forecast will be available much earlier than the long-term forecast.

5.1.1 Autoregression model

To forecast the lead hourly indoor particulate matter concentration, an autoregression method was tested. One model was implemented for each of the sensors monitored.

Only nodes with at least 50 hourly measurements available were used. A total of 260 autoregressive models were implemented. The historical hourly indoor $PM_{2.5}$ concentration was used as input. The model was implemented using the autoregression model statsmodels.tsa.ar_model provided with the Python statistical module Statsmodels.

The appropriate number of lag values for each model was automatically selected by a method provided with the Statsmodels library, which uses statistical tests and trains a linear regression model.

Table 5. Results using an autoregression model to predict one lead time indoor $PM_{2.5}$. The ovarall time to train all the models is presented.

Avg. no. of lags used	**Max. no. of lags used**	**Min. no. of lags used**	**Mean Squared Error**	**Mean Average Error**	**Training time (secs)**
29	9	44	1.42	0.28	80

5.1.2 Random forest

As an alternative method to autoregression, Random Forest was tested. One model was implemented for each of the sensors monitored. Only nodes with at least 50 hourly measurements available were used. A total of 260 Random Forest models were implemented. The dataset for each node was splitted into train and test set in the identical way of the previous method. To find the optimal number of lag variables for the models, different values have been tested.

Lags tested: 2, 4, 6, 8, 10, 12, 14, 16, 20, 25, 30, 35, 40, 50, 55, 60, 65

If 40 lags variables are used, it means that to predict the Indoor concentration of the next hour, the previous 40 observations of indoor $PM_{2.5}$ were used. The model was implemented using the Random Forest Regressor provided with scikit-learn, the machine learning library for the Python programming language.

rf =RandomForestRegressor(n_estimators=n)

The parameter “number of estimators” is the number of trees in the forest. Different number of estimators were tested (using 2 lags): 10 (MSE 1.34), 30 (MSE 1.32) and

100 (MSE 1.30), with the latter being the best parameter. Quantitative results and execution times are described in the table below.

Table 6. A comparison of the number of features used in the model, the results and the time in minutes to train all the 260 models

lags	**2**	**4**	**6**	**10**	**12**	**14**	**16**	**20**	**25**	**30**	**35**	**40**	**50**	**55**	**60**	**65**
MSE	1.30	1.26	1.48	2.22	2.11	2.29	2.23	1.91	2.46	3.03	2.37	1.50	1.24	**1.23**	1.24	1.23
MAE	0.28	0.28	0.29	0.32	0.32	0.32	0.32	0.31	0.32	0.35	0.34	0.31	0.27	**0.27**	0.27	0.27
Time	3	4	6	9	10	12	14	17	21	25	30	34	44	**49**	54	60

The previous table shows that the lowest error is obtained using 55 features. Also 4 features gave good results, but with much lower time needed to train the models. Month and city were initially added as categorical variables but didn't bring any improvement to the model.

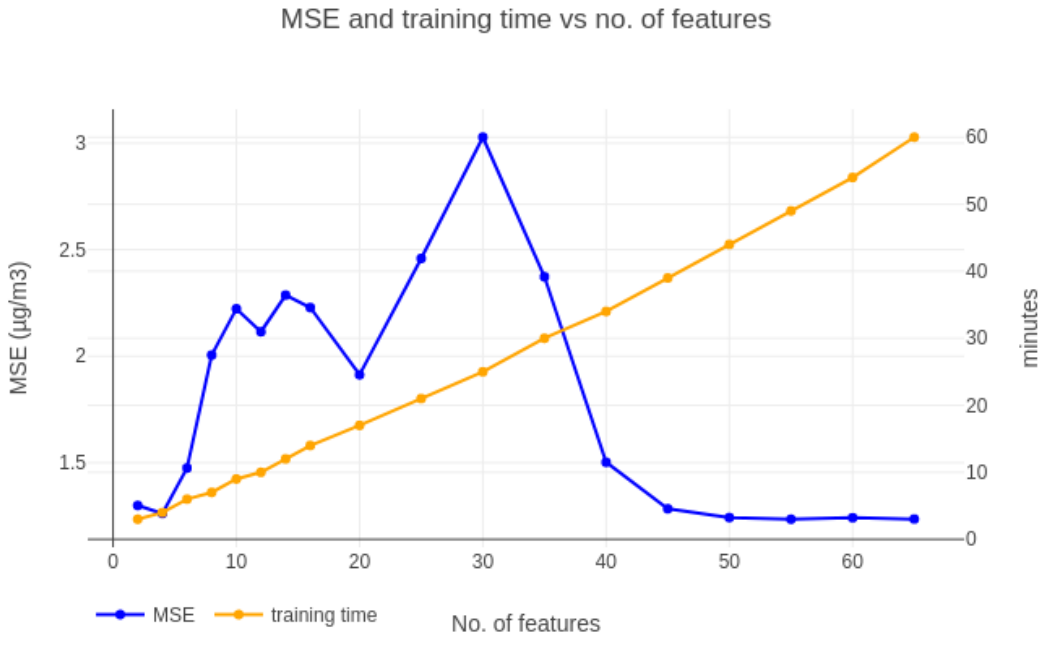


Figure 3. A comparison between the number of features (lag variables used), the Mean Squared Error and the overall time needed to train the models (in minutes).

The previous picture shows that increasing the number of lag variables more than a certain level doesn't improve significantly the accuracy of the model. The lowest MSE is obtained when using the 55 previous hourly observations.

5.1.3 Long-Short Term Memory neural network

A Long-Short Term Memory network was also tested to predict the indoor concentration of particulate matter at time t+1. The network takes as input the n lag values of $PM_{2.5}$ at time t...t-n. One model was implemented for each sensor with at least 500 observations. The LSTM model expects the input as [samples, timesteps, features], where samples is the number of observations available and it varies for each sensor. Timesteps is the number of lag variables. The number of features in this case is equal to one, since we are using a univariate input.

Training a LSTM network for each sensor is a time-consuming task and testing a high number of different lags for all the sensors would require too much time. To have an estimate of which number of lags produces the best results, only one sensor was selected. The best results were obtained using 2 lags, so it was selected as the number of timesteps used to train all the LSTM models. The models were trained using a EC2 Amazon instance p2.xlarge with for TensorFlow(+Keras2) with Python3 (CUDA 9.0 and Intel MKL-DNN) and GPU NVIDIA.

LSTM model parameters: the LSTM model implemented contains a single hidden LSTM layer with 200 units, followed by a fully connected layer with 200 nodes to interpret the LSTM layer. The output layer will predict in this case a single value, the indoor $PM_{2.5}$ a time t. The mean squared error is used as loss function. The number of epochs selected was 70 and the batch size 16, relatively small to let the model learn different mapping of inputs to outputs each time it is trained. Adam was selected as optimizer to update the network weights (Brownlee J 2019).

Table 7. Descriptive statistics for MSE and MAE using a LSTM network with 2 lags for all nodes

	Mean	**Std.**	**Min.**	**25%**	**50%**	**75%**	**Max.**	**Training time**
MSE	0.85	4.18	0.01	0.07	0.13	0.34	40.49	16.3 hrs. (avg 3.7 minutes per model)
MAE	0.21	0.27	0.01	0.1	0.14	0.24	3.39	

The highest prediction error (MSE = 40.49) happened in a space with several cases of indoor generated particulate matter, probably by occupants. These cases are more difficult to predict.

5.1.4 Comparison of results

Table 8. Comparison of the MSE and MAE obtained with different methods

Method	**Lags**	**MSE**	**MAE**	**Training time**
AR	29	1.42	0.28	80 secs.
Random Forest	10	1.26	0.28	27 mins. (avg 0.16 minutes per model)
LSTM	2	0.85	0.21	16.3 hrs. (avg 3.7 minutes per model)

As we can see from table 9, the best results when predicting indoor concentration at lead time t+1 were obtained using a LSTM network with 2 time-steps. On the other end, even though it is the method that provided the best results, LSTM requires several hours to train all the models and a bigger amount of data compared to the other methods.

5.2 Eight hours particulate matter concentrations forecast

Once a sensor has collected enough data, longer predictions can be made. The second part of the study focuses on predicting the average indoor concentration of $PM_{2.5}$ for the following 8 hours. The goal is to predict 8 output variables, using a multi-output forecast strategy. Only nodes with at least 500 observations, for a total of 240 models were implemented, one for each sensor. Random Forest and LSTM were selected as methods, testing both methods first using only historical data (indoor and outdoor), then adding the 8 hours forecasts to the model.

5.2.1 Random forest

Results using Random forest for regression using multivariate time series as input are presented. Two methods were tested: using only historical data and adding forecasts of outdoor $PM_{2.5}$ (8 lead variables).

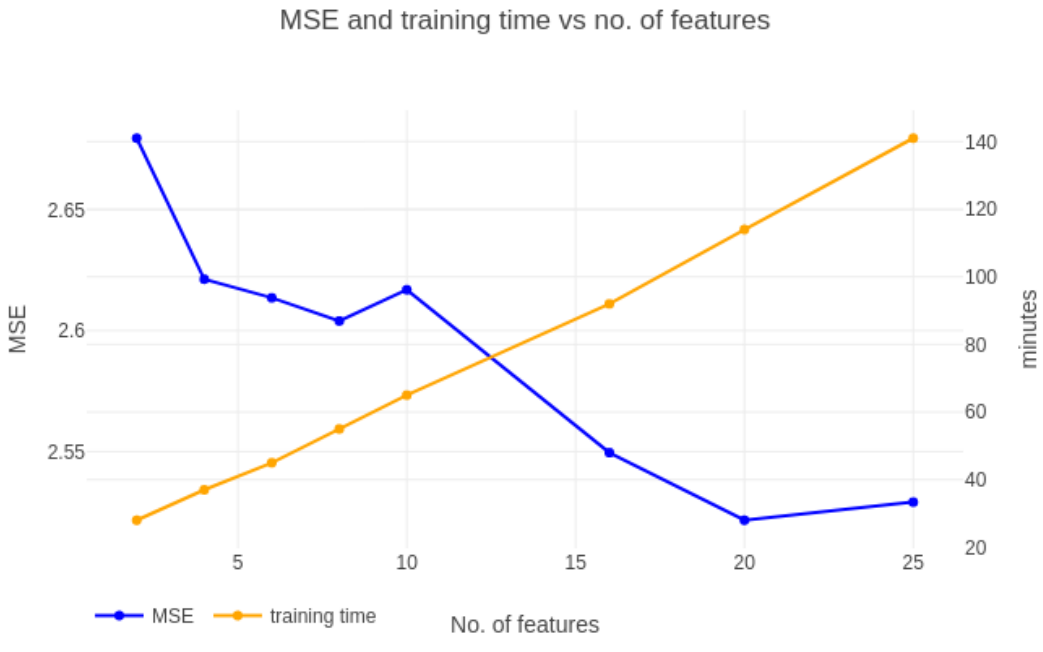
Table 9. MSE using Random forest with historical indoor, outdoor and forecasts

lags	t+2	t+3	t+4	t+5	t+6	t+7	t+8	Total	Training mins.
2	1.85	2.29	2.62	2.89	3.14	3.48	3.87	2.68	28
4	1.85	2.23	2.54	2.82	3.10	3.42	3.75	2.62	37
6	1.81	2.27	2.57	2.83	3.07	3.37	3.71	2.61	45
8	1.81	2.26	2.56	2.82	3.06	3.34	3.68	2.60	55
10	1.83	2.29	2.58	2.82	3.07	3.35	3.70	2.62	65
16	1.81	2.25	2.51	2.74	2.99	3.25	3.55	2.55	92
20	**1.82**	**2.23**	**2.48**	**2.70**	**2.95**	**3.21**	**3.49**	**2.52**	**114**
25	1.80	2.18	2.44	2.71	2.97	3.25	3.58	2.53	141

Different number of estimators were tested (using 2 lags): 10 (MSE 2.97), 30 (MSE 2.93) and 100 (MSE 2.89), with the latter being the best parameter.

Table 10. MAE using Random forest with indoor, outdoor and forecasts

lags	t+2	t+3	t+4	t+5	t+6	t+7	t+8	Total	Training mins.
2	0.379	0.462	0.524	0.574	0.624	0.677	0.732	0.529	28
4	**0.379**	**0.461**	**0.523**	**0.574**	**0.624**	**0.679**	**0.733**	**0.530**	**37**
6	0.382	0.466	0.527	0.577	0.627	0.680	0.734	0.532	45
8	0.384	0.466	0.527	0.577	0.626	0.680	0.735	0.533	55
10	0.386	0.468	0.529	0.580	0.629	0.681	0.736	0.535	65
16	0.389	0.470	0.528	0.577	0.624	0.676	0.727	0.533	92
20	0.389	0.470	0.526	0.574	0.622	0.672	0.725	0.532	114



5.2.2 Long-short Term Memory

Long-Short Term Memory network was tested to predict the indoor concentration of particulate matter at time t+1, t+2, …, t+8, using as input only historical data and then adding the forecast time series to the model. In the first case, the number of lag variable used was 2 and the output variables 8. For the second model including forecasts, the number of lag and lead variables used was 8.

The network takes as input the n lag values of $PM_{2.5}$ at time t, t-1...t-n. One model was implemented for each sensor with at least 500 observations.

Table 12. MSE and MAE obtained using LSTM with 8 time-steps , indoor-outdoor and forecasts time series as input

	lags	**t+2**	**t+3**	**t+4**	**t+5**	**t+6**	**t+7**	**t+8**	**Total**	**Training time**
MAE	8	0.36	0.44	0.49	0.54	0.58	0.62	0.66	0.49	36 hours (avg 9 min-utes per model)
MSE	8	1.26	1.58	1.80	2.01	2.19	2.37	2.56	1.86	

Table 13. MSE and MAE obtained using LSTM with t2o time-step and indoor-outdoor time series as input

	lags	**t+2**	**t+3**	**t+4**	**t+5**	**t+6**	**t+7**	**t+8**	**Total**	**Training time**
MAE	2	0.37	0.47	0.56	0.63	0.68	0.73	0.77	0.50	18 hrs. (avg 4.5 minutes per model)
MSE	2	1.67	2.20	2.61	2.95	3.19	3.41	3.66	2.58	

5.2.3 Comparison of results

Table 14. MSE comparison obtained using Random Forest and LSTM

Method	**lags**	**t+2**	**t+3**	**t+4**	**t+5**	**t+6**	**t+7**	**t+8**	**Total**	**Time**
RF Ind-Out	20	1.29	1.68	2.03	2.35	2.58	2.78	3.00	20.7	114 mins
RF ind-out-forecasts	4	1.82	2.23	2.48	2.70	2.95	3.21	3.49	2.52	37 mins
LSTM Ind-Out	2	1.67	2.20	2.61	2.95	3.19	3.41	3.66	2.58	18 hrs.
LSTM ind-out-forecasts	8	1.26	1.58	1.80	2.01	2.19	2.37	2.56	1.86	36 hrs.

The previous table shows the hourly mean squared error obtained using Random Forest with only historical indoor and outdoor $PM_{2.5}$ data, RF with outdoor forecasts, LSTM with only historical indoor and outdoor $PM_{2.5}$ data and LSTM with outdoor forecasts.

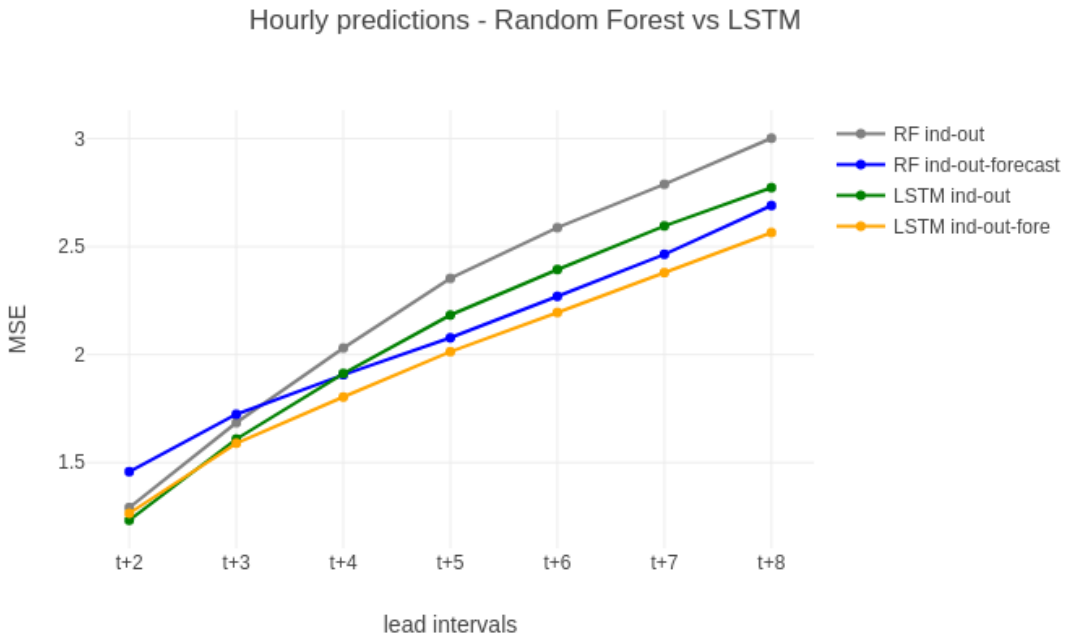


Figure 5. Long-term predictions using Random Forest with outdoor forecasts vs LSTM with only historical data.

From the previous graph we can see that best results for long term predictions were obtained using a LSTM Neural Network including the outdoor forecasts.

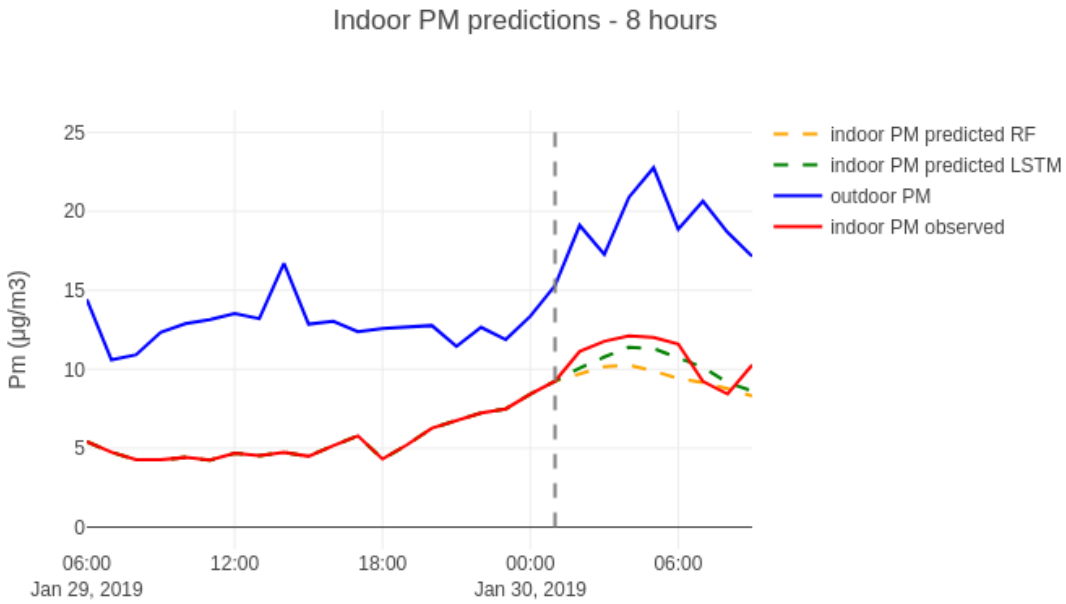


Figure 6. An example of predicting the indoor $PM_{2.5}$ *concentration 8 hours ahead using Random Forest.*

6 DISCUSSION

6.1 Managing expectations

Fine particulate matter ($PM_{2.5}$) levels were monitored in 120 Finnish buildings during the period 09.2014 – 02.2019. Each building had one or more sensor installed, for a total of 280 sensors. For each sensor monitored, short- and long-term indoor $PM_{2.5}$ forecasts were tested using three different methods: Autoregression, Random Forest and Long-Short Term Neural Network.

For fine particulate matter short-term predictions (+ 1 hour) the best results were obtained using LSTM with a univariate time series (historical indoor $PM_{2.5}$) with 2 lag variables. One model was used for each one of the sensors in the dataset with enough data available, for a total of 260 models. The average time to train one model was 3.7 minutes, much higher than using Random Forest. Both MAE and MSE for +1 hour predictions were in average 0.21 μg/m³ and 0.85 μg/m³.

For long-term predictions (up to +8 hours), the best results were obtained using LSTM with multivariate time series as input (historical indoor and outdoor $PM_{2.5}$ and outdoor forecasts). Results show as the MSE increased when predicting further intervals for all the methods. For long term predictions, the initial hypothesis that using outdoor $PM_{2.5}$ forecast would increase the accuracy of the model was confirmed. While the method wasn't the most accurate in predicting a short interval (+ 2 hours), results show that the long-term forecasts are more accurate when using the outdoor $PM_{2.5}$ forecasts in the model. It was demonstrated that using multivariate time series improved the long-term forecasts, but the 8 hours forecast error is still higher compared to the 1 hour forecast. Using a single model for each sensor gave better results than using a single model for all the sensors.

Fine particulate matter ($PM_{2.5}$) is considered one of the most harmful air pollutants. While a large proportion of the particles is originating from outdoor sources, people are mostly exposed while indoors. Predicting future trends of $PM_{2.5}$ concentrations using the previously describe methods could help building owners and operators developing better control strategies, and minimizing delays in responding to potential indoor air quality (IAQ) issues. High particulate matter levels might signify that the filter used in the ventilation system are no longer efficient and they need to be changed. Suggesting to occupants to keep the windows closed could reduce their risk to be exposed to harmful pollutants. The forecasts could also be used as input by automatic ventilation systems in modern buildings to adjust the amount of air introduced in the building.

6.2 Solution Limitations

The model implemented failed to forecast accurate $PM_{2.5}$ values in cases where fine particulate matter was generated by indoor sources. The highest indoor predictions error (MSE about 40µm/m³) was observed for a space with multiple events of $PM_{2.5}$ generated from indoor: predicting occupants' behavior is much more challenging.

The resampling technique used for the missing data points is reliable when only a few hours are missing, but in cases of longer periods of missing observations, the filling method adopted might have caused some problems to the models. Different resampling techniques could be tested (for instance KNN) (Junninen et al. , 2004).

The indoor values registred seem to be lower than expected. All the currently available particulate matter sensors underestimate the actual concentrations. This explain the low values showed in the descriptive statistics. This limitation doesn't affect the relative comparison between different methods and models.

Even though LSTM was the best method tested, the model requires much more time to be trained. This can be a problem when implementing the forecasts as a service in a production environment, due to the more resources needed to train the models in an acceptable time.

6.3 Error analysis

The model implemented failed to forecast accurate $PM_{2.5}$ values in cases where fine particulate matter was generated by indoor sources. The highest indoor predictions error (MSE about 40µm/m³) was observed for a space with multiple events of $PM_{2.5}$ generated from indoor: predicting occupants' behavior is much more challenging.

6.4 Recommendations for future research

Different types of LSTM Neural Network can be tested for long-term indoor fine particulate matter predictions: encoder-Decoder LSTM, grid LSTM, CNN-LSTM and ConvLSTM could improve the results and can be used for 24 hours predictions. Temporal pattern attention for multivariate time series forecasting could also be tested (Shih et al. 2018).

Building models using group of sensors in the same building or with similar behavior could lead to a faster training time. Including variables like ventilation rate, the rooms square meters and the location of the building could help in reducing the error for long-term forecasts.

REFERENCES

Ahn, Jaehyun & Shin, Dongil & Kim, Kyuho & Yang, Jihoon, 2017. *Indoor Air Quality Analysis Using Deep Learning with Sensor Data*. Sensors. 17. 2476. 10.3390/s17112476.

Allen, J. G., MacNaughton, P., Satish, U., Santanam, S., Vallarino, J., & Spengler, J. D. 2016. *Associations of Cognitive Function Scores with Carbon Dioxide, Ventilation, and Volatile Organic Compound Exposures in Office Workers: A Controlled Exposure Study of Green and Conventional Office Environments*. Environmental Health Perspectives, 124(6), 805–812.

Allen, R.; Larson, T.; Sheppard, L.; Wallace, L.; Liu, L.J.S, 2003, *Use of Real-Time Light Scattering Data to Estimate the Contribution of Infiltrated and Indoor-Generated Particles to Indoor Air. Environ*. Sci. Technol. 37, 3484–3492.

Allen, Ryan & Larson, Timothy & Sheppard, Lianne & Wallace, Lance & J Sally Liu, L. , 2003. *Use of Real-Time Light Scattering Data To Estimate the Contribution of Infiltrated and Indoor-Generated Particles to Indoor Air. Environmental science & technology*. 37. 3484-92. 10.1021/es021007e.

Breezometer, Air Quality Index, available from *https://breezometer.com/air-quality-map* [accessed 21 Mar, 2019].

Brownlee, J, 2019, Gentle Introduction to the Adam Optimization Algorithm for Deep Learning, https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/, 2017. [Accessed: 21.3.2019]

Cho, K, Van Merriënboer, B, Gulcehre, C, Bahdanau, D, Bougares, F, Schwenk, H, Bengio, Y 2014. *Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar; pp. 1724–1734.

Chu & , Kiu-fung & , Truman & , 朱喬鋒, 2003. *Guidance Notes for the Management of Indoor Air Quality in Offices and Public Places*. http://sunzi.lib.hku.hk/hkuto/record/B3125486X [accessed 21 Mar, 2019].

Coco Liu, Jia & Mickley, Loretta & Sulprizio, Melissa & Dominici, Francesca & Yue, Xu & Ebisu, Keita & Brooke Anderson, Georgiana & F. A. Khan, Rafi & Bravo, Mercedes & L. Bell, Michelle, 2016. *Particulate Air Pollution from Wildfires in the Western US under Climate Change*. Climatic Change. 138. 10.1007/s10584-016-1762-6.

Directive 2008/50/EC of the European Parliament and of the Council of 21May 2008 on ambient air quality and cleaner air for Europe OJ L 152, 11.6.2008, p. 1–44.
EMC Data Science Global Hackathon (Air Quality Prediction), available from https://www.kaggle.com/c/dsg-hackathon, [accessed 20 apr, 2019].
Finnish Meteorological Institute, Fine Particles data, available from https://ilmatieteenlaitos.fi/pienhiukkaset-ilmansaasteena, [accessed 21 Mar, 2019].
Guo, Tian & Lin, Tao & Lu, Yao, 2018. *An interpretable LSTM neural network for autoregressive exogenous model.*
Hänninen, Otto & Knol, Anne & Jantunen, Matti & Lim, Tek-Ang & Conrad, André & Rappolder, Marianne & Carrer, Paolo & Fanetti, AC & Kim, Rokho & Buekers, Jurgen & Torfs, Rudi & Iavarone, Ivano & Claßen, Thomas & Hornberg, Claudia & Mekel, Odile, 2014. *Environmental Burden of Disease in Europe: Assessing Nine Risk Factors in Six Countries. Environmental Health Perspectives*. 10.1289/ehp.1206154.
Health and Environmental Effects of Particulate Matter (PM), available from https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm [accessed 21 Mar, 2019].
Hochreiter, S.; Schmidhuber, J,1997 Long short-term memory. Neural Comput. 9, 1735–1780. [CrossRef] [PubMed].
Hoppe, P, Martinac, *I, 1998; Indoor climate and air quality-Review of current and future topics in the field of ISB study group 10*. Int JBiometeorol42(1):1-7.
Junninen, Heikki & Niska, Harri & Tuppurainen, Kari & Ruuskanen, Juhani & Kolehmainen, Mikko, 2004. *Methods for imputation of missing values in air quality data sets. Atmospheric Environment. 38. 2895-2907. 10.1016/j.atmosenv.2004.02.026.*
Lai, H.K.; Bayer-Oglesby, L.; Colvile, R.; Götschi, T.; Jantunen, M.J.; Künzli, N.; Kulinskaya, E.; Schweizer, C.; Nieuwenhuijsen, M.J. *Determinants of indoor air concentrations of $PM_{2.5}$, black smoke and NO_2 in six European cities (EXPOLIS study).* Atmos. Environ. 2006, 40, 1299–1313.
Li, Xiang & Peng, Ling & Yao, Xiaojing & Cui, Shaolong & Hu, Yuan & You, Chengzeng & Chi, Tianhe, 2017. *Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution* (Barking, Essex : 1987). 231. 997-1004. 10.1016/j.envpol.2017.08.114.

Makridakis, Spyros & Spiliotis, Evangelos & Assimakopoulos, Vassilis, 2018. *Statistical and Machine Learning forecasting methods: Concerns and ways forward.* PLoS ONE. 13. 10.1371/journal.pone.0194889.

Müller, A, Guido,S, 2016., *Introduction to Machine Learning with Python: A Guide for Data Scientists.*

Namieśnik, Jacek & Górecki, Tadeusz & Kozdroń-Zabiega ła, Bożena & Łukasiak, Jerzy, 1992. *Indoor air quality (IAQ), pollutants, their sources and concentration levels. Building and Environment.* 27. 339-356. 10.1016/0360-1323(92)90034-M.

Olah, C., *Understanding lstm networks*, http://colah.github.io/ posts/2015-08-Understanding-LSTMs/, 2015. [Accessed: 21.3.2019].

Omidvarborna et al. 2015,*Recent studies on soot modeling for diesel combustion. Renewable and Sustainable Energy Reviews.* 48: 635–647. doi:10.1016/j.rser.2015.04.019.

Rodriguez-Galiano, 2015, Modelling interannual variation in the spring and autumn land surface phenology of the European forest - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-flowchart-of-random-forest-RF-for-regression-adapted-from-Rodriguez-Galiano-et_fig3_303835073 [accessed 21 Mar, 2019].

Seppänen, O., Fisk, W. 2006. *Some quantitative relations between indoor environmental quality and work performance or health.* HVAC&R Research 12(4): 957-973.

Shih, Shun-Yao & Sun, Fan-Keng & Lee, Hung-yi. 2018. *Temporal Pattern Attention for Multivariate Time Series Forecasting.*

SMU faculty, 2019, Augmented Dickey-Fuller Unit Root Tests, http://faculty.smu.edu/tfomby/eco6375/BJ%20Notes/ADF%20Notes.pdf [accessed 21 Mar, 2019].

SPSS Tutorials: Pearson Correlation, available from https://libguides.library.kent.edu/SPSS/PearsonCorr [accessed 21 Mar, 2019].

University of Vaasa, Testing for cointegration, 2019 http://lipas.uwasa.fi/~sjp/Teaching/Afts/Lectures/etsc32.pdf [accessed 21 Mar, 2019].

US EPA , Particulate Matter (PM) Pollution, available from https://www.epa.gov/pm-pollution [accessed 21 Mar, 2019].

WHO Air quality guidelines for particulate matter,ozone, nitrogen,dioxide and sulfur dioxide, Global update 2005 Summary of risk assessment.