

**Sensing Spaces: An Online
Spatio-Temporal Model to Predict
Air Pollution With Mobile and
Stationary Data**

Ryan Egan

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2019

Abstract

As air pollution levels rise in urban areas and in developing countries, methods for predicting pollution levels in both the spatial and temporal aspect are necessary. Air pollution is a leading cause of cardiovascular diseases, strokes, cancer and asthma. Proposed is a new model to predict spatio-temporal air pollution, specifically Particulate Matter (PM) levels. Included are a state-of-the-art Convolutional Neural Network (CNN) to spatially interpolate PM values, and a novel method of building a model that predicts temporally using a single Artificial Neural Network (ANN) which uses land usage, and road type as input features. A combination of the two are used to predict spatio-temporally. Both mobile data and stationary data are used in training of the model. The model uses online learning to improve adaptability as well as ensembles to improve generalization. Experiments on different grid size, online update rules, larger datasets, and mobile data are included.

This work shows that ANNs can effectively be used to predict PM values temporally, spatially, and spatio-temporally. The proposed model ensemble was the only model that was able to perform better than the baseline in temporal predictions with an MAE of 0.124 on the 2018 dataset. The spatial CNN introduced in this work improved upon previous best MAPE spatial interpolation by 3 percentage points and the previous best CNN model by 17 percentage points with an MAPE of 39.10%. The spatio-temporal predictions on the 2018 datasets also had positive results. For the 2018-A dataset, the proposed model predicted better than the previous best with an MAE of 1.50. On the 2018-B dataset, the proposed model's MAE was 1.16 which was also better than the previous best MAE of 1.21.

A new dataset was also collected for central Edinburgh that used ten stationary sensors and up to four concurrent mobile sensors. The proposed model set a baseline MAPE of 17.20% for temporal predictions and 56.89% for spatio-temporal predictions.

Acknowledgements

I would like to D.K. Arvind for proposing this topic as well as providing all resources necessary. He was a great help each week in our weekly meetings offering assistance and guidance.

Additionally, thank you to Zoë Petard for the consistent assistance on direction and implementation. Andrew Bates for help setting up and deploying the stationary and mobile sensors. Sitthinut Kumpalanuwat for assistance collecting mobile data and building stationary sensors.

Contents

1	Introduction	1
1.1	Major Contributions	3
2	Background	4
2.1	Particulate Matter	4
2.2	Data Collection	5
2.3	Open Street Map	7
2.4	Online Learning	7
2.5	Models	8
2.5.1	Statistical Interpolation	8
2.5.2	Passive Aggressive Regressor	8
2.5.3	Artificial Neural Network	9
2.5.4	Convolutional Neural Network	9
2.5.5	Long Short-term Memory	9
3	Related Work	10
3.1	Least Polluted Paths for Transit	10
3.2	Sensing Spaces: Improving Spatial and Temporal Prediction of Air Pollution Using Online Learning Algorithms	10
3.3	High Resolution Spatial Predictions of Air Pollution Levels in Edinburgh	11
3.4	Sensing Spaces: An online spatio-temporal model for PM _{2.5} predictions using a combination of stationary and mobile airborne particulate sensors	11
3.5	Fine-grained Spatiotemporal prediction of air pollutants in New Delhi, india	12
3.6	Spatio-Temporal Air Pollution Prediction Using Deterministic Models	12
3.7	Spatio-Temporal Air Pollution Prediction Using ANNs	13

4 Methodology	15
4.1 Dataset and Test Area	15
4.1.1 Calibration	17
4.1.2 Preprocessing Data	18
4.2 Proposed Methods	19
4.2.1 Many Model Method	19
4.2.2 Master Model Method	20
4.2.3 Ensemble	20
4.2.4 Bias Correction	21
4.3 Performance Metrics	21
4.3.1 Validation	22
5 Results and Discussion	25
5.1 Hyperparameter Search	25
5.2 Feature Selection	26
5.3 Spatial CNN	26
5.4 Update Decision	27
5.5 Temporal Prediction	29
5.6 Spatio-temporal Prediction	31
5.6.1 Master Model Comparison to PAR Grid	31
5.6.2 Grid Size	35
5.6.3 Utility of Mobile Data For Training	36
5.6.4 2019 Dataset Results	37
5.7 Results Summary	38
6 Conclusions	39
Bibliography	41
A Additional Dataset Information	47
A.1 Stationary Sensor Trends	47
A.2 Calibration Factors	47
A.3 2019 Deploy Dates	48
A.4 50 × 50 Grid	49
A.5 Mobile Vs. Stationary Data	49
B Update Decision	53

C Temporal Prediction Without Mobile Data	54
D 2019 Temporal Results	55
E Information for Data Collection Participants	56

Chapter 1

Introduction

Around the world, especially in urban areas, air pollution is a leading cause of cancers, heart disease, stroke, asthma and other forms of respiratory diseases [22] [33]. A major source for the dangers of air pollution comes from Particulate Matter (PM) leading to the importance of spatio-temporal PM predictions. Accurate spatio-temporal models are rare and usually expensive, this project aims to solve both issues.

According to the World Health Organization (WHO) air pollution is the world's largest single environmental risk [34]. In 2015, an estimated 3.2 million deaths and 5.1 trillion USD was spent in the 41 countries making up Organisation for Economic Co-operation and Development (OECD). OECD mainly comprised of Western Europe and North America and BRIICS (Brazil, Russia, Indonesia, India, China, and South Africa) and account for an estimated 70% of worldwide deaths from air pollution [37]. For such a large burden of cost for healthcare it is in the best interest of governments and individuals to reduce air pollution.

PM can be anthropogenic and are normally emitted from factories, power plants, refuse incinerators, motor vehicles, construction activity, and fires [22]. Air pollution increases global heating and thus increases the effects of climate change [48]. In return, air pollution increases as the global temperature rises. Since the effects of air pollution are felt relatively quicker (i.e. asthma, heart attack) than climate change (i.e. change of global temperature) there have been more measures by governments and the general populace to reduce air pollution; for example, the shift from fossil fuels and coal to wind and solar energy [48].

The dangers and prevalence of air pollution make timely predictions of high pollution areas important. A person at risk of having an asthma attack due to PM concentrations would be interested in seeing the cleanest route to a destination. This could

also apply to someone that wishes to lead a healthy life as PM exposure lowers the expected lifespan [33]. The most practical application of predicting future air pollution levels would be determining clean, safe routes either for walking, biking, or even driving. Such an application can be used in any route planning application from Google, Apple, or Bing.

To further study the effects and spread of air pollution, the Centre for Speckled Computing at the University of Edinburgh developed low cost air pollution sensors [5]. Since their inception these sensors have been used for a number of applications, including a study which has been run multiple times: trying to understand how air pollution moves spatially as well as dissipates temporally [35] [3] [6]. Data from the sensors has been used in the creation of a framework for clinical studies on the effect of PM on the respiratory system as well as projects creating algorithms and applications for cleaner transportation routes [4] [21] [41] [26].

Previous work in predicting air pollution spatio-temporally have mostly used statistical regression models to achieve best results [6] [50] [54]. More recent models have started to use Artificial Neural Networks (ANNs). ANNs that include more meteorological data, land usage, and road types perform the best [1] [49]. Recent work has shown that a Convolutional Neural Network (CNN) can produce similar results to the best deterministic spatial predictors [35] [49].

This project introduces a novel approach of first using an ANN to predict temporally then spatially interpolating over an area using a CNN. Using one ANN for all locations within the test area was used to reduce data sparseness issues faced in the past. Interpolation is necessary because it is not feasible to have sensors covering the entirety of the test area at all times. The use of mobile data in training for a spatio-temporal problem is a novel concept as most projects do not include mobile data for training and use it for validation. Online learning in the past has been shown to allow models to react to volatile data more accurately, so this project will also apply online learning [3] [6].

Unique to this work is the use of ensembles of online and pretrained models. The hypothesis is that the online portion will be able to predict spikes in PM while the pretrained portion will predict the cyclical parts of the dataset. Ensembles have been proven to generalize better in ANNs and other machine learning models [52] [16]. Even if the ensemble does not have the intended effect of following cyclical data with the pretrained and spikes with the online model, it should have better results.

For some experiments, this work relied on a dataset collected by Petard in 2018

in order to compare to some existing benchmarks [35] [6]. For other experiments, a new dataset was collected by the author with the help of colleagues at the University of Edinburgh, Sitthinut Kumpalanuwat, DK Arvind, and Andrew Bates. The new dataset has a larger area of focus, longer collection time, and more stationary and mobile sensors deployed simultaneously.

While this work does not make significant use of the 2019 dataset due to time restrictions the dataset collected for this work will be able to be used for the foreseeable future. The data more than doubles the mobile and stationary data collected in Central Edinburgh. Future work can compare the datasets, be used for projects similar to this, attempt transfer learning across datasets, and be used to generate adversarial data.

The structure of this work is as follows: The next chapter is a background chapter describing PM in more detail, tools used, and the components of the model. The third chapter will describe in detail related work. The fourth chapter will describe the datasets, methodology, and model structure. The fifth chapter will show results of the experiments carried out as well as discussion of results.

1.1 Major Contributions

Main contributions from this project are as follows:

- A novel online ensemble approach to predicting PM values spatio-temporally which uses only one model for temporal predictions and one model for spatial interpolation.
- A new best MAE for temporal predictions for PM_{2.5} on the 2018 dataset using the proposed model, 0.124.
- A new best MAPE for spatial interpolation for PM_{2.5} on the 2018 dataset using the proposed model, 39.10%.
- Of the two existing 2018 datasets, proposed model performed better than existing benchmark for one and similar to benchmark on the other using the proposed model, 1.50 and 1.16 for datasets 2018-A and 2018-B.
- A new dataset using ten stationary and up to four concurrent mobile sensors over 22 days with temporal and spatio-temporal benchmarks set by the proposed model.

Chapter 2

Background

2.1 Particulate Matter

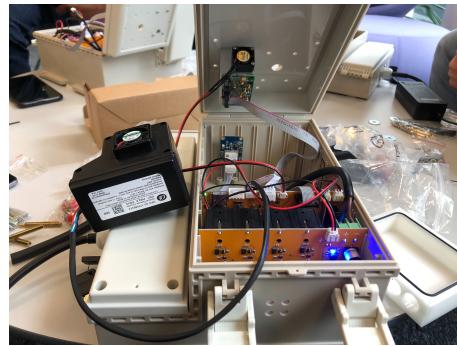
PM are defined as “small solid or liquid particles suspended in the atmosphere which can be emitted from natural, anthropogenic sources, or can be formed through chemical reactions in the atmosphere [36].” PM can be classified as coarse and fine particles. The border is often considered to be at $2.5\mu m$, fine particles being under $2.5\mu m$ and coarse particles bigger than $2.5\mu m$. The two types generally have a different origin. The larger particles come from dust from road and industry, and contain earth crust. The smaller particles come from gasses such as car exhaust, barbecue or bonfire smoke, cigarette smoke, as well as non-smokeless fuels for heating [33]. This work focuses on $PM_{2.5}$, the fine particles, as their origin is more traceable by the information available to this project. The density of PM in the air is measured in micrograms per cubic meter ($\mu g/m^3$). This work focuses on $PM_{2.5}$ as PM_{10} is often more volatile and therefore less likely for a model to predict. PM_1 is a further classification of the fine particles and thus a subset of $PM_{2.5}$. The correlation between $PM_{2.5}$ and PM_1 is 92% so if $PM_{2.5}$ can be predicted so can PM_1 [29].

$PM_{2.5}$ is a pollutant known to be controlled by both local conditions and long range transport of PM [44]. Since PM values are known to come more often from man-made sources, land usage and road type do help to predict PM values [35]. Long range transport of PM raises an issue for this project because PM can travel from outside of the test area to raise the air pollution. There would be no explanation for why PM increased in our dataset as the dataset contains no meteorological data such as wind speed and wind direction.

2.2 Data Collection

To gather data, two types of sensors were developed by the Centre for Speckled Computing at the University of Edinburgh [5]. The two sensors developed are the Airspeck-S and Airspeck-P. The Airspeck-S is used as a stationary monitor whereas the Airspeck-P is designed to be carried around. The data retrieved from the Airspeck-S and the sensor itself will be referred to as the “stationary data” and “stationary sensor” respectively for the rest of this paper. Likewise the data from the Airspeck-P will be referred to as “mobile data” and the Airspeck-P sensor will be referred to as the “mobile sensor”.

The Airspeck-P and Airspeck-S sensors are examples of a trend to create cheaper air pollution and gas sensors within the research community. Other works include bicycle-mounted sensors to observe traffic-related air pollution, personal exposure estimation, and indoor air pollution estimator [43] [53] [13]. These sensors can cost as much as three orders of magnitude less than high accuracy sensors [31]. The hope is that with cheaper sensors the effects and characteristics of air pollution can be better monitored and analyzed. With the price being significantly cheaper they are able to be deployed en masse leading to more data which drives innovation in predicting air pollution.



(a) The inside of an Airspeck-S as it is being assembled.



(b) A photo of an Airspeck-S being assembled.



(a) A photo of an Airspeck-S being deployed.



(b) A deployed Airspeck-S in the Meadows.

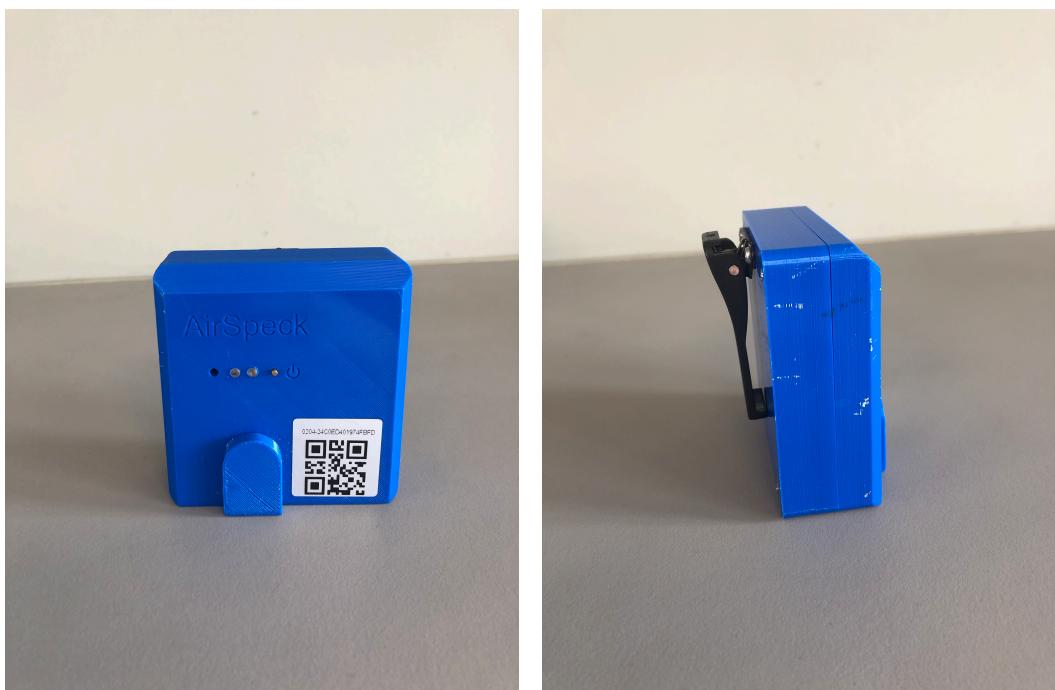


Figure 2.3: Airspeck-P

The sensors use an optical particle counter to classify the PM that enters the sensor to the appropriate size. In total there are 16 bins to which PM can be classified ranging from $0.38\mu\text{m}$ and $17\mu\text{m}$. Based on the amount in each bin the PM values for PM_1 ,

$\text{PM}_{2.5}$ and PM_{10} can be determined. This results in a relative PM count that is simply based on the previous PM count and how much it has changed. This means that the absolute value is not known until the sensor is calibrated using a ground truth provided by an already calibrated sensor.

For the new dataset gathered in 2019 the author spent approximately 25 hours building, deploying, maintaining and calibrating the stationary sensors. 19 hours were spent by the author collecting data using a mobile sensor walking around central Edinburgh. Contributions from colleagues, with extra collection from DK Arvind and Sittinut Kumpalanuwat added another 45 hours. The dataset will be discussed further in Section 4.1: Dataset and Test Area.

2.3 Open Street Map

OpenStreetMap (OSM) was used to collect data for the road type and land usage [32]. OSM is an open source database that contains a plethora of information about roads, land, types of neighborhoods, and more. Since it is open source and extensive it does include extra superfluous information which had to be stripped away to access the data useful to this project such as road type and land usage.

The use of land usage and road types are necessary for intelligent interpolation. Both have been shown to increase interpolation accuracy [35] [1]. Land usage can be broken down into three types: green space, residential, commercial, and industrial. Air pollution rates increase around industrial space, decrease near green space, while residential and commercial have no effect on air pollution [15]. There are five road types: primary, secondary, unclassified, residential/service, pedestrian/cycle/noroad. The main classification for road type is that primary is correlated to highest air pollution, then secondary, followed by the other three [35].

2.4 Online Learning

In this work a timestep will refer to each time a full spatial-temporal prediction is made. While a datapoint will refer to one cell in the grid of a spatio-temporal prediction. For example, if there is an $n \times n$ grid of values at timestep t then there will be a total of $n \times n \times T$ datapoints, where T is the total time in the dataset. Therefore, the example dataset will have T timesteps.

Online learning is shown to learn and adapt to volatile data better than a pre-trained model in a time-series setting [3]. Online learning works by continually updating the model after each new data point. Constant updates make it able to adjust parameters to better predict spikes in data. Online learning has been used in predicting spikes in neural cultures, nonstationary time-series problems, and can produce results in a time-series setting as good as autoregressive moving average (ARMA) [18] [46] [2].

The difference between online learning and the more common batch training is that batch training takes a batch of samples and updates the weights on the cumulative error. Going through one epoch using batch training is one update per epoch whereas online is k updates per epoch, where k is the number of timesteps. Reduction in the cost function per update for online learning is often high since it updates after each datapoint, which should help identify spikes or “noise” in the data.

2.5 Models

A number of models are mentioned and used in this work. This section will give a brief overview of each model.

2.5.1 Statistical Interpolation

The only statistical interpolation method used in this work is ordinary kriging (OK). Kriging are a group of spatial estimators in which spatial predictions are made when a variogram is known. In other words it uses nearby values to predict a different location’s value. The goal of OK is to minimize the error variance which should lead to accurate estimations [28]. OK also makes no assumptions about the underlying distribution of values.

2.5.2 Passive Aggressive Regressor

Passive Aggressive Regressors (PAR) are a part of a group of online learning algorithms created by Crammer et al. [9]. PAR has been shown to perform better than other online algorithms such as online perceptrons and MIRA [3] [9]. PAR makes predictions similarly to a linear regression where each variable has an associated weight. Unique to PAR is how it learns to update weights after new data. When the model prediction is close to the real value the model will not update, but if the model is wrong

it will update its weights. A slack variable is used to determine the threshold for when to update the weights.

2.5.3 Artificial Neural Network

Artificial Neural Networks (ANN) are a way to make predictions or classifications for almost all types of problems and domains. An ANN is a collection of neurons and connections which take in a vector of numerical input and output a vector of values. Depending on the task these numbers can represent anything from values of a pixel to numerical features, and from classification features to probability distributions. In the context of this work the input will be a vector of inputs containing both numerical values as well as binary classification features. The output will be a single value which represents a future PM value. ANNs are adept at finding patterns in large datasets, as well as estimating non-linear functions, which are both useful to this task.

An ANN was used to make temporal predictions in this project. ANN was selected because of its ability to approximate non-linear functions, which temporal PM values follow.

2.5.4 Convolutional Neural Network

Convolutional Neural Networks (CNN) are a type of ANN. The major distinction of a CNN, the convolutional layers are the reason they are good at image recognition. Convolutions allow the CNN to look at small sections of an image and put all the small parts together to form context. CNNs are mostly used for image classification ranging from tasks of identifying numbers to help autonomous cars recognize objects [27]. As the map of central Edinburgh can be interpreted as an image at any given time, a CNN can be applied to interpolate spatially.

2.5.5 Long Short-term Memory

Another type of ANN is the Long Short-term Memory (LSTM). LSTMs unique functionality comes from its ability to use previous hidden layers and output as input. The use of previous data and predictions make LSTMs appropriate for time-series data. LSTMs are only briefly used in this work for reasons mentioned in Section 5.5, Temporal Prediction.

Chapter 3

Related Work

3.1 Least Polluted Paths for Transit

A major application of the proposed model in this work is to be able to map clean routes to use for walking or biking. Such an application is being developed in Edinburgh at the time of writing by Sitthinut Kumpalanuwat titled “The Healthy City Tour Route Planner in the Area of Central Edinburgh” [26]. The purpose of his work is to find the route with the least air pollution given a user’s preferences. The user is allowed to input whether they are walking or biking, how long they intend to walk or bike, a threshold for the number of turns, and amount of hills. His algorithm uses the model in this work to predict future air pollution levels to determine where to direct the user.

Another project being developed simultaneously similar to Kumpalanuwat’s work is Enhao Sun’s “Least Polluted Route: Algorithms and Application” [41]. The objective of Sun’s work is to find the least polluted routes from one point to another. Sun also developed a mobile app to make the algorithm accessible to users. Sun focused on a larger area than Kumpalanuwat. Sun also uses the proposed model in this work to provide predicted PM levels to route through.

3.2 Sensing Spaces: Improving Spatial and Temporal Prediction of Air Pollution Using Online Learning Algorithms

The work done by Angelinin in 2018 provided the first look into using online learning for temporal predictions of PM [3]. Angelinin’s work showed that online methods

proved better predictions on volatile data than offline models. The study showed that a Passive Aggressive Regressor (PAR) was the best online model for predicting PM temporally. Angelinin also provided a less in depth exploration of spatial predictions than Petard's work in the next section.

3.3 High Resolution Spatial Predictions of Air Pollution Levels in Edinburgh

This study was conducted by Petard in 2018 using a CNN to make accurate spatial prediction of air pollution in central Edinburgh [35]. While the study found that the CNN, when compared to statistical models, was slightly less accurate using quantitative results it made more sense when looking at a map of predictions. For example it predicted higher pollution along roads than on green space far from roads. Petard's model used granular land usage, road types, and humidity as the only four features. Petard broke central Edinburgh into a grid of 20x20 consisting of about a 0.96km by 0.97km space. This makes each cell approximately 50x50 meters. The dataset created in Petard's study was also instrumental in creating baselines for previous work and this proposed model.

3.4 Sensing Spaces: An online spatio-temporal model for PM_{2.5} predictions using a combination of stationary and mobile airborne particulate sensors

Catarino's work was finished just before this project started. His work focused on the same problem described here: to predict PM online spatio-temporally using mobile and stationary data. However, Catarino faced data scarcity which this project sought to improve upon in a number of ways. Catarino's approach was to use a grid and use a PAR model within each of those cells to predict temporally and then trained on data that was interpolated using OK. This allows for the usage of mobile data in training the model as well as predictions which was the first to use mobile data in training. Catarino also introduced a bias correction to stop the model from underpredicting. Catarino's work sets a baseline for this work to compare against as it was testing on Petard's 2018 dataset. A problem Catarino faced was the sparseness of data within each cell seen

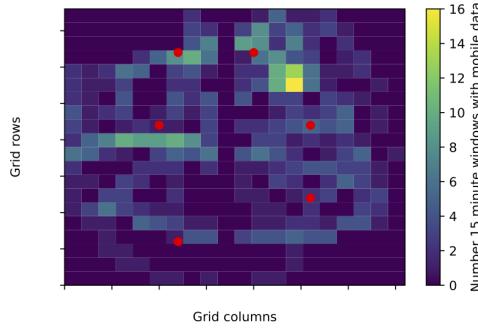


Figure 3.1: 2D visualization of mobile data collected in 2018 dataset [6]. The area represents the 0.97km by 0.98km in central Edinburgh captured by Petard.

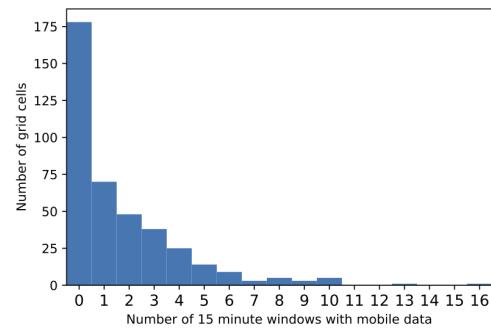


Figure 3.2: Histogram of grid cells by amount of mobile data collected in 2018 dataset [6].

in **Figure 3.1** and **Figure 3.2**. A PAR model like any other machine learning model requires a large amount of data. This is one problem my proposed model looks to address along with adding ensembling to generalize better and use a more accurate spatial interpolation method.

3.5 Fine-grained Spatiotemporal prediction of air pollutants in New Delhi, india

M Thirukkumaran’s work is similar to the work undertaken in this thesis, but focuses on spatio-temporal predictions in New Delhi, India [42]. Created in the work is a novel way to retrieve average traffic density for the time of day and day of the week. Thirukkumaran’s work made one minute predictions using LSTMs and classical machine learning approaches in a continuous setting, trying to predict air pollution on a mobile sensor regardless of where the mobile sensor is moving.

3.6 Spatio-Temporal Air Pollution Prediction Using Deterministic Models

Originally spatio-temporal predictions were deterministic models that heavily relied on aerodynamic theory, and the theoretical movement and dissipation of PM. Predic-

tions are made using numerical methods with simulations and in-time predictions that calculate the dynamics of the environment [50] [54]. Some common models following this approach are the Community Multi-scale Air Quality (CMAQ), the WRFChem model, and the Nested Air Quality Prediction Modeling System (NAQPMS) [7] [39] [47]. The issue with these models is that they require features and data that is often unobtainable, such as emissions from construction dust, fugitive dust, soil dust, transportation and biomass burning [45]. They are also generally inaccurate as they require estimation for some elements, such as regression coefficients [11].

3.7 Spatio-Temporal Air Pollution Prediction Using ANNs

There is a fairly recent interest in using ANNs to predict air pollution in the past few years. Previously a top choice in predicting air pollution spatio-temporally was to make a land regression maps with statistical models [1]. The reason for choosing this method over an interpolation method is because of the sparseness and variability of the spatial data collected. Adams *et al.* created an ANN model to make spatial predictions in 2016 using mostly mobile data to allow for interpolation [1]. The model used included land usage as that is an important factor in predicting PM values, a feature included in this proposed work. The model also used wind speed, wind direction, traffic congestion, and altitude which are all valuable features when predicting PM values [1].

Wen *et al.* created a novel approach to PM prediction in China in early 2019 [49]. They created a spatio-temporal model named convolutional long short-term memory neural network extended (C-LSTME). This was essentially a combination of an LSTM to predict temporally, and a CNN, to predict spatially. The data used came from 1233 air quality monitoring stations across all of China for a full year. The stations were able to obtain meteorological data, planetary boundary layer height, and aerosol optical depth. All of these features have been proven to improve PM predictions [12] [39] [24]. The stations took measurements hourly and therefore the model was used for hourly predictions. Prediction intervals ranged from one hour to twelve hours, getting less accurate the farther into the future the prediction was made.

The model used nearest neighbors to add input features to the model from nearby stations. So not only will the model receive input information for the current location being evaluated but the closest other stations information as well. The proposed model in this work looked to emulate using current events from more than one source when

predicting.

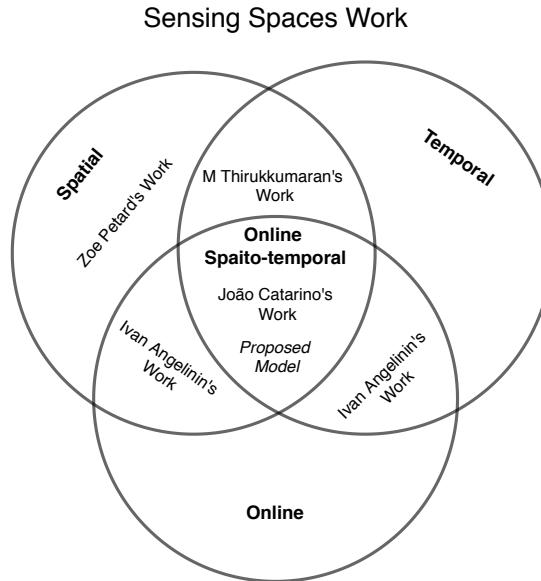


Figure 3.3: A description of the work done within the Centre for Speckled Computing by past students. Credit João Catarino for original image [6]

<i>Feature</i>	<i>Proposed Model</i>	<i>Thirukkumaran</i>	<i>Catarino</i>
Spatio-temporal	✓	✓	✓
Online	✓		✓
Neural Network	✓	✓	
Land Usage	✓		
Road Type	✓		
Traffic Congestion		✓	
Bias Correction	✓		✓
Ensemble	✓	✓	

Table 3.1: A further distinction between the work proposed in this work, Catarino and Thirukkumaran [6] [42]

Chapter 4

Methodology

The purpose of this research is to use an online learning method to predict air pollution levels in the near-term. Online learning has been shown to predict rapidly changing as well as more volatile air pollution levels temporally [3]. It has also been shown to help make more accurate air pollution predictions spatio-temporally [6].

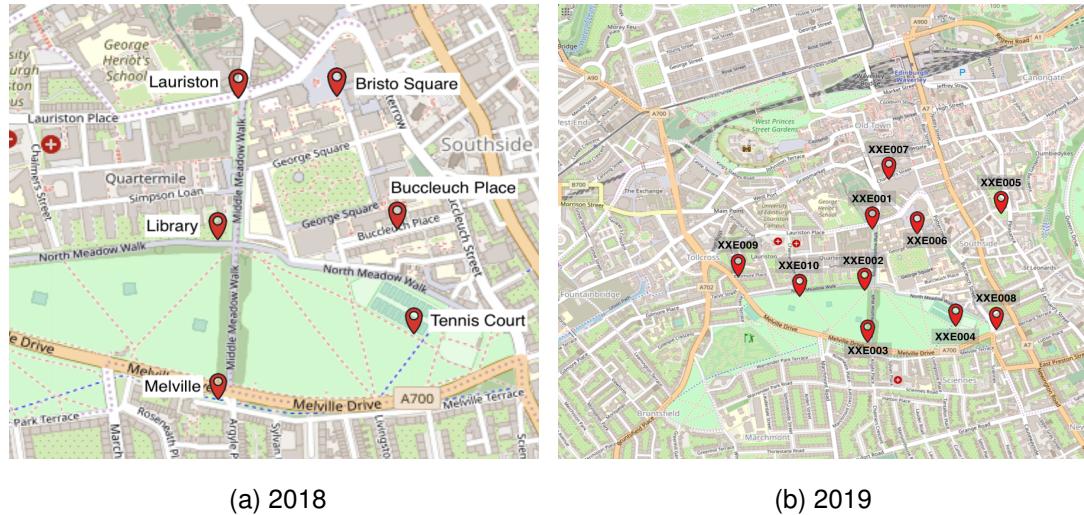
Catarino used an individual PAR to make predictions on each grid cell which allows two natural options for improvement: using ANNs in place of PAR, or to use a continuous model that makes use of just one ANN [6]. This work will explore both options with an emphasis on the latter as that is unexplored using online learning and an ensemble. Since online models perform better in volatile environments and batch trained models perform better in cyclical environments, this work proposes to use an ensemble of batch and online models.

This chapter will explain the dataset including the preprocessing that took place, the features in the dataset, the test area in central Edinburgh and calibration necessary for the dataset to be used. This chapter will also propose a novel method of predicting PM values spatio-temporally. Lastly, this chapter explains the performance metrics used as well as the method for comparing predictions to the ground truth.

4.1 Dataset and Test Area

The data was collected using both Airspeck-S and Airspeck-P over two summers. Data was collected from June 28 to August 8 of 2018 by Petard and June 28 to July 29 of 2019 by the author. This is the second work to use mobile data for spatio-temporal prediction within the the Centre for Speckled Computing. In the past that data was solely used for validation of spatial predictions. In this work both Airspeck-S and

Airspeck-P will be used for both training and validation.



In 2018 there were six Airspeck-S setup in Central Edinburgh as seen in **Figure 4.1a**. **Figure 4.1b** shows where the ten Airspeck-S were setup in Central Edinburgh for 2019 while **Table A.3** shows the dates each sensor was deployed. Example Airspeck-P collection maps can be seen in **Figure 4.2a** for 2018 and **Figure 4.2b** for 2019.

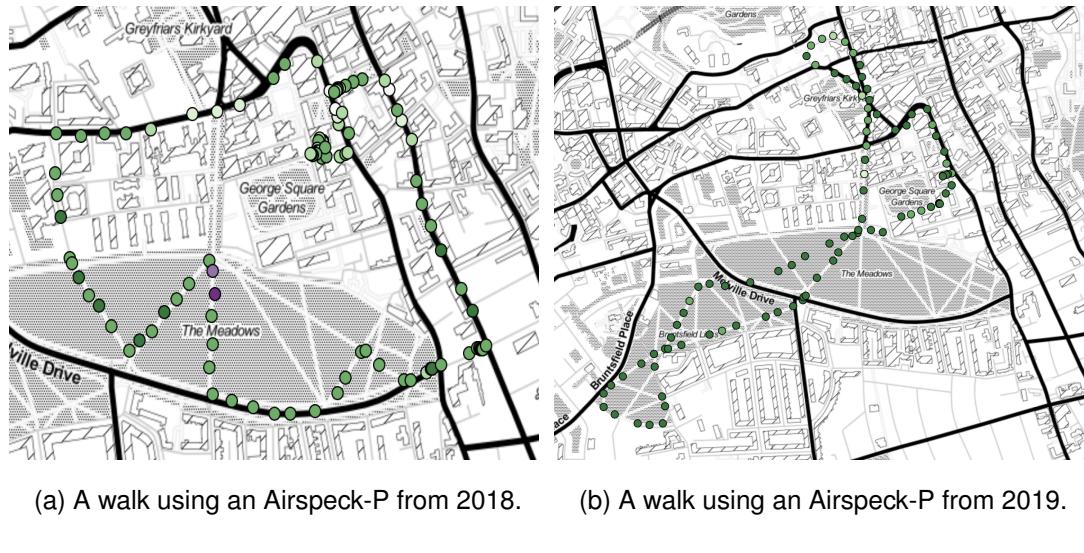


Figure 4.2: Each dot represents a point when the sensor retrieved data. The green dots are clean and the purple have more air pollution.

The test area for the 2018 data is a rectangle of size 0.97km (latitude) by 0.96km (longitude) in the central area of Edinburgh, for a total area of 0.93km^2 . The area contains grassland, major roads, minor roads, walking paths, and a University. Data collected in 2019 was collected from a larger area with more sensors. The 2019 area is 1.35km (latitude) by 1.66km (longitude), resulting in a total area of 2.24km^2 .

For most experiments a 20×20 grid was used to divide the area into cells as seen in **Figure 4.3a** and **Figure 4.3b**, when a 20×20 grid is not used it is replaced by a 50×50 grid. The size of each cell in the 2018 dataset is $48m$ by $49m$ and each cell in the 2019 dataset is $67m$ by $83m$.

Sensors XXE001, XXE002, XXE003, XXE004, and XXE006 in 2019 were placed in roughly the same location as 2018 sensors LT, LY, MV, TC, and BS respectively.

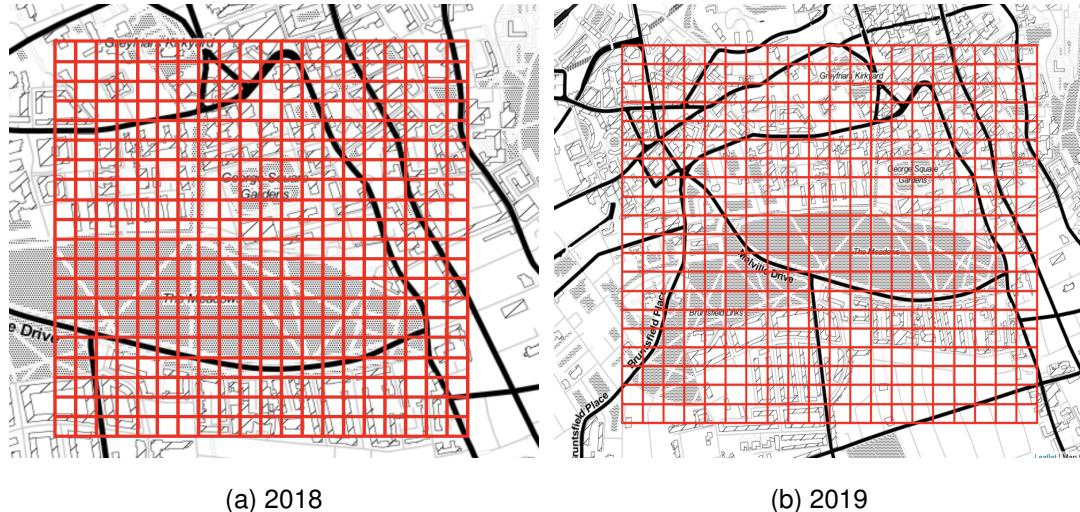


Figure 4.3: An overview of the 2018 and 2019 area with the 20×20 grid overlayed.

4.1.1 Calibration

As mentioned in background Section 2.2 Data Collection, both the Airspeck-S and Airspeck-P collected relative PM values. The data collected by Petard by the mobile Airspeck-P's were calibrated by using an already calibrated Airspeck-S as the ground truth. This was done by leaving the sensors together in a lab for approximately 20 hours. Because the sensors show relative air pollution levels they all show different levels of air pollution over this period. Therefore, the sensors needed to be calibrated to show the same levels. A linear scaling can be used to make all sensors show the same values during the calibration period as long as relative humidity is below 80% [10]. A different more complex non-linear method involving neural networks and two weeks of training is used to calibrate data when relative humidity is greater than 80%. For this project only data with relative humidity less than 80% was used.

4.1.2 Preprocessing Data

To make use of the data in a grid format, it was necessary to map each real coordinate to a cell in the grid. The size of the grid varied from 20x20 to 100x100 cells. A 20x20 grid can be seen in **Figure 4.3a** and a 50×50 grid can be found in **Figure A.3**. The real latitude and longitude coordinates were converted to “grid” latitude and longitude based on which cell they fell in. Each data point from both the mobile and stationary data was converted into grid format. When there was more than one reading in a cell during a timestep, all values were averaged using the arithmetic mean.

As Mentioned in the background Section 2.3 Open Street Map, land usage and road type play important roles in predicting air pollution levels. Each cell in the grid included a land usage and road type. Both land usage and road type were found using OSM. To determine which category of land usage to assign each cell the most prominent land type was chosen. For road type, the highest level road contained in a cell was chosen. Highest meaning largest, for example if a pedestrian path and a primary road are in the same cell the cell will be labelled as a primary road.

Each datapoint has an associated timestamp. In its raw form this feature cannot be used but was converted to make features such as “hour” and “minute”. The idea being that if data is cyclical “hour” and “minute” should help predict a given PM level based on time.

Like most time-series data this dataset was noisy. To account for noise, each timestep was an average of 15 minutes windows within the grid cell. Stationary sensors collect data every 5 minutes and mobile sensors collect data about every 10 seconds so the cells with stationary sensors averaged three points in each window (unless a mobile sensor was also collecting data in that cell).

When training with mobile data, it is necessary to estimate the next PM value to use as the prediction label. For mobile data there were two options: using the nearest neighbor, or only using the datapoints that had data in the next timestep. The first option reduced the accuracy of the next PM value, whereas the second option significantly reduced the number of datapoints. Both methods were tested while training the model and the full nearest neighbor approach was used. This nearest neighbors approach is only necessary for mobile data, because stationary sensors as they always have a next data point.

In some test cases the previous PM value was also used in prediction in an attempt to leverage more of a history. The previous value was found in the same way the next

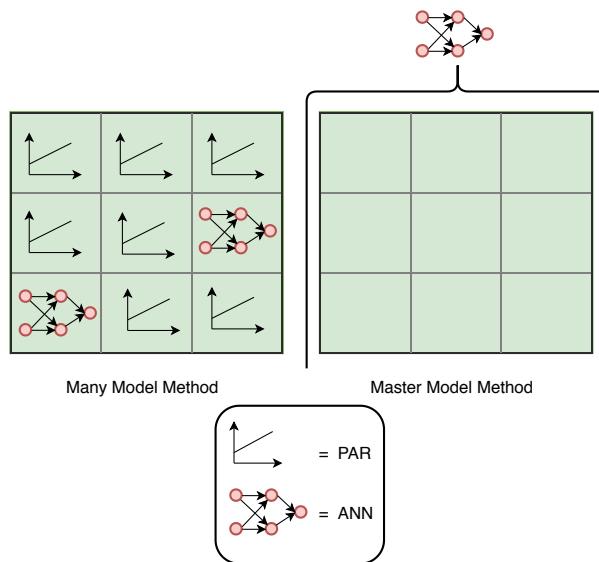


Figure 4.4: The difference between the master and many model methods. The many model method uses the same number of models as there are grid cells with PAR models in cells with no stationary sensor and ANNs in cells that do have stationary sensors. Whereas the master model just uses one ANN.

PM value was, using a nearest neighbor search.

4.2 Proposed Methods

There are two methods this work created and explored. Both methods require the area of experimentation to be split into a grid. The first method, has a predictor model on each cell of the grid, with an ANN on all cells that have a stationary sensor and a PAR on the remaining cells. This will be known as the “Many Model” method or “PAR Grid” when referring to Catarino’s work. The second method, a novel approach of using one single ANN to predict for all cells of the grid, then interpolating using a CNN to fill in the spaces between that had no data from mobile or stationary sensors the previous timestep. This method will be referred to as the “Master Model” method. Both models incorporated online learning and the temporal aspect will use an ensemble to help generalize results. **Figure 4.4** shows the difference between the two methods.

4.2.1 Many Model Method

For a grid of 20x20 there are 400 cells and therefore in the 2018 dataset there were 6 ANNs (one for each cell where a stationary model is deployed) and 394 PAR models.

Ideally an ANN would be used on every cell because ANNs are able to predict better and catch more nuanced patterns than a PAR [25]. Unfortunately, training ANNs require much more computing power and time to train so using an ANN on each cell of the grid is unfeasible.

4.2.2 Master Model Method

The second model is a novel approach and the main focus of this work. The proposed model will still use a grid as it lends itself to better interpolation and more accurate validation. The steps of prediction are to first predict temporally then interpolate to get spatial estimations. When any new data arrives it will be used to predict fifteen minutes in the future which will lead to a sparsely filled matrix. The predicted values are then used to interpolate using a CNN.

This model is suggested because it not only trains faster and requires less memory than the grid method, but because it is able to seamlessly apply mobile data to the training data. The model allows any new data to be added to the model both for training and validation since there is not one model per cell on the grid. This was necessary because of the way mobile data is sporadic and unpredictable in its availability. The grid method cannot fully utilize mobile data because there are 400 models so that the amount of data on each of those models is extremely sporadic and sparse. This method also helps in that for the 2019 dataset all stationary sensors were deployed at different times and in different locations than the 2018 sensors.

This work is missing features that have strong predicting power such as wind speed, wind direction, congestion, and elevation [38] [20]. All are features that could easily be added into the model in the future as this model allows for easy addition or removal of features.

Similar to the “many model” method this model will update the temporal aspect online. There were two options in selecting when to update the model: after each datapoint or after each timestep. The final model updates after each timestep, and will be discussed further in section 5.4: Update Decision.

4.2.3 Ensemble

Unique to this work in predicting spatio-temporally is the use of ensembles to help generalize and better fit to spikes and cyclical datapoints. Online learning should help in catching the spikes and more accurately predicting them. However, since the model

learns off of possible noise there needs to be a counter measure to ensure it still predicts the cyclical or typical data. This work proposes to use an ensemble of the pretrained model and online model to ensure that both spikes and cyclical data are predicted. The batch model should handle the cyclical data and the online model should predict the volatile data. Both methods employed an ensemble to attempt to get the benefits of both the online model and the batch pretrained model.

There has been work to show that ensembles can generalize better and add robustness in machine learning algorithms [51] [52]. This means that even if the ensemble does not account for both cyclical data as well as spikes the model should still perform better due to more robustness. The stochastic nature of ANNs mean that when using the same data to train and validate as well as all the same parameters there will be different results. Ensembles use the stochastic nature to provide more robustness similar to a Random Forest.

4.2.4 Bias Correction

A secondary method to help learn broader patterns is the implementation of a bias correction. It is complimentary to online learning in that it attempts to fix flaws in the model in real time. Online learning updates the parameters of the model to better fit the data after each timestep whereas bias correction is updating the results of the model to better fit historical real values. This is an approach used in the past by previous researchers in climate and pollution spatio-temporal settings [19] [6] [30]. This can be used if the model is consistently predicting above or below the true value.

In this work a simple linear bias correction is used. After the model predicts a PM value, the predicted PM value is updated by adding the average previous errors. For example if the model predicts $1.10\mu g/m^3$, but the model is wrong by an average of $0.22\mu g/m^3$ the bias correction will change the predicted value to be $1.32\mu g/m^3$.

4.3 Performance Metrics

To measure the success of each model Mean Absolute Error (MAE) was used as both the loss function while training and the benchmark for comparison. The main reason MAE was chosen is because Catarino's work used MAE and that work is the best benchmark to use for comparison [6]. MAE can be seen in **Equation 4.1**, where n is the total number of datapoints being predicted, f_t is the $PM_{2.5}$ value predicted by the

model, and y_t is the real value at the predicted time and location.

$$MAE = \frac{1}{n} \sum_{t=1}^n |f_t - y_t| \quad (4.1)$$

A good MAE will be low, therefore a MAE of 0 is a perfect prediction. MAE can be interpreted as the average absolute difference between f_t (the prediction) and y_t (the true value). The downside of MAE is that it is impossible to relate results from one dataset to another [14].

Mean Absolute Percentage Error (MAPE) on the other hand allows for comparison of error rates from one dataset to another. Although, as some datasets are harder to predict on than others, a direct comparison might not be appropriate. For example, predicting in the Meadows in Edinburgh is easier than predicting in downtown Delhi. MAPE is not perfect either for comparing from one dataset to another. MAPE is shown in **Equation 4.2**

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|f_t - y_t|}{y_t} \quad (4.2)$$

f_t and y_t are the same for MAPE as MAE, they represent the prediction and the ground truth respectively. The difference between MAPE and MAE is that MAPE divides the absolute difference between the prediction and ground truth by the ground truth to standardize the error. The error is multiplied by 100 to get a percentage. A perfect MAPE is 0% and has no upper limit. MAPE is asymmetric in that equal errors above the actual value result in a greater MAPE than those below the actual value which can incentivize a model to predict low numbers if it is trained using that as the loss function [17].

4.3.1 Validation

To validate the results the data was split 80/20 between a training and validation set. Because this is time-series data it was necessary to keep the data in the same order, which does cut down on the amount of data that can be used for cross-validation. For experiments in searching for hyperparameters on the 2018 dataset the training set was the first 1450 data points and the validation was the next 358. The models would be trained on the training portion then fed one step at a time of the validation data. This allows for the models to update after each new datapoint. At the end, the 358 predictions are compared to the ground truth to calculate MAE and or MAPE.

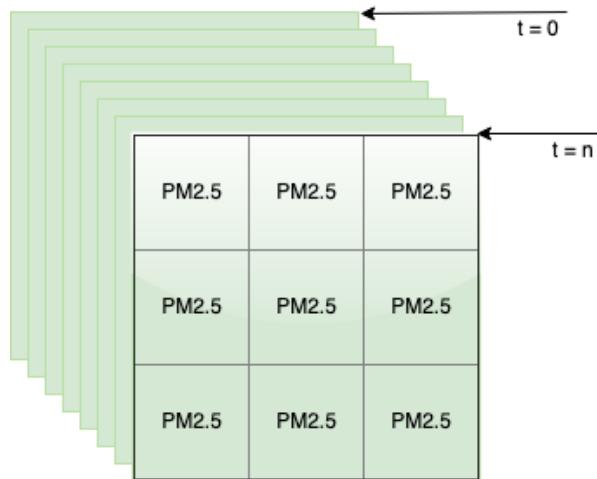


Figure 4.5: A visual representation of the 3D real value matrix. Each box represents a grid of real PM_{2.5} values at a timestep.

A validation grid was created for each timestep. This is essentially a 3-D array of timestep, latitude grid coordinate, longitude coordinate, and the ground truth value. **Figure 4.5** is a visual representation of the array of validation grids. Predicted values were compared to its corresponding grid label values to find MAE and MAPE.

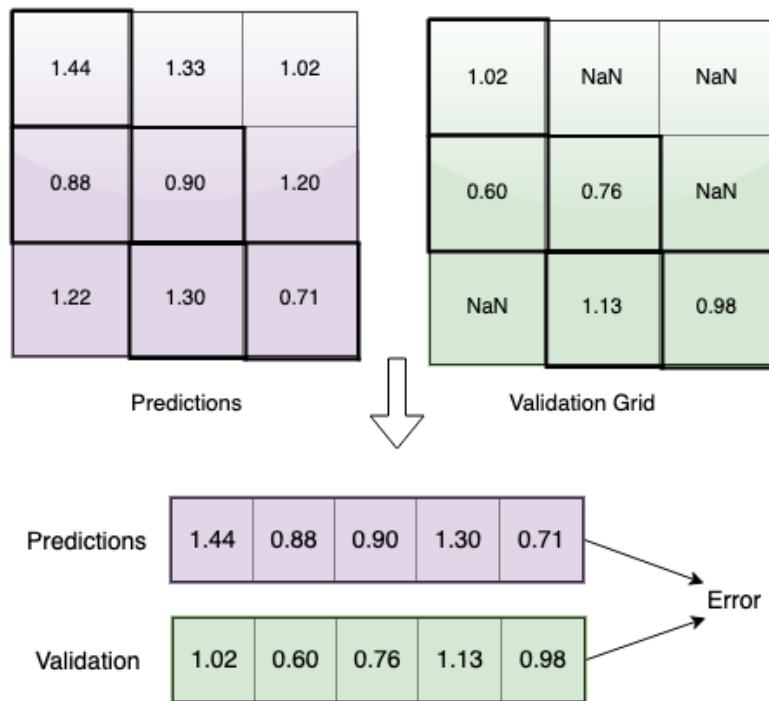


Figure 4.6: A visual representation of how prediction and validation grid values are transformed to arrays and then used to calculate error. This happens at each timestep. Cells on the prediction and validation grid are matched. When both have a real value the value is transferred to separate arrays but maintain the same position in their respective arrays to be evaluated.

Chapter 5

Results and Discussion

5.1 Hyperparameter Search

Before any experiments were done, ideal hyperparameters needed to be selected for all models. This includes PAR, spatial CNN, and temporal ANN for “many model” and “master model” approaches. Optimal PAR and CNN hyperparameters were already found by Catarino and Petard respectively [6] [35].

A grid search was conducted for the ANN hyperparameters. It was found that a simpler model performed better than a deep complex model. A feed-forward model consisting of two layers of 8 nodes performed better than deeper and wider models. An Adam optimizer was used with a learning rate of .0001 and default betas and epsilon (beta1=0.9, beta2=0.999, epsilon=1e-08). The results of the one and two layer grid search for learning rate and number of cells in each layer can be seen in **Figure 5.1**.

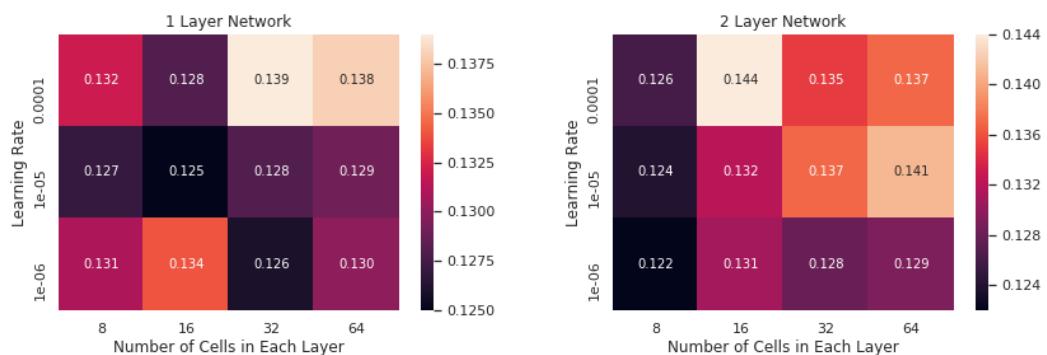


Figure 5.1: The results of the grid search for temporal aspect of the master model. The left is using a model of 1 layer and the left is a model of 2 layers.

A quick look at **Figure 5.1** shows that there is not much pattern to the hyperparam-

eters used, nor the number of layers. There is one trend and that is that simple models seem to be preferable to a complicated deep model.

5.2 Feature Selection

Tests were conducted to find which features were important to the master model. In the end the current $PM_{2.5}$, humidity, lat grid, long grid, land use, and road type were used. Hour, minute, temperature, true latitude, true longitude and any past $PM_{2.5}$ values were dropped as features. The true latitude and longitude were dropped because the measurements were not precise enough and data was already being averaged in a grid so it made more sense and produced better results to use the grid locations as a feature rather than the true latitude and longitude. The humidity used as a feature is relative humidity meaning that temperature is implicitly in it. Keeping both would cause some collinearity as these two variables are highly correlated.

Hour and minute also caused a decrease in performance according to MAE. Minute was not expected to be a good indicator for prediction because of the way the data is averaged over 15 minute timesteps. The loss of performance when using hour was surprising considering that an assumption before was that patterns of cyclical data will be found using hour. Perhaps with a different model that would be the case but the ANN mostly used the current PM value with road type and land usage to predict the next PM value. This also suggests that a time of day does not affect PM levels. We know that is not the case by looking at the plot of PM values over the month of data collection (**Figure A.1** and **Figure A.2**). It is more likely there simply is not enough data to find minute patterns like hour as there are only nineteen days in the dataset and therefore each hour does occur frequently enough. A second factor affecting hour is that there is a high concentration of mobile data only from one time period in the day (2-4 PM) because in 2018 the goal of this data was to test spatial interpolation.

5.3 Spatial CNN

When implementing the spatial CNN similar to Petard's work it was found that Petard ran the code with a small bug [35]. The output of the model would be run through a sigmoid function. A sigmoid will squash a result between 0 and 1. Often times this is used to represent a probability distribution, but in this case the CNN should have been predicting a real valued number above zero. While the results Petard showed are good

and relatively close to Ordinary Kriging that is because Edinburgh generally has a low PM count between 0 and 1. The model was also helped by the way the dataset in 2018 was calibrated. The sensors were calibrated relative to each other rather than a true ground truth and the resulting PM values were lower than the ground truth.

The sigmoid function was removed and a grid search for the best hyperparameters was run (learning rate = 0.001, kernels = 1). The resulting CNN predicts better than OK by a slight margin and the MAPE of the CNN decreased by 18 points. Table 5.1 compares the new CNN against Petard’s CNN as well as the best performing statistical interpolation model Ordinary Kriging [35].

<i>Spatial Interpolation Model</i>	<i>Dataset 2018 19 Days</i>
Petard CNN Baseline	57.1%
Ordinary Kriging	43.7%
Proposed CNN	39.1%

Table 5.1: The MAPE results of Petard’s CNN against Ordinary Kriging interpolation and the updated CNN from this work. The new CNN performs better than both existing models on the 19 days that were tested in Petard’s work from the 2018 dataset [35]

The updated MAPE value reflects what Petard hypothesized in 2018 that the CNN should perform better than any statistical interpolation model [35]. These results are the quantitative proof of what Petard’s figures were qualitatively showing. Petard showed that along smaller roads such as footpaths air pollution is lower than along major roads.

5.4 Update Decision

As mentioned in Section 2.4, Online Learning, online learning relies on the instantaneous updating of the model after new data arrives. Standard time-series data is 2 dimensional as in there is time and a value. Spatio-temporal forecasting is unique in that it is 3 dimensional, time, space and value. When using online learning in a standard time-series forecast the update to the model would happen right away when the new datapoint for the timestep arrives. Since spatio-temporal is 3 dimensional each timestep has as many datapoints as there is space in the forecast area. A question faced was whether to update after each timestep or datapoint is predicted. The number of

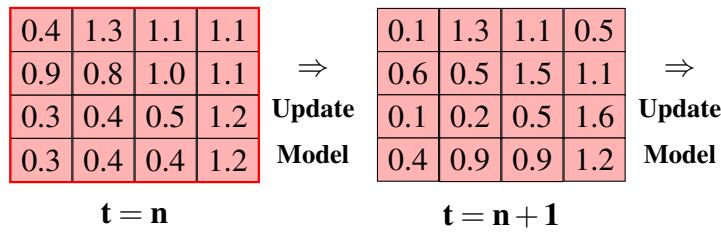


Figure 5.2: A demonstration of how updating after each timestep works. The grid represents a spatial grid used for the master model. Each cell represents a PM value at a coordinate at time t . Each red cell is a cell that will be updated. All cell values are used to update the model at each update.

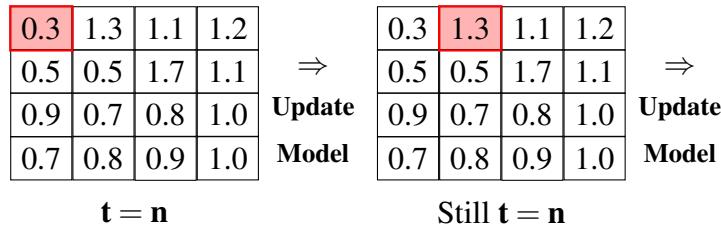


Figure 5.3: A demonstration of how updating after each datapoint works. The grid represents the spatial aspect used for the master model. Each cell represents a PM value at a coordinate at time t . Each red cell is a cell that will be updated. Only one cell value is used to update the model at a time.

datapoints in a given timestep could be between 6 (only stationary sensor data) and somewhere in the low 100s (many people collecting data in the same timestep).

Updating after each datapoint is predicted seemed like the logical answer because since this is online learning the idea is to update after each datapoint. This would allow the updates in a given timestep to adapt to the errors made by the preceding predictions in the timestep. The downside is this could lead to a localized change in PM value at one sensor affecting the next sensor making a prediction.

The other option is to update the model after all datapoints in a timestep have made a prediction. If the problem is viewed as each timestep being a multi-dimensional output this makes sense. Update after each full timestep is predicted so as to only influence the next timestep and not the current timestep.

The results can be seen in table **Table B.1**. This shows that the second option of updating at the end of each timestep provides better results.

A second decision needed to be made for the number of epochs to run when updating the model. The number of epochs at each update acts as a proxy for learning rate.

The higher the number of epochs the more often the model will be adjusted. Therefore a high number of epochs can make the model overfit for the data in the last update. This can be useful when there is a spike in the data as it would be helpful for the model to shift predictions to the spike. Performance time is not an issue even if running 1000 epochs because of the small number of updates being made at each time of update. Epochs from 1 to 1000 were tested. It was found 20 was a good compromise between overfitting yet being able to adapt to changing data. This was not necessarily the best result for the online ANN alone, but provided the best boost in performance when used in an ensemble.

5.5 Temporal Prediction

A goal of this work was to see if ANNs can increase performance of temporal predictions, and how effective online learning in an ensemble is in helping predictions. The validation set is the final 358 timesteps from the 2018 dataset for temporal prediction. To measure the model's accuracy only the predictions made in the cells with stationary sensors were used. **Table 5.2** shows the results of the “many model” method.

The pretrained model mainly just predicts the same value at all timesteps meaning that the ANN was not able to find a pattern to fit the data. The online predictions move more with the trends and fit the cyclical data better as it is mostly trying to mimic the last datapoint. This goes against the original hypothesis and Angelinin’s work that pretrained models perform better on cyclical data than online models [3]. While this is the case for ANNs, it may not be the case for a statistical method. A Baseline of using the past PM value performs well as seen in **Table 5.2**. A surprising result is that the PAR preforms better than all ANN models. This is most likely due to the ANN’s inability to learn on 1450 datapoints, this is discussed further in the “master model” temporal results section.

The ensemble models do have the desired effect of having better prediction power than the model’s that it is made up of. This is probably due to better generalization and not so much of the original hypothesis that online methods can find spikes and the pretrained model finds the cycles. No model or ensemble performs as well as the baseline.

An LSTM was tried to leverage more of the historical data than the feedforward model or PAR. The LSTM performed slightly worse than both the online, batch pretrained, and PAR models.

	<i>Model</i>	<i>Mean Absolute Error</i>
1	Baseline	0.128
2	PAR	0.146
3	ANN Online	0.173
4	ANN Pretrained	0.154
5	LSTM	0.164
6	Ensemble of 2 & 3	0.150
7	Ensemble of 2 & 4	0.141
8	Ensemble of 3 & 4	0.152

Table 5.2: Results of temporal predictions on the 2018 dataset using the first 1450 datapoints as training and remaining 358 as test. Ensembles are even 50/50 splits of the components.

The baseline performed better than all the machine learning models. This is not rare in time-series data when using a naive forecast as the baseline [23]. Naive forecasts are best when following a random walk, which is a dataset that is often hard to predict because of random spikes. Naive forecasting also does well when the dataset is not decomposed for seasonality, or other trends. To make accurate predictions there would need to be a “leading indicator” for example economists use Purchasing Managers Index (PMI) in order to predict growth in a nation’s gross domestic product (GDP) [40]. It can be concluded that there are not enough useful features in the model. With only humidity, $PM_{2.5}$, and hour being collected for the “many model” method it is clear no pattern is being found in the data to make accurate predictions.

While the “many model” method does not predict better than the baseline, the “master model” does. It was hypothesized that the “master model” would predict similar to the “many model” method due to having more datapoints even though it is not individually created for each cell. The “master model” is able to perform at about the baseline, and slightly better using an ensemble. The results can be seen in table 5.3.

The improved results are most likely due to two conditions: The model has more data, and the model can learn from other locations. There is more data because instead of just using data that occurs in one cell it is able to learn from all data, stationary and mobile. Therefore the model also learns from what is happening in other cells. So if a pattern is observed by one sensor the model has the ability to adapt that for all locations. These factors give the master model the ability to learn more nuanced

	<i>Model</i>	<i>Mean Absolute Error</i>
1	Baseline	0.128
2	Master PAR	0.213
3	Master ANN Online	0.131
4	Master ANN Pretrained	0.126
5	Ensemble of 3 & 4	0.124

Table 5.3: Results of temporal predictions on the 2018 dataset using the first 1450 datapoints as training and remaining 358 as test. Ensembles are even 50/50 splits of the components.

patterns in the data.

5.6 Spatio-temporal Prediction

5.6.1 Master Model Comparison to PAR Grid

The “master model” was compared to the PAR Grid using the 2018 dataset. In Catarino’s experiments the data was split into two three-day windows: July 3 to July 6 (Dataset 2018-A) and July 23 to July 26 2018 (Dataset 2018-B). Dataset 2018-A has 457 datapoints and Dataset 2018-B has 287 datapoints [6].

The first attempt was to use the “master model” as is, making temporal predictions at each datapoint available then interpolating. Interpolating using OK was used as the master model baseline. The results from predicting at every point then interpolating using OK were poor as seen in **Table 5.4**. There were some erratic predictions made after interpolation, such as extremely large values and negative numbers. Both of which are bad predictions in Edinburgh or impossible. There were 55 datapoints where the model predicted an unreasonable value. When inspecting the data there is nothing strange except above average PM_{2.5} values for the mobile data as well as the mobile values being approximately 4 times as large as the stationary values (exact average values can be seen in **Table A.6**). The reason for the erratic predictions was coming from a combination of the mobile temporal predictions and using OK for interpolation. **Figure 5.4** shows why sometimes OK interpolations are bad. One map looks good with smooth transitions, whereas the other shows a grid with only one value. This occurs when OK cannot find other values that minimize the kriging estimator due to sparsely

filled grid with high variance values.

When predicting temporally strictly from the stationary sensors, accuracy was significantly better. **Table 5.4** shows how much better predictions from only stationary sensors are than predictions from all sensor data. Later, interpolating using the CNN to interpolate is explored, and predictions during the same 55 datapoints were reasonable.

When the predictions were examined next to the ground truth it was found that the predictions were almost always under the true values, this is due to the difference in PM values between the mobile data and stationary data since the mobile data is being used as the validation set. **Figure 5.6a** shows that most predictions were too low in that they were mostly to the left of the zero error line. To account for this the simple bias correction method described in Section 4.2.4, Bias Correction was put in place. The method used was to calculate the previous average error across the entire grid, then add the average error to the next prediction. This produced significantly better results because of the way it scaled the predicted values toward the ground truth values as seen in **Figure 5.6b**. The addition of the error factor greatly helped the models as seen in **Table 5.4**. Similarly, the predictions on a graph without bias correction can be seen in **Figure 5.7a** and **Figure 5.7b** with bias correction. The simple bias correction was enough to linearly shift the predictions closer to the real values to reduce error.

<i>Data Used</i>	<i>Model W/ OK</i>	<i>Dataset 2018-A</i>	<i>Dataset 2018-B</i>
	Catarino Baseline	2.28	1.21
Mobile & Stationary	Online	19.68	1.40
	Pretrained	25.85	1.26
	Ensemble	25.66	1.33
Stationary Only	Online	2.39	1.69
	Pretrained	2.41	1.69
	Ensemble	2.40	1.69
Stationary Only w/ Bias Correction	Online	1.48	1.19
	Pretrained	1.47	1.20
	Ensemble	1.47	1.20

Table 5.4: Spatio-temporal MAE results for master model. The first section the model predicts on all available datapoints then interpolates. The second section predicts only on stationary data then interpolates. The last section adds the bias correction. All are compared to the baseline from Catarino's work on the same dataset [6].

<i>Data Used</i>	<i>Model W/ CNN</i>	<i>Dataset 2018-A</i>	<i>Dataset 2018-B</i>
	Catarino Baseline	2.28	1.21
	Ordinary Kriging Best	1.47	1.19
Mobile & Stationary	Online	1.99	1.30
	Pretrained	1.94	1.36
	Ensemble	1.93	1.32
Mobile & Stationary	Online	1.71	1.38
w/ Bias Correction	Pretrained	1.75	1.35
	Ensemble	1.71	1.33
Stationary Only	Online	2.36	1.54
	Pretrained	2.45	1.44
	Ensemble	2.40	1.47
Stationary Only	Online	1.51	1.18
w/ Bias Correction	Pretrained	1.50	1.16
	Ensemble	1.50	1.16

Table 5.5: Spatio-temporal MAE results for master model. The first section the model predicts on all available datapoints then interpolates. The second section predicts only on stationary data then interpolates. The last section adds the bias correction. All are compared to the baseline from Catarino’s work on the same dataset [6].

Section 5.3, Spatial CNN, showed that a CNN is better than OK at interpolating spatially. Therefore it was hypothesized that spatio-temporal predictions using a CNN would also be better. **Table 5.5** show the full spatio-temporal prediction results using CNN for interpolation.

The main difference between using a CNN and OK for interpolation is when both mobile and stationary sensors are used for temporal prediction on dataset 2018-A. The CNN is able to keep the MAE closer to a reasonable value than OK. The issue with OK is in many cases it fills all interpolated cells with the same value seen in the map in **Figure 5.4b**. Other times it fills in all cells smoothly transitioning from one value to the next, but has no regard for land usage, or road type of those cells shown in the map in **Figure 5.4a**.

The CNN predictions are less smooth, but that is okay and expected because it uses road type and land usage as features while interpolating. The map in **Figure 5.5** shows how CNN interpolated maps can look somewhat random, but notice that the north and

east where there are busier roads and more congestion it is more polluted than the center and around the park where there is green space.

The results also showed that using online, pretrained, or ensemble do not matter much. This is most likely because the amount of cells on the grid determined by the temporal prediction are between 6 and 100 out of 400. That means at most 25% of the cells being evaluated are determined strictly by the temporal prediction. The interpolation portion has more of an effect on the total MAE.



Figure 5.4: The left image is a good example of spatio-temporal predictions using OK for interpolation. An example of a bad ordinary kriging interpolation on the right. Notice the six cells that are different are the cells that have stationary sensors.

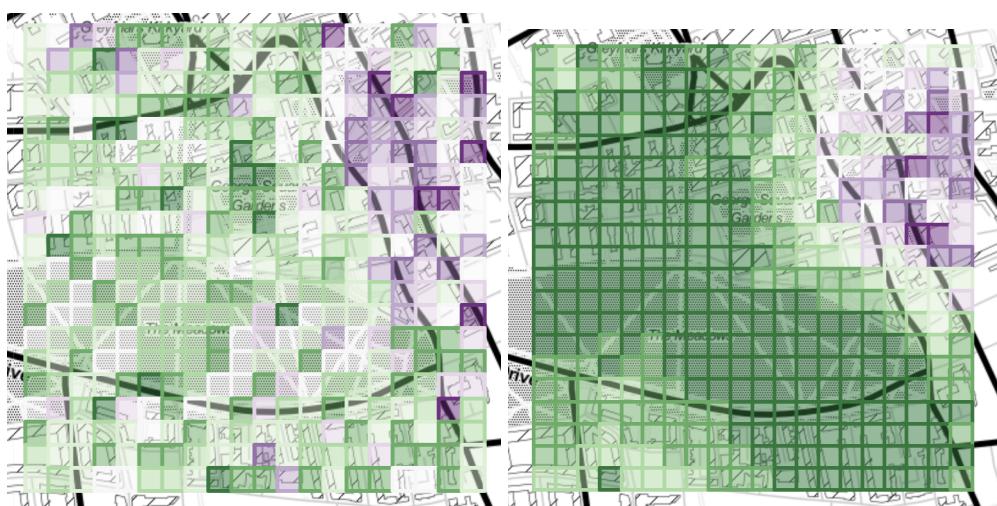


Figure 5.5: Two examples of spatio-temporal predictions using CNN for interpolation.

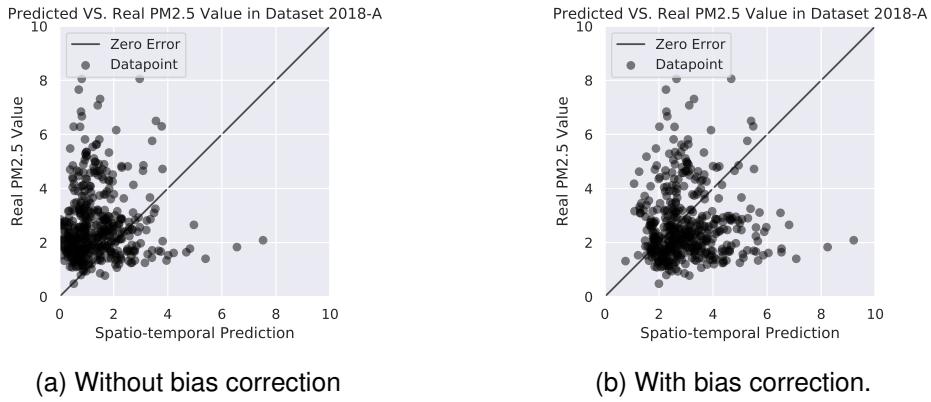
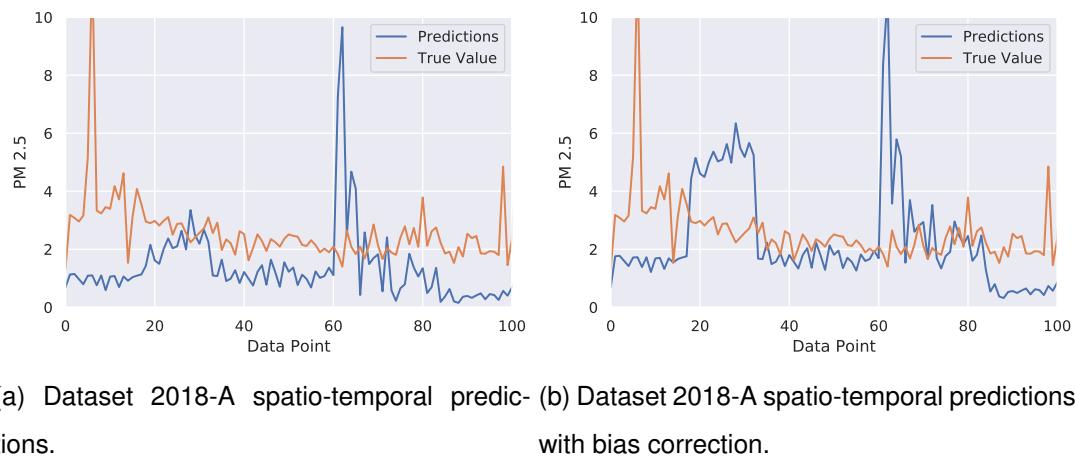


Figure 5.6: The figure shows error the models predicted PM_{2.5} value against the real PM_{2.5} value. The line through the middle represents a perfect prediction with zero error. Notice with bias correction the datapoints shift to the right signifying a higher model prediction.



5.6.2 Grid Size

To try to determine the importance of land use and road type the grid size was increased to 50×50 . Changing the grid size changed the size of each cell from $48m \times 49m$ to $19m \times 19m$. This allows a more precise estimation as to what type of land and road is contained within the cell.

The hyperparameters for the CNN and Master Model were tuned for a 50×50 grid and the same experiments from the last section were run again. Using datasets 2018-A and 2018-B it was found that results were the same if not worse than the 20×20 grid. One of the factors determined in the last section to be preventing the model from

making more accurate predictions is sparse data. Even though it was hypothesized that the master model should handle sparse data better it does not. By increasing the grid size it makes the data even more sparse, as there are now 2500 cells compared to 400 in the 20×20 grid.

Results were roughly the same on the 100×100 grid. This is more of an indictment on the model's inability to handle sparse data rather than the lack of information land usage and road type add.

5.6.3 Utility of Mobile Data For Training

This experiment determines how helpful mobile data is to this model. Since there is a rich amount of stationary data in both datasets compared to mobile data, there is potential that the model makes all decisions based on the stationary data. In a sense the mobile data could be so infrequent that it does not alter the model enough to make an impact positively or negatively.

Training without mobile data negatively affects the temporal aspect of the master model. The results of just predicting on the stationary sensors dramatically drops off when mobile data is removed. Mentioned in the original temporal section, the model is able to learn features helpful to prediction of PM values from other sensors.

As expected, if the temporal predictions are worse without mobile training data, the spatio-temporal are worse as well. Results of spatio-temporal results can be seen in **Table 5.6**. The spatio-temporal predictions are significantly better for the model that uses mobile data to train and make predictions. This is because the model learns more from more datapoints and when predicting from mobile datapoints more of the grid is filled accurately before interpolation.

<i>Spatio-Temporal Model</i>	<i>Dataset 2018-A</i>	<i>Dataset 2018-B</i>
All Data Best	1.50	1.16
Master Model Ensemble w/o Mobile	2.37	1.66

Table 5.6: Spatio-temporal MAE results for the best model with mobile data for training and the best model without mobile data for training. The model performs better on both datasets with mobile data.



Figure 5.8: A spatio-temporally predicted and interpolated map using the “Master Model” on the 2019 dataset with one sensor’s mobile collection overlayed.

5.6.4 2019 Dataset Results

Experiments done on the 2019 dataset were brief, but the main focus was on the increased size of the area. Predictions within an area of approximately 1km by 1km is probably not too useful other than showing a concept works, so a larger area was tested as well. Only the “master model” was tested on the 2019 dataset because it was built with scaling in mind and showed it performs best. It can adapt to new data and a larger area without significant increase of computation time or increase in memory usage. Within this new range there were additionally more stationary sensors and more concurrent mobile sensor data.

Strictly temporal results showed that when using the first 80% of the dataset to train and the final 20% as test data the model performs better than the same model trained and tested on the smaller dataset when compared using MAPE (results in **Table D.1**). Direct comparison between two different datasets is not perfect, but it is a good way to get an idea of the quality of the model on a different dataset.

Spatio-temporal results were the same as temporal where results were better on the 2019 dataset. The map seen in **Figure 5.8** shows the same types of patterns from the 2018 maps. The cells are filled less smoothly than the OK maps. The green space in the middle does contain the most unpolluted air and the predictions are mostly in line with the mobile data.

The slightly better accuracy on the 2019 dataset probably has more to do with the increase in timesteps and datapoints than an increase in performance on a larger

area. The model is able to find realistic patterns with more data. Since the temporal predictions do not interpolate this means that at sensors this model predicts well and again better than the baseline even in a larger environment.

<i>Spatio-Temporal Dataset</i>	<i>MAPE</i>
2018 Master Model	62.94%
2019 Master Model	56.68%

Table 5.7: Temporal results for Master model on 2018 and 2019 datasets. Test set was final 20% of timesteps.

5.7 Results Summary

For temporal results the novel “master model” with an ensemble of a pretrained and online ANN had the best results on both the 2018 and 2019 datasets. It was the only model that was able to predict better than the baseline of a naive forecast.

The updated spatial CNN showed the best results for interpolation with a MAPE of 39.1%. The model was an improvement over Petard’s CNN from 2018 and OK.

Spatio-temporal predictions for the “master model” are better than 2018-A baseline and about equal to the 2018-B baseline. The model with OK performed slightly better than the model with a CNN for interpolation.

Table 5.8 shows the best MAPE and MAE where applicable for each prediction type, temporal, spatial interpolation, and spatio-temporal.

<i>Prediction Type</i>	<i>Dataset</i>	<i>Model</i>	<i>MAE</i>	<i>MAPE</i>
Spatial	2018	CNN	N/A	39.10%
Temporal	2018	Master Model Ensemble	0.124	18.51%
	2019	Master Model Ensemble	0.532	17.20%
Spatio-Temporal	2018-A	Master Model W/ BC	1.50	58.87%
	2018-B	Master Model W/ BC	1.16	71.70%
	2019	Master Model W/ BC	3.32	56.68%

Table 5.8: A summary of the best results using the models created in this work. All have best accuracies on their respective datasets.

Chapter 6

Conclusions

The novel model proposed in this work was shown to predict better than previous methods for PM values temporally, spatially, and spatio-temporally. The success of the “master model” comes from its ability to learn patterns observed by multiple sensors at once, allowing the “master model” to learn more nuanced patterns in the data. This leads to the “master model” to make more accurate predictions at all locations in the test area and scale to larger areas or increase the number of sensors.

A unique feature of the proposed model is that it is adept at using new data streams in the area of focus both for training and prediction. This was shown by using mobile data for training and prediction in the 2018 dataset. While also being able to use more stationary and mobile sensors deployed at different times in the 2019 dataset without any modification.

The “master model” ensemble was the only model that was able to perform better than the baseline in temporal predictions with an MAE of 0.124 on the 2018 dataset. The spatial CNN introduced in this work improved upon previous best MAPE spatial interpolation by 3 percentage points and the previous best CNN model by 17 percentage points with an MAPE of 39.10%. The spatio-temporal predictions on the 2018 datasets also had positive results. For the 2018-A dataset, the “master model” predicted better than the previous best with an MAE of 1.50. On the 2018-B dataset, “master model’s” MAE was 1.16 which was also better than the previous best MAE of 1.21.

A new dataset was collected for central Edinburgh for the Centre for Speckled Computing at the University of Edinburgh to use in the future. On this dataset a baseline MAPE for the final 20% of the data was set. A baseline of 56.68% was found. The MAPE on the 2019 dataset was better than MAPE on both 2018 datasets mainly due

to the increase in number of timesteps and datapoints despite the larger area of focus in Edinburgh.

Unfortunately, due to timing not much experimentation was done on the new 2019 dataset. While this work does set a baseline for it, it can be improved upon. Baselines of 17.20% for temporal predictions and 56.68% for spatio-temporal predictions were found.

This work shows that exact values for training are not entirely necessary, as the nearest neighbor approach for training worked better than using the real next datapoint and improved upon the baseline for temporal predictions. But as more data is added there would come a point where using the real next value will train more accurately.

Like past research, this work has shown that the “master model” also does not handle sparse data as seamlessly as expected. To fix this, in the future a Generative Adversarial Network (GAN) could be used to generate fake data similar to the dataset to improve training. This is something that is done in Machine Translation research when working with low-resource languages that do not have much source material for labelled training data [8]. There is also research that shows that this type of training can help generalization [16].

If there was more time to continue with this project the next source of improvement would be to integrate the CNN and temporal ANN into one model. The temporal ANN is able to learn online, but the CNN is pretrained and does not get updated after that. An attempt to further smooth the training data would be made to better learn patterns of the cyclical data. Additionally, a better method for bias correction and “weighting” the ensemble could yield benefits in the future.

In conclusion, this work determined that a single model can be used to predict air pollution across a grid temporally better than any previous method. Secondly, this work produced a spatial interpolation method with MAPE lower than the previous state-of-the-art. When used together these two models were able to predict spatio-temporally better than the best on both 2018 datasets. Lastly, this work created a new air pollution dataset in central Edinburgh using mobile and stationary data.

Bibliography

- [1] Matthew D Adams and Pavlos S Kanaroglou. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. *Journal of environmental management*, 168:133–141, 2016.
- [2] Oren Anava, Elad Hazan, Shie Mannor, and Ohad Shamir. Online learning for time series prediction. In *Conference on learning theory*, pages 172–184, 2013.
- [3] Ivan Angelinin. Sensing spaces: Improving spatial and temporal prediction of air pollution using online learning algorithms. Master’s thesis, The University of Edinburgh, 4 2018.
- [4] DK Arvind, CA Bates, DJ Fischer, and Janek Mann. A sensor data collection environment for clinical trials investigating health effects of airborne pollution. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 88–91. IEEE, 2018.
- [5] DK Arvind, Janek Mann, Andrew Bates, and Konstantin Kotsev. The airspeck family of static and mobile wireless air quality monitors. In *2016 Euromicro Conference on Digital System Design (DSD)*, pages 207–214. IEEE, 2016.
- [6] João Catarino. Sensing spaces: An online spatio-temporal model for pm2.5 predictions using a combination of stationary and mobile airborne particulate sensors. Master’s thesis, The University of Edinburgh, School of Informatics, 4 2019.
- [7] Jianjun Chen, Jin Lu, Jeremy C Avise, John A DaMassa, Michael J Kleeman, and Ajith P Kaduwela. Seasonal modeling of pm2. 5 in california’s san joaquin valley. *Atmospheric environment*, 92:182–190, 2014.

- [8] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.
- [9] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(Mar):551–585, 2006.
- [10] Leigh R Crilley, Marvin Shaw, Ryan Pound, Louisa J Kramer, Robin Price, Stuart Young, Alastair C Lewis, and Francis D Pope. Evaluation of a low-cost optical particle counter (alphasense opc-n2) for ambient air monitoring. *Atmospheric Measurement Techniques*, pages 709–720, 2018.
- [11] Koen De Ridder, Ujjwal Kumar, Dirk Lauwaet, Lisa Blyth, and Wouter Lefebvre. Kalman filter-based air quality forecast adjustment. *Atmospheric environment*, 50:381–384, 2012.
- [12] Luis A Díaz-Robles, Juan C Ortega, Joshua S Fu, Gregory D Reed, Judith C Chow, John G Watson, and Juan A Moncada-Herrera. A hybrid arima and artificial neural networks model to forecast particulate matter in urban areas: The case of temuco, chile. *Atmospheric Environment*, 42(35):8331–8340, 2008.
- [13] R du Plessis, Anuj Kumar, Gerhard P Hancke, and Bruno J Silva. A wireless system for indoor air quality monitoring. In *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*, pages 5409–5414. IEEE, 2016.
- [14] Ryan Egan. Twitter for electoral predictions. *Informatics Research Review INFR11136*, 2019.
- [15] Nicolas L Gilbert, Mark S Goldberg, Bernardo Beckerman, Jeffrey R Brook, and Michael Jerrett. Assessing spatial variability of ambient nitrogen dioxide in montreal, canada, with a land-use regression model. *Journal of the Air & Waste Management Association*, 55(8):1059–1063, 2005.
- [16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [17] Paul Goodwin and Richard Lawton. On the asymmetry of the symmetric mape. *International journal of forecasting*, 15(4):405–408, 1999.

- [18] Tayfun Gürel, Ulrich Egert, Steffen Kandler, Luc De Raedt, and Stefan Rotter. Predicting spike activity in neuronal cultures. *BMC neuroscience*, 8(2):P185, 2007.
- [19] Sabrina Hempel, Katja Frieler, Lila Warszawski, Jacob Schewe, and Franziska Piontek. A trend-preserving bias correction—the isi-mip approach. *Earth System Dynamics*, 4(2):219–236, 2013.
- [20] Gerard Hoek, Rob Beelen, Kees De Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33):7561–7578, 2008.
- [21] Stefan Ivanov. Sensing spaces: Least polluted routes. Master’s thesis, The University of Edinburgh, School of Informatics, 1 2018.
- [22] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental pollution*, 151(2):362–367, 2008.
- [23] Lutz Kilian and Mark P Taylor. Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, 60(1):85–107, 2003.
- [24] RBA Koelemeijer, CD Homan, and J Matthijsen. Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over europe. *Atmospheric Environment*, 40(27):5304–5315, 2006.
- [25] Nowrouz Kohzadi, Milton S Boyd, Bahman Kermanshahi, and Iebeling Kaastra. A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, 10(2):169–181, 1996.
- [26] Sitthinut Kumpalanuwat. The healthy city tour route planner in thearea of central edinburgh. Master’s thesis, The University of Edinburgh, School of Informatics, 8 2019.
- [27] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

- [28] Allen S Lefohn, H Peter Knudsen, and Douglas S Shadwick. Using ordinary kriging to estimate the seasonal w126, and n100 24-h concentrations for the year 2000 and 2003. *ASL & Associates*, 111, 2005.
- [29] Yang Li, Quanliang Chen, Hujia Zhao, Lin Wang, and Ran Tao. Variations in pm10, pm2. 5 and pm1. 0 in an urban area of the sichuan basin and their relation to meteorological factors. *Atmosphere*, 6(1):150–163, 2015.
- [30] Douglas Maraun. Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *Journal of Climate*, 26(6):2137–2143, 2013.
- [31] Lidia Morawska, Phong K Thai, Xiaoting Liu, Akwasi Asumadu-Sakyi, Godwin Ayoko, Alena Bartonova, Andrea Bedini, Fahe Chai, Bryce Christensen, Matthew Dunbabin, et al. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment international*, 116:286–299, 2018.
- [32] OpenStreetMap contributors. Planet dump retrieved from <https://planet.osm.org>. <https://www.openstreetmap.org>, 2017.
- [33] World Health Organization et al. Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: report on a who working group, bonn, germany 13-15 january 2003. Technical report, Copenhagen: WHO Regional Office for Europe, 2003.
- [34] World Health Organization et al. 7 million premature deaths annually linked to air pollution. *World Health Organization, Geneva, Switzerland*, 2014.
- [35] Zoe Petard. High resolution spatial predictions of air pollution levels in edinburgh. Master’s thesis, The University of Edinburgh, 8 2018.
- [36] Hannah Ritchie and Max Roser. Air pollution, Apr 2017.
- [37] Rana Roy and Nils Axel Braathen. The rising cost of ambient air pollution thus far in the 21st century. (124), 2017.
- [38] Patrick H Ryan and Grace K LeMasters. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhalation toxicology*, 19(sup1):127–133, 2007.

- [39] Pablo E Saide, Gregory R Carmichael, Scott N Spak, Laura Gallardo, Axel E Osses, Marcelo A Mena-Carrasco, and Mariusz Pagowski. Forecasting urban pm₁₀ and pm_{2.5} pollution episodes in very stable nocturnal conditions and complex terrain using wrf–chem co tracer model. *Atmospheric Environment*, 45(16):2769–2780, 2011.
- [40] James H Stock and Mark W Watson. A procedure for predicting recessions with leading indicators: econometric issues and recent experience. In *Business cycles, indicators and forecasting*, pages 95–156. University of Chicago Press, 1993.
- [41] Enhao Sun. Least polluted route: Algorithms and applications. Master’s thesis, The University of Edinburgh, School of Informatics, 8 2019.
- [42] M Thirukkumaran. Fine-grained spatiotemporal prediction of air pollutants in new delhi, india. Master’s thesis, The University of Edinburgh, School of Informatics, 8 2019.
- [43] Joris Van den Bossche, Jan Peters, Jan Verwaeren, Dick Botteldooren, Jan Theunis, and Bernard De Baets. Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset. *Atmospheric Environment*, 105:148–161, 2015.
- [44] Lili Wang, Zirui Liu, Yang Sun, Dongsheng Ji, and Yuesi Wang. Long-range transport and regional sources of pm_{2.5} in beijing based on long-term observations from 2005 to 2010. *Atmospheric Research*, 157:37–48, 2015.
- [45] Tijian Wang, Fei Jiang, Junjun Deng, Yi Shen, Qinyan Fu, Qian Wang, Yin Fu, Jianhua Xu, and Danning Zhang. Urban air quality and regional haze weather forecast for yangtze river delta region. *Atmospheric Environment*, 58:70–83, 2012.
- [46] Xinying Wang and Min Han. Online sequential extreme learning machine with kernels for nonstationary time series prediction. *Neurocomputing*, 145:90–97, 2014.
- [47] Z Wang, T Maeda, M Hayashi, L-F Hsiao, and K-Y Liu. A nested air quality prediction modeling system for urban and regional scales: Application for high-ozone episode in taiwan. *Water, Air, and Soil Pollution*, 130(1-4):391–396, 2001.

- [48] Nick Watts, W Neil Adger, Paolo Agnolucci, Jason Blackstock, Peter Byass, Wenja Cai, Sarah Chaytor, Tim Colbourn, Mat Collins, Adam Cooper, et al. Health and climate change: policy responses to protect public health. *The Lancet*, 386(10006):1861–1914, 2015.
- [49] Congcong Wen, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu, and Tianhe Chi. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Science of The Total Environment*, 654:1091–1099, 2019.
- [50] M.c. Woody, H.-W. Wong, J.j. West, and S. Arunachalam. Multiscale predictions of aviation-attributable pm 2.5 for u.s. airports modeled using cmaq with plume-in-grid and an aircraft-specific 1-d emission model. *Atmospheric Environment*, 147:384–394, 2016.
- [51] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.
- [52] Tom Zahavy, Bingyi Kang, Alex Sivak, Jiashi Feng, Huan Xu, and Shie Mannor. Ensemble robustness and generalization of stochastic deep learning algorithms. *arXiv preprint arXiv:1602.02389*, 2016.
- [53] Ruizhe Zhang, Daniele Ravi, Guang-Zhong Yang, and Benny Lo. A personalized air quality sensing system-a preliminary study on assessing the air quality of london underground stations. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 111–114. IEEE, 2017.
- [54] Guangqiang Zhou, Jianming Xu, Ying Xie, Luyu Chang, Wei Gao, Yixuan Gu, and Ji Zhou. Numerical air quality forecasting over eastern china: An operational application of wrf-chem. *Atmospheric Environment*, 153:94–108, 2017.

Appendix A

Additional Dataset Information

A.1 Stationary Sensor Trends

A.2 Calibration Factors

Sensor Name	Sensor Identifier	Grid Row	Grid Column	Calibration Factors		
				PM _{2.5}	Temperature	Humidity
Melville (MV)	02E5F77764B873DA	16	6	1.0	1.0	1.0
Lauriston (LT)	E5FD8C55EAA37555	3	6	1.22	1.19	1.01
Library (LY)	200A7CED9D597407	8	5	1.55	1.51	1.00
Tennis Court (TC)	AA0E63CF5118F98F	13	13	2.38	2.31	0.98
Bristo Square (BS)	B61241EF668DBC2C	3	10	2.34	2.22	0.98
Buccleuch Place (BP)	E786F1568F65C296	8	13	7.08	7.05	0.97
Mobile 1	XXM007	-	-	4.86	1.16	0.74
Mobile 2	XXM008	-	-	4.91	1.19	0.74

Table A.1: Description of sensors and calibration factors used, with Melville sensor as reference for 2018.

Sensor Name	Sensor Identifier	Grid Row	Grid Column	Calibration Factors		
				PM _{2.5}	Temperature	Humidity
XXE001	B2BC10D2B9F14328	4	11	1.96	1.0	1.0
XXE002	FD368D7D6B815C1B	8	11	2.89	1.0	1.0
XXE003	229B8913D6A5B970	12	11	1.84	1.0	1.0
XXE004	DA1F257AD20C85A7	11	16	2.58	1.0	1.0
XXE005	9C4F842B28FC5971	3	19	2.04	1.0	1.0
XXE006	3AB8A58A952B89B6	5	14	2.42	1.0	1.0
XXE007	AE93FD8053BFDDFD	1	12	2.38	1.0	1.0
XXE008	3BC10C97D2490E7B	12	18	2.76	1.0	1.0
XXE009	660CB9BC2A97231A	7	4	2.36	1.0	1.0
XXE010	F753DAAE895C0DEB	9	8	1.54	1.0	1.0
Mobile 1	XXE101	-	-	1.79	1.0	1.0
Mobile 2	XXM102	-	-	1.13	1.0	1.0
Mobile 1	XXM103	-	-	1.80	1.16	1.0
Mobile 2	XXM104	-	-	1.27	1.19	1.0

Table A.2: Description of sensors and calibration factors used, with a sensor not used in the dataset as reference for 2019.

A.3 2019 Deploy Dates

Sensor Name	Sensor Identifier	Grid Row	Grid Column	Start Date	End Date
XXE001	B2BC10D2B9F14328	4	11	June 28	July 30
XXE002	FD368D7D6B815C1B	8	11	June 28	July 30
XXE003	229B8913D6A5B970	12	11	June 28	July 30
XXE004	DA1F257AD20C85A7	11	16	June 28	July 30
XXE005	9C4F842B28FC5971	3	19	June 28	July 30
XXE006	3AB8A58A952B89B6	5	14	July 3	July 30
XXE007	AE93FD8053BFDDFD	1	12	July 3	July 30
XXE008	3BC10C97D2490E7B	12	18	July 3	July 30
XXE009	660CB9BC2A97231A	7	4	July 12	July 30
XXE010	F753DAAE895C0DEB	9	8	July 12	July 30

Table A.3: Date each stationary sensor was deployed and when it was removed for the 2019 sensors.

A.4 50×50 Grid

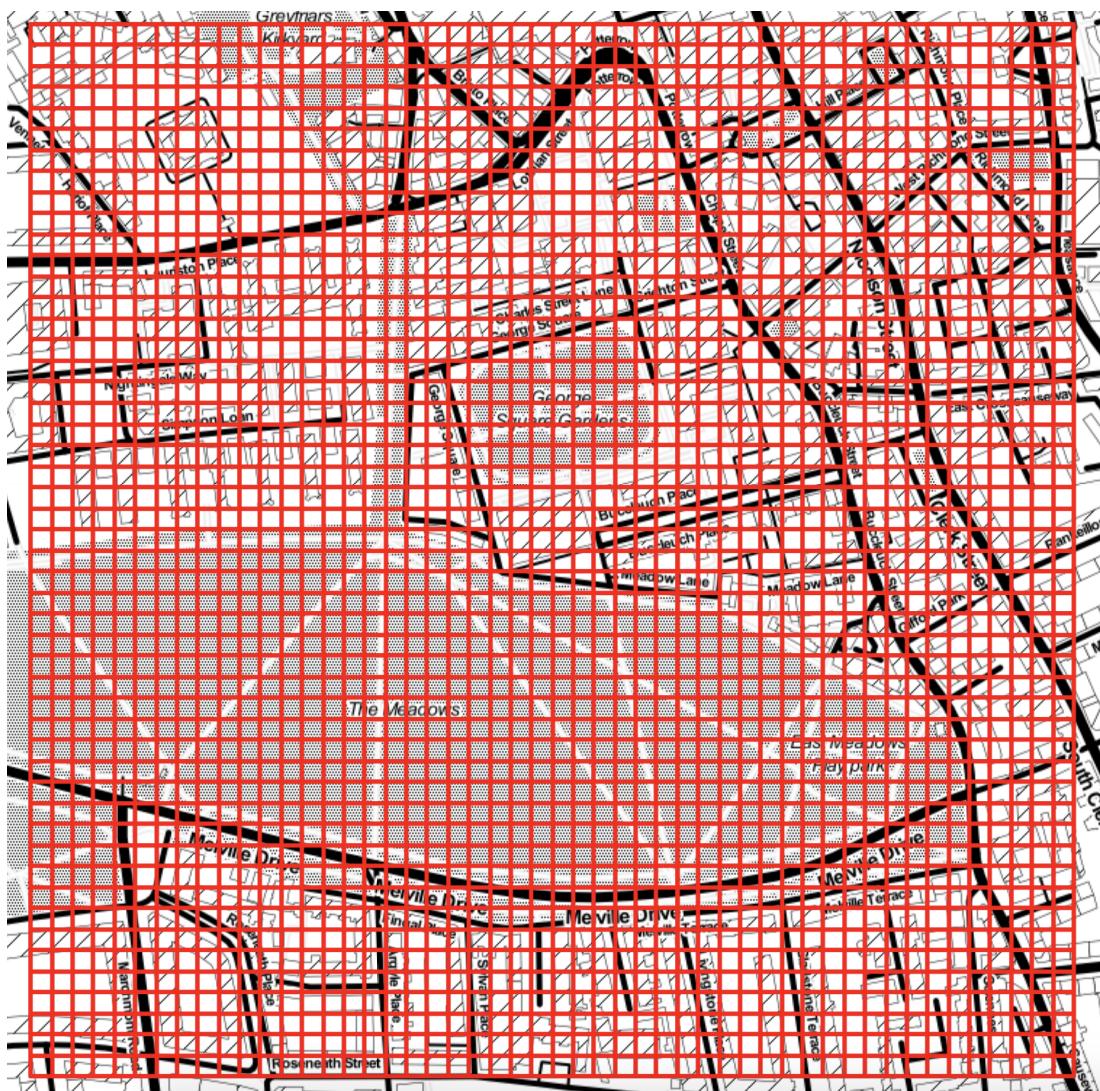


Figure A.3: A map of the 50×50 grid.

A.5 Mobile Vs. Stationary Data

The table below shows the difference in mean values between mobile and stationary data. This led to stationary predicting below validation values.

<i>Feature</i>	<i>Mobile</i>	<i>Stationary</i>
Humidity	43.72	64.51
PM 2.5	2.086	0.588

Table A.4: The mean values of PM_{2.5} and humidity for the mobile and stationary data from Dataset 2018-A

<i>Feature</i>	<i>Mobile</i>	<i>Stationary</i>
Humidity	42.52	59.44
PM 2.5	3.082	1.032

Table A.5: The mean values of PM_{2.5} and humidity for the mobile and stationary data from Dataset 2018-B

<i>Feature</i>	<i>Mobile</i>	<i>Stationary</i>
Humidity	57.95	46.55
PM 2.5	4.619	4.999

Table A.6: The mean values of PM_{2.5} and humidity for the mobile and stationary data from Dataset 2019

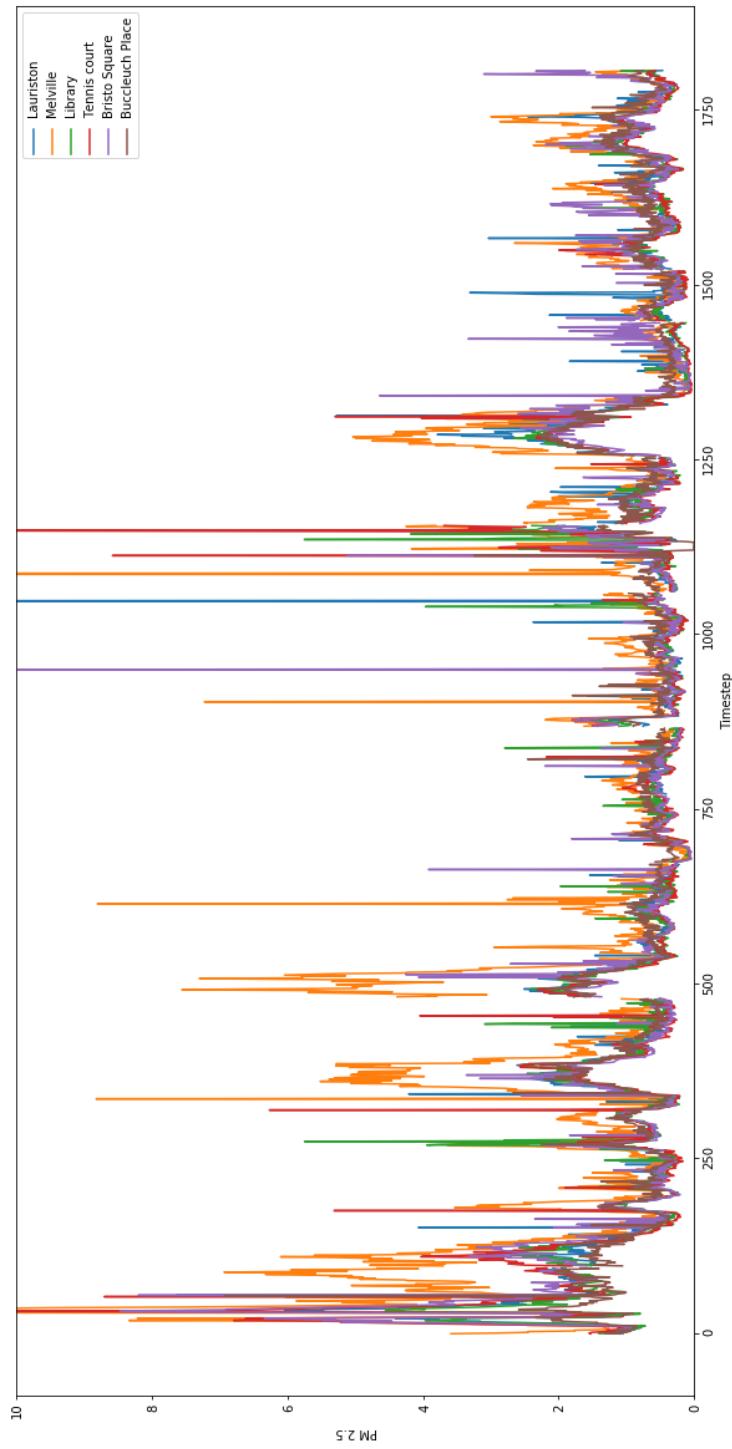


Figure A.1: All stationary data for 2018 dataset.

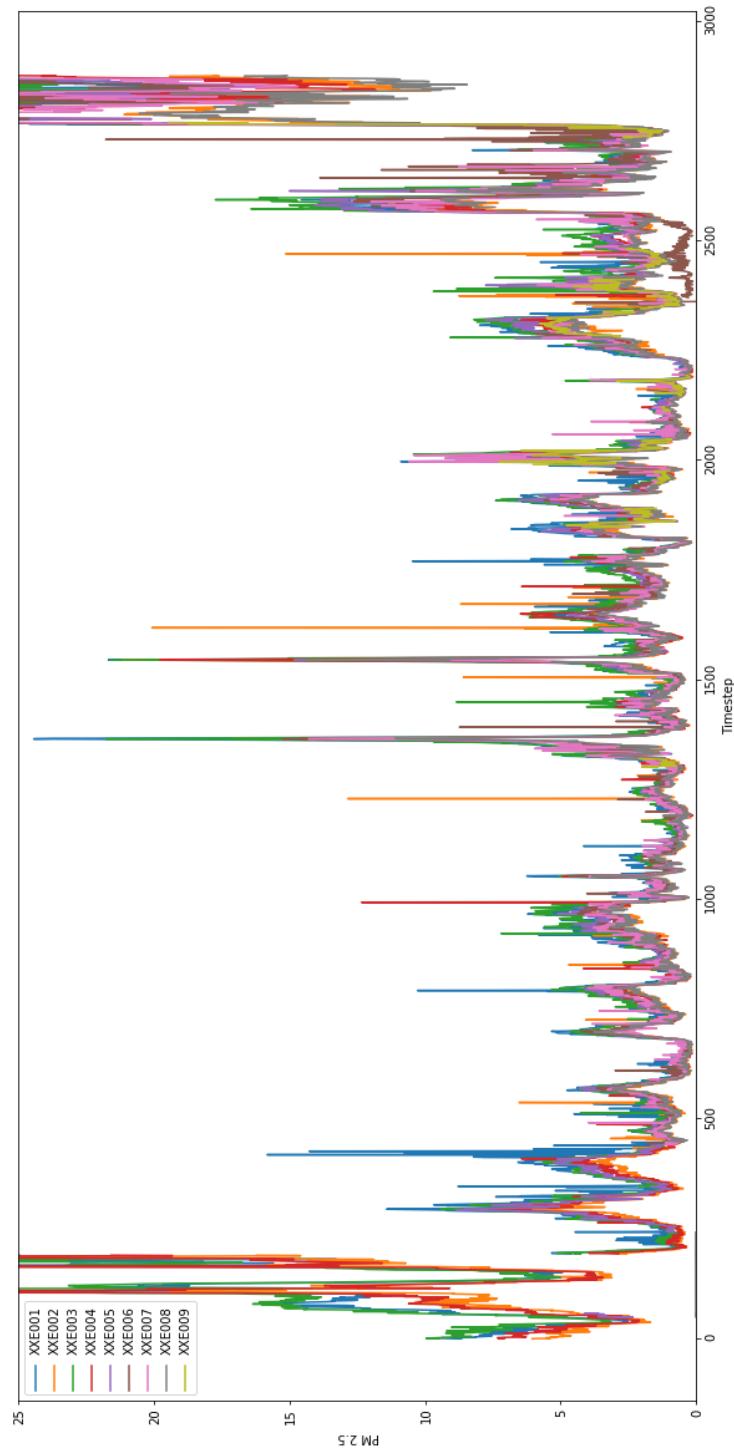


Figure A.2: All stationary data for 2019 dataset.

Appendix B

Update Decision

<i>Update Time</i>	<i>MAE</i>
After Datapoint	0.176
After Timestep	0.131

Table B.1: Temporal results when updating after each datapoint or timestep.

Appendix C

Temporal Prediction Without Mobile Data

<i>Temporal Model</i>	<i>Mean Absolute Error</i>
Master Ensemble w/ Mobile	0.124
Master Ensemble w/o Mobile	0.390

Table C.1: Results of temporal predictions on the 2018 dataset using the first 1450 datapoints as training and remaining 358 as test. Ensembles are even 50/50 splits of an online and pretrained master model.

Appendix D

2019 Temporal Results

<i>Temporal Dataset</i>	<i>MAPE</i>
2018 Master Model	18.51%
2019 Master Model	17.20%

Table D.1: Temporal results for Master model on 2018 and 2019 datasets. Test set was final 20% of timesteps.

Appendix E

Information for Data Collection

Participants



The University of Edinburgh

Online Spatial-Temporal Predictions Using
Mobile and Stationary Data

Ryan Egan
s1876158@sms.ed.ac.uk

Information Sheet

Data Collection For Online Spatial-Temporal Air Pollution Predictions Using Mobile and Stationary Data

I am an MSc in Artificial Intelligence Student at the University of Edinburgh. My supervisor is Professor D K Arvind in the School of Informatics. In conjunction with the Centre for Speckled Computing in the School of Informatics, I am using mobile and stationary data of particulate matter to make near term spatial and temporal predictions of pollution.

What is the purpose of this study?

We aim to develop a model that can effectively use mobile and stationary data to accurately predict particulate matter levels spatially and temporally.

Why have you been asked to take part?

You have been asked to take part in this study because you are a student in close proximity to the University of Edinburgh where data must be collected.

How do I let the researcher know I want to take part?

Before you decide to take part it is important you understand why the research is being conducted and what it will involve. Please take time to read the following information carefully.

If you choose to take part in the study, you can sign the attached consent form. It ensures that you are aware of your rights and explains the extent of your participation.

What happens if you choose to take part?

You will be asked to wear one sensor, the Airspeck-P, which will collect information about particulate matter levels in your proximity and send the data to an Android Application developed for data collection.

The Airspeck-P device can be clipped anywhere to the person as it just needs to be mobile.

You will be provided an Android phone for collection the Airspeck-P data. The only task necessary is that the participant go outside and walk around the designated area. I will give the participant a map of the designated area and verify that the data is being collected appropriately.

At any point in time, if you feel that you do not wish to continue with the study, then please feel free to let me know and the study will be stopped immediately.

What are the potential risks and how have they been mitigated?

You'll be asked to wear the Airspeck-P device. This is a CE-marked device and as part of this process it has undergone the necessary safety tests. There are no significant risks involved in participating in this study.

How will my data be processed?

The data is anonymised and no personal information is stored on the computers. The storage and processing of your data complies with the new GDPR rule. Your data is only identified by your Airspeck-P sensor ID which will be used by multiple participants.

What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance with Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk (<http://www.ico.org.uk>). Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?

If you have any further questions about the study, please contact Ryan Egan (s1876158@sms.ed.ac.uk) or the lead researcher, Professor D K Arvind (dka@inf.ed.ac.uk). In your communication, please provide the study title. If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk.