White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/134435/

Version: Accepted Version

Proceedings Paper:
Makrai, Gabor and Bate, Iain (2018) Signal selection in a complex environmental distributed sensing problem. In: Proceedings of the 13th International Conference on Distributed Computing in Sensor Systems, DCOSS 2017. 13th International Conference on Distributed Computing in Sensor Systems, DCOSS 2017, 05-07 Jun 2017 IEEE, CAN, pp. 155-162.

https://doi.org/10.1109/DCOSS.2017.24

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Signal Selection in a Complex Environmental Distributed Sensing Problem

Gabor Makrai
Real-Time Systems Group
Department of Computer Science
University of York, UK
gabor.makrai@york.ac.uk

Iain Bate
Real-Time Systems Group
Department of Computer Science
University of York, UK
iain.bate@york.ac.uk

***Abstract*—Supporting sustainable development for the urban environment is crucial in the age of rapid urbanisation. Air pollution modelling is one of the key tools for researchers, scientists, and urban planners to understand pollution behaviour. Recent updates in air quality regulations are challenging the state-of-the-art air pollution modelling techniques by requiring accurate predictions on a high temporal level, i.e. predictions at the hourly level rather than the annual level. Current state-of-the-art models designed to have good prediction accuracy on the low temporal resolution by assuming that the pollution is in steady state. Making predictions on higher temporal resolution violates this assumption and causing inaccurate predictions. We introduce a novel statistical regression based air pollution model which produces accurate hourly predictions by using data with high temporal resolution and advanced regression algorithms. We conducted an analysis which shows that the state-of-the-art evaluation techniques (e.g. RMSE) do not describe the nature of the mispredictions of the models built on different data subsets. We carried out an extensive input data evaluation experiment where we concluded that our approach could achieve further accuracy improvement by training the models on a carefully selected subset of the input data.**

I. Introduction

There is a recent interest in understanding the hourly changes of the NO_2 air pollution driven by updates in the air quality directives [1]. Modelling pollution concentration level on high temporal resolution (hourly levels prediction instead of annual levels prediction) is the key to doing that [2], [3]. The current state-of-the-art air dispersion pollution models struggle to make accurate predictions on the hourly level because the models depend on unreliable data (e.g. estimated traffic amount on the roads, vehicle emission data) therefore they are unable to identify significant episodes when the concentration levels are temporarily high [2], [4]. Annual concentration level predictions with these models are close to the observations as uncertainties in the hourly data smooth out on the annual time scale [5] however, models using this data suffer to make accurate hourly predictions.

Statistical regression approaches have been proposed against the state of the air dispersion models to achieve higher prediction accuracy on annual level [6], [7]. These models consider topographical, geographical and pollution-related information around the monitoring location and calculate pollution concentration levels based on these features using a statistical regression algorithm. Land use data (e.g. number of the buildings or the length of the roads around the monitoring stations), however, has a very low temporal resolution which makes the land use regression approach insufficient for hourly concentration level prediction. Simply including high temporal resolution data (e.g. weather data) would result in a complex regression problem [6]. The complex non-linear correlation relationship in the data [8] makes the predictions hard for the traditionally used linear regression and it results in low prediction accuracy using the state-of-the-art regression algorithm [9].

Recent studies [10], [9], [11], [8] use the fundamentals of land use regression methods to improve the prediction accuracy of the standard land use regression model for monthly and yearly concentration level predictions. To achieve the desired accuracy improvement, they use different regression algorithms such as neural network regression [12] or support vector machine regression [13] however, these studies focused only on prediction at lower temporal resolution (monthly and yearly) than hourly concentration level modelling. Applying these algorithms for prediction of hourly concentration levels needs the careful investigation on how the non-linear complex relationship in the data has been exploited to avoid mispredictions.

We propose a novel method for hourly prediction of NO_2 concentration levels which exploits the combination of the usage of complex data and the usage of advanced regression algorithms. This approach has the advantage of discovering statistical patterns in the data which are relevant to the regression problem and does not directly rely on inaccurate datasets.

We implemented this approach for York, United Kingdom where the local council operates five regulatory NO_2 monitoring stations. We used one of the state-of-the-art air dispersion models to compare its prediction accuracy with our statistical regression approach. We created a validation framework to systematically determine the accuracy of each method.

The contributions of this paper are

- Showing that a statistical regression approach can achieve the same or even better prediction accuracy as the state of the art air dispersion methods

- Understanding the data requirements for the statistical regression approaches
- Describing the benefit of context-dependent combinations of statistical models trained on various subset of the input data

The rest of this paper is organised as follows. Section 2 introduces the studies related to our work. Section 3 describes our methodology to develop the framework to validate our model. Section 4 explains the results of the validation and following a discussion about the importance of the results. The conclusion is given in Section 5.

II. Related works

Air dispersion models are modelling the air pollution distribution by using their physical properties combining them with weather conditions. They assume that the distribution of the pollution follows a multi-dimensional Gaussian process [14], [15]. It is the most commonly used method for air pollution modelling and many extensions of this method developed in the past (e.g. ADMS-Urban [14], OSPM with canyon mode for urban canyons [15]). Owen et al. [4] evaluated the ADMS model and concluded that it shows good annual prediction accuracy and prediction correlation with observation data in the London area considering 24 monitoring sites. Hourly prediction evaluation shows considerable errors however they did not investigate the root cause of these errors. Vardoulakis et al. [2] investigated the prediction accuracy of the OSPM model and the model shows good annual concentration level prediction accuracy however it underpredicts hourly concentration levels as it uses incorrect emission inventory data.

Briggs et al. [16] developed a statistical regression approach to annual air pollution modelling. Their method considers topographical, geographical and pollution-related information of the monitoring location and predicted pollution concentration levels based on these features using statistical regression algorithm. The motivation for their method is to determine the most relevant features contribute to the annual concentration level observed by a monitoring station using statistical approaches. They and later studies used the following datasets as input data to the regression: building numbers, building geometry, road length, road geometry, traffic volumes, land use and topographical information. With the most relevant features to the annual concentration level, they could investigate the major contributors to the air pollution in the investigated area. Cyrys et al. [7] developed a similar statistical regression method and their approach could achieve satisfying prediction accuracy in Munich for annual NO_2 levels. Marshall et al. [3] developed a regression model for the Greater Vancouver area and their evaluation shows good correlation to annual observation data.

These methods show good prediction accuracy on annual temporal level but recent studies [17], [5] suggest that this approach would suffer to make accurate prediction on hourly temporal level because the data used to train the regression models only include data has low temporal resolution (e.g. number of buildings, length of road around the monitoring station).

Land use regression models are known to limited only to predict annual and monthly averages, because all the features are insufficient to be able to predict hourly changes of concentration levels. Hoek et al. [17] stated that developing Land Use Regression model which can produce prediction with high temporal and spatial resolution is the interest of study. Isakov et al. [5] indicated that predicting hourly averages of pollutant concentration levels is challenging. They stated that one fundamental problem for predicting hourly averages of concentration levels was to collect data with the necessary temporal resolution but they were not considering the regression algorithm quality used for the prediction.

Recent studies [10], [9], [11], [8] however use the fundamentals of land use regression methods to improve prediction accuracy of the standard Land Use Regression model on monthly and yearly concentration level predictions using different regression algorithms such as neural network regression [12] or support vector machine regression [13].

Tree induction based regression algorithms [18] are powerful tools for regression problems. They were applied in the past with success to learn the relationship between the input data and the observations. Tso et al. [11] used decision tree regression technique to predict electricity energy consumption and compared it with other algorithms. They reported that this algorithm has the advantage of using complex datasets from different data sources and can discover hidden patterns in the data. Ensembling the decision tree regression algorithm (such as the widely used random forest regression [19]) was used for predicting yearly averages of NO_2 by Champendal et al. [8]. They reported good prediction accuracy against standard linear regression methods.

Our proposed method uses the idea of the statistical regression models for hourly NO_2 concentration level predictions. Using this technique allows us to avoid the direct usage of data where uncertainties (e.g. vehicle emission inventory database) outweigh the benefit of using it. This technique, however, was not designed to make hourly predictions and simply applying high temporal resolution data results in a complex regression problem. To tackle this issue, we propose the usage of tree induction based regression algorithms instead of the state-of-the-art linear regression algorithm.

III. Methodology

The initial step of implementing an air pollution model is collecting the necessary data. This step provides the raw data for the modelling task. We mainly used publicly available datasets to make our work easily reproducible. The second step is transforming the raw data into a format which makes the data processable for the algorithms. We used similar data transformation steps to other studies [9], [5], [20], [8], [6]. The modelling and evaluation are the last steps which allow us to systematically determine the accuracy of the different methods. We implemented the state-of-the-art validation framework. We introduce the details of these steps in this section.

Fig. 1. Heworth monitoring station (left), map of the modelling area (centre), Buffer area for the Fishergate monitoring station (right)

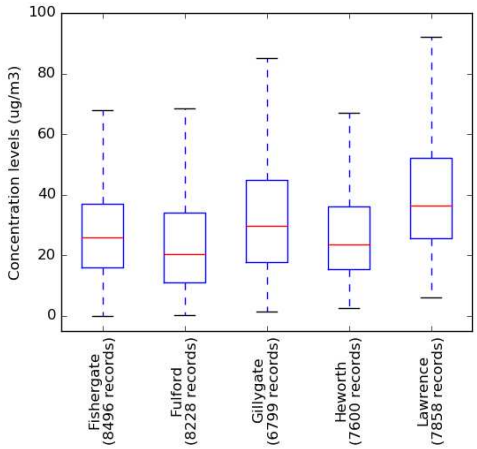

Fig. 2. Distribution and availability of the monitoring data

A. Monitoring and other data

1) Monitoring data: The most important data is the NO2 concentration level data. The City of York Council (CYC) operates a network of high precision (chemiluminescence-based) instruments in York to monitor the air quality. We acquired monitoring data from 5 stations between 1st January 2013 and 31st December 2013. Figure 1 shows the area of interest with the location of each monitoring station (red stars). Figure 2 shows a boxplot of these measurements for each station.

2) Traffic data: Traffic data was provided by the City of York Council's Transportation Managment Group. This group operates a network of automated traffic counters (ATC) across the city. For our modelling method, we selected the five units which co-located to the NO2 monitoring stations. The instruments only provide a simple vehicle count and cannot give detailed information about different vehicle types (such as cars, LGVs, HGVs, buses, etc.).

3) Road data: We collected road data using the Open Street Map database which contains detailed lane information about each road segment (e.g. lane numbers, allowed directions, speed limits).

4) Building data: We acquired building data from the Ordnance Survey's 2009 version of Mastermap^{TM} Topography layer. This layer gives spatial information (e.g. geometry, surface area, etc.) about buildings within the area of interest.

5) Land use data: We collected land use data from the Open Street Map database. The available data describes the areas (in polygons format) usage scenarios (e.g. leisure, green areas, farm, etc.).

6) Meteorological data: We acquired meteorological data from the Weather Underground database using its API to download data. This database contains observations for cities and includes temperature, relative humidity, wind speed, wind direction, and pressure measurements.

7) Time related data: Time-related indicators (e.g. hour of the day, day of the week, bank holiday, etc.) for our statistical regression model are important because the regression models can use this information to discover temporal patterns in the input data. We also included some York specific event indicator such as event (e.g. York horse races when tens of thousands of visitors come to the city leading to significantly higher traffic volumes than the normal at the certain time of day) indicator which affects the traffic pattern in the whole city.

B. Data preprocessing

The core idea of the statistical regression approaches [6] is to extract information around the monitoring station. We executed the following steps to transform the available data into useful data for the regression models.

First, we created a 100-meter wide rectangular area (called the buffer area) and extracted all the spatial information for each monitoring station. We followed the guidelines (size and extraction technique) of many previous studies to create the buffer areas e.g. as followed by [16]. Using the available road data, we extracted the feature "road_length" that represents the

amount of the road in the buffer area. In addition, we generated another road data feature called "lane_length" which weights the road with their lane numbers (so multi-lane roads gives more value to this feature). We processed the building data and calculated the number of the buildings and area of the buildings covered by each buffer area and generated "buildings" and "buildings_area" features. We used the available land use data to find out the area of the used land and leisure spaces in the buffer area and we generated the "landuse_area" and "leisure_area" features. After we merged all the generated features for each station based on the stations' locations, we generated an hourly timestamp feature runs from 1st January 2013 to 31st December 2013 and multiplied the dataset to give all the timestamps for each station. In the final step, we merged the weather data and the time-related data based on the hourly timestamp. Table 1 shows a summary about the generated dataset.

Feature	Unit	Source	Data group
no_2 level	ug/m3	CYC	-
road_length	meter	Open Street Map	R
lane_length	meter	Open Street Map	R
buildings	-	OS Mastermap	B
buildings_area	area	OS Mastermap	B
landuse_area	area	Open Street Map	L
leisure_area	area	Open Street Map	L
atc	vehicle/hour	CYC	A
wind_direction	degree (angle)	Weather Underground	W
wind_speed	m/s	Weather Underground	W
temperature	celsius degree	Weather Underground	W
rain	indicator	Weather Underground	W
pressure	hPa	Weather Underground	W
hour	-	Generated	T
day_of_week	-	Generated	T
month	-	Generated	T
bank_holiday	indicator	Generated	T
race_day	indicator	Generated	T

TABLE I
SUMMARY OF THE COLLECTED DATA

C. Methods

We implemented a validation framework to determine the general accuracy of the proposed model against the state-of-the-art. This framework consists the state-of-the-art location based leave one out cross validation (LOOCV) [6], [7], [3]. This validation method allows to build the statistical model using data from four stations and validate the accuracy comparing the prediction of the model at the location of the fifth station and the observation of the fifth station. Within the cross-validation method, we applied the root mean squared error ($RMSE$) as the accuracy indicator similarly to [7], [3], [20]. Root mean squared error ($RMSE$) is defined by the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{\forall i}(y_i - \hat{y_i})^2} \quad (1)$$

where n is the number of the observations, y_i is the observed target value, $\hat{y_i}$ is the prediction produced by the model.

We run the WinOSPM 5.1.90 software to determine the accuracy of one of the state-of-the-art air dispersion model for our modelling scenario. This is the latest implementation of the OSPM model developed by the Department of Environmental Science at Aarhus University [15]. This model only needs weather data and manual entry of street geometry around the modelled area. All the other data were already prepared in the software package (e.g. emission inventory data, sun radiation data, etc.).

We propose to use tree induction based regression algorithms for hourly NO_2 concentration level predictions. We include two of these algorithms: the decision tree regression and the random forest regression algorithms. We used the scikit learn library [21] to implement and run the proposed statistical method as well as other statistical regression approaches. Using this library gives the advantage of using a well-established and extensively tested implementation of the required machine learning algorithms.

Linear regression [22] is a method to create prediction based on the following equation: $\hat{y}(w,x) = w_0 + w_1x_1 + ... + w_mx_m$, where $\hat{y}$ is the prediction for the input feature vector $x = \{x_1, x_2, ..., x_m\}$, x_i are the features, w_0 called the intercept and w_i are the coefficients. There are multiple ways to calculate weights and scikit learn framework uses the Ordinary Least Squares optimization where it solves the mathematical equation $argmin(\sum_{\forall x_i \in X}(\hat{y_i}(w,x_i) - y_i)^2)$ where $X = \{x_1, x_2, ..., x_n\}$ is the set of the feature vectors of the observations and y_i are the target value for each observation. Linear regression can only discover linear relations between the target value of the observation and the features, however, these relations (represented by the coefficients) can be ranked and described very well if the input data is normalised. This property of the algorithm established its popularity, because researchers could understand the main factors of predictions. The sickit learn framework implementation of this algorithm has two parameters: "intercept" what we can choose to include the intercept to the equation and "normalize" which forces to train algorithm to normalize the data before the training. We refer to this method as *LR*.

Decision tree regression [18] is a decision tree induction based regression technique where tree induction algorithms create a decision tree and every leaf of this tree contains a prediction value and every other internal node has decision criteria (for example $x_4 < 0.5$). The decision tree is built to have the best fit for the training dataset and every prediction starts at the root, evaluates it, then decides to take the left or right children (if it is a binary decision tree) then evaluate all the internal node until it ends at a leaf node where there is a prediction value. The implementation of the tree induction algorithm in the scikit learn library contains a parameter ("leaf") which stops the tree induction if an internal node has only the given number of observations (this is a general technique to avoid overfitting called back-pruning technique). We refer to this method as *DTR*.

Random forest regression [19] is an ensemble method based on the decision tree regression. Instead of training one large

decision tree for the regression, it follows the idea of the ensemble methods where the algorithms train models (the parameter ”n_estimators” defines the number of the models) on the subset of the train data (in terms of observations as well as features) and rank the created sub-models on the efficiency based on the other part of the training data. With this procedure, the method can randomly pick up an interesting part of the data and have a large number of efficient sub-models. The prediction is based on a voting procedure, where each sub-model has a vote and based on their weighted average the final prediction is calculated. We refer to this method as *RFR*.

We started to evaluate the OSPM method with the implemented validation framework first. The accuracy of this model provides a baseline for the statistical regression methods

The first regression approach is only using the LR algorithm and the land use related data (precisely the road, the building, and the land use data). Investigating such an approach can provide information about the difficulties of predicting NO2 with existing methodologies [6]. The second regression approach is using the LR algorithm with all the available data. Analysis of the accuracy of this approach can provide details of prediction difficulties of the LR algorithm facing complex non-linear data. Tree induction based regression techniques have been already sued to make predictions on complex non-linear data. We included the decision tree regression and the random forest regression algorithms to determine the accuracy of these methods on this environmental prediction task. In the last step, we investigated the data requirements for the statistical regression methods. It is not clear that which data sources provide the most relevant data for the regression approach. Also, it is unknown what is the quality of the collected data. Using only the most relevant data can increase the accuracy as the algorithms do not have to deal with data contains errors, outliers, anomalies which could lead to mispredictions.

IV. Evaluation

We introduce and discuss the results of the validation framework. Firstly, we focused on the analysis of the accuracy of the described methods. Secondly, we investigated the data requirements for the proposed regression method. Lastly, we analysed the nature of the prediction error using the proposed regression method generated on different data subsets.

A. Prediction accuracy of the different air pollution modelling approaches

We executed a grid parameter search to tune each statistical regression approach for this regression problem. We used the normalise and intercept options for the LR algorithm, leaf = 15 for the DTR algorithm and leaf = 9, n_estimator = 59 for the RFR algorithm.

Figure 3 shows the summary of the outcome of the validation framework. As Briggs et al. [6], and Hoek et al. [17] assumed, using the LR algorithm with land use related data introduces more error in the prediction than the state-of-the-art methods (OSPM in our case). Using data with high temporal resolution (weather and time-related data) does not help as the data becomes more complex and the LR algorithm suffers to make accurate predictions.

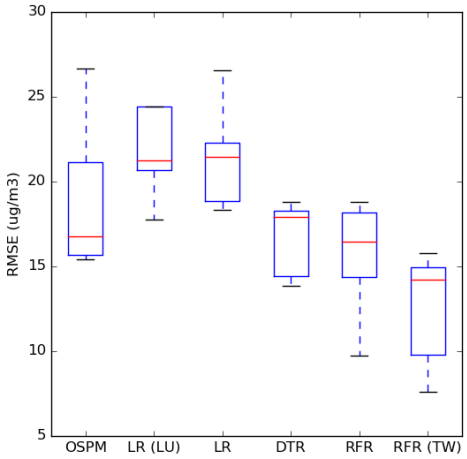

Fig. 3. RMSE error plot of the investigated methods

Using the complex data with the DTR method, however, resulted in an air pollution model which provides the same accuracy as the OSPM model. Using the RFR method increases the accuracy even further. These algorithms could exploit their ability to discover the complex non-linear relationship in the data which makes them appropriate for this regression task as it produced more accurate predictions than the state-of-the-art method.

B. Data requirement analysis of the statistical regression approach

As RFR shows the best accuracy, we were interested in determining which data (combination of the different subsets of the originally available data) is the most relevant for this algorithm. We divided the data into smaller data categories based on the data sources: ”R”, ”B”, ”L”, ”A”, ”W”, ”T” represent the road data, building data, land use data, ATC data, weather data and time-related data, respectively. We then run our validation method for each combination of the data groups.

The best data subset is the combination of time (”T”) and weather (”W”) related data only according to our experiment (Figure 3). This result, however, was unexpected as the RFR model was trained on 4 stations data still the most accurate model only uses T+W data which are global and have the same information at each station.

Figure 4 shows the relative RMSE error analysis of each combination to the case when we only applied T+W data for the RFR method. There are only a few cases where adding more data could partially increase the accuracy, but the overall (mean of the error of the other data combinations) accuracy is always worse than the T+W case.

We discovered the same trend during the individual investigation of all the 64 combinations. Figure 5 shows that not

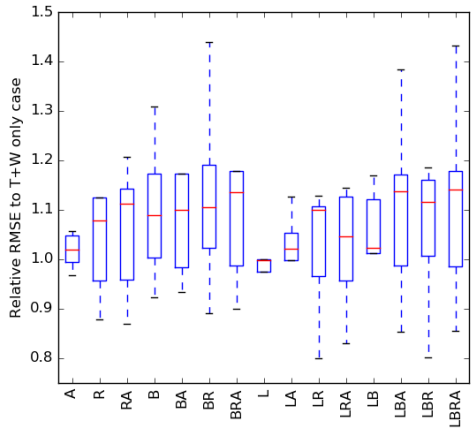

Fig. 4. RFR+TW method RMSE accuracy considering other input data subsets

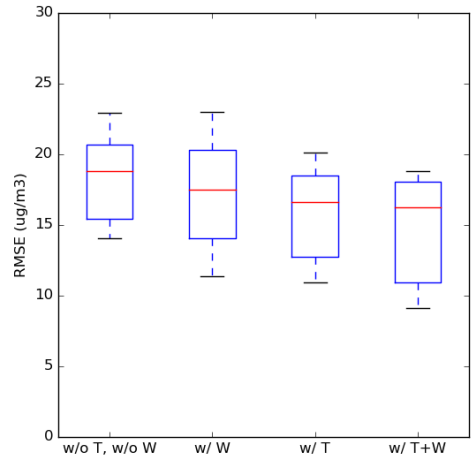

Fig. 5. RFR method RMSE accuracy considering different input data subsets

using T+W data introduces error for the prediction. Using only the weather ("W") or time ("T") related data helps to make more accurate predictions and using T+W data resulted always in the most accurate predictions.

These results tell us that our proposed statistical regression approach can achieve better accuracy if we can carefully subset the input data and use only the time and weather-related data for our modelling scenario.

C. Detailed error analysis of the proposed method

As the main pollution source is the traffic according to the local air quality experts, we started to focus on the usage of ATC ("A") data. Adding ATC to the RFR+TW model resulted in increased RMSE error (RFR+TW provides RMSE of 12.79 ug/m3 meanwhile RFR+TWA provides RMSE of 13.45ug/m3). This is the reason why our previous data optimisation procedure did not select the ATC data. It is important to note that the general RMSE does not give any details about the nature of the errors. To understand the error introduced by the use of the ATC data, we analysed the absolute prediction error and compared the results of the RFR+TW and RFR+TWA models. Interestingly, both models produced error episodes where within short time windows, they couldn't make accurate predictions, however, these error episodes of the two models do not overlap. Figure 6 shows an example where the RFR+TWA model has smaller prediction errors during the week between 22nd July 2013 and 28th July 2013 at the Fulford monitoring station.

Figure 6 shows that the RFR+TWA model makes more accurate predictions than the RFR+TW model in some cases. Having established that RFR+TW and RFR+TWA models make non-overlapping error episodes, the next objective was to investigate the accuracy of the RFR+TW and the RFR+TWA models in some specific prediction situations. To do this, we analysed the predictions according to specific rules, because it allows the systematic assessment of the prediction error of the two models. To find such rules, we used our prior knowledge about the modelled area. In general, the RFR+TW model provides the most accurate predictions, however, it does not use information about the traffic. In cities, traffic peaks twice a day when commuters flood the roads (so they called morning and afternoon traffic peak period). We then separated two different time windows focusing on days where the weather does not effect the pollution (e.g. the wind speed is low):

- *morning*: before the morning traffic peak period, when the pollution has been cleaned out during the night (4AM-7AM)
- *afternoon*: during the afternoon traffic peak period, where traffic is high on the roads and traffic jams are highly likely (4PM-7PM)

Figure 7 shows the results of analysis of absolute error in prediction during these time windows using the model RFR+TW, RFR+TWA, and RFR+WA. We included RFR+WA for this analysis to investigate the accuracy of a model which does not have information about the time-related data. In the morning case, there is no benefit of using more data than the T+W. Using RFR+TWA model, however, shows less error in prediction when the traffic is peaking (afternoon case). Moreover, in this situation, using time-related ("T") data does not show relevance as the RFR+TWA and RFR+WA show similar prediction accuracy.

This result motivates the usage of complex modelling system where multiple random forest statistical regression models are being trained on different subsets of the input data and a model selector decides what model to use in which situation to exploit the non-overlapping error episodes of the different models.

V. Conclusion

In this paper, we proposed a novel statistical regression approach for hourly NO_2 concentration level modelling. This model exploits the random forest regression algorithm and it

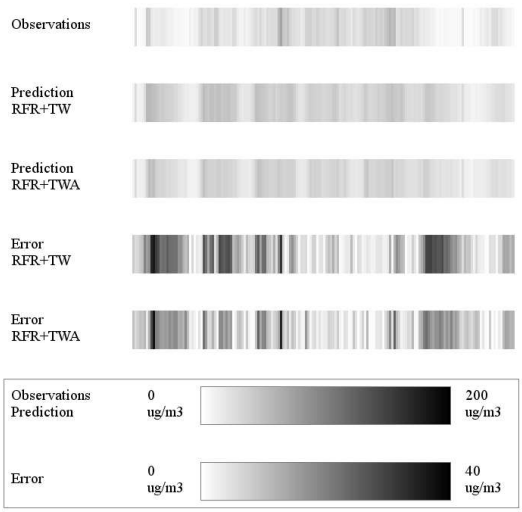

Fig. 6. Heatmap of observations, predictions, and errors at the Fulford station between 22nd July 2013 and 28th July 2013

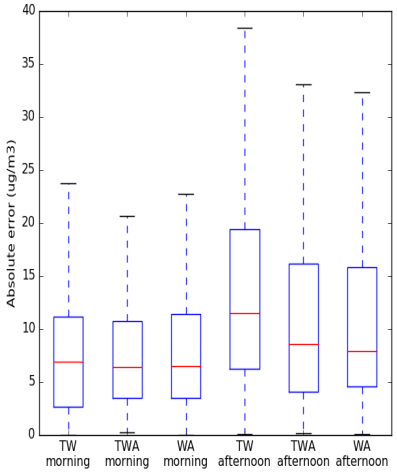

Fig. 7. Absolute error plot of RFR+TW, RFR+TWA, and RFR+WA at the morning and afternoon time windows

uses complex input data to make accurate predictions. We compared our method against the state of the art air dispersion method. During the evaluation, we confirmed the challenges of using statistical regression methods for this prediction problem [17], [5]. The results show that the proposed method produces more accurate predictions than the state-of-the-art model in our modelling scenario. Using only the time and weather-related data to generate the random forest regression model led us to more accurate predictions than using all the available data. This result motivated us to further investigate to prediction errors. The detailed analysis of the prediction errors of the RFR+TW and RFR+TWA models revealed that RFR+TWA model makes more accurate predictions in some situations despite the of the larger general RMSE error of the RFR+TWA model. To exploit the non-overlapping error episodes produced by the two models, we investigated two scenarios where we concluded that using the RFR+TWA or RFR+WA models consistently provide more accurate predictions on hours where the traffic is peaking than the RFR+TW model.

Our results show that our statistical regression approach trained on different subsets of the input data (RFR+TW, RFR+TWA, etc.) produced different error episodes. We will investigate the development of the stacking of these models where the stacking procedure generates the models on different input data subsets (not only a few variations, e.g. RFR+TW,RFR+TWA) systematically and this procedure will find the appropriate rules to use the best model from the existing model set to make all the hourly predictions as accurate as possible.

Moreover, integrating this approach into a Geographic Information Systems (GIS) can give a more accurate modelling tool for urban city planners to make better decisions considering the environmental effect of the urban processes and it gives them a better understanding of the air pollution of the modelling area.

VI. Acknowledgement

Authors would like to thank Francesco Pilla and Steve Cinderby for the helpful comments during the development of the work.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 608014.

References

[1] W. H. Organization et al. Air quality guidelines for europe. 2000.

[2] S. Vardoulakis, M. Valiantis, J. Milner, and H. ApSimon. Operational air pollution modelling in the UK - street canyon applications and challenges. *Atmospheric Environment*, 41(22):4622–4637, 2007.

[3] J. D. Marshall, E. Nethery, and M. Brauer. Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmospheric Environment*, 42(6):1359–1369, 2008.

[4] B. Owen, HA. Edmunds, DJ. Carruthers, RJ. Singles. Prediction of total oxides of nitrogen and nitrogen dioxide concentrations in a large urban area using a new generation urban scale dispersion model with integral chemistry model. *Atmospheric Environment*, 34(3):397–406, 2000.

[5] V. Isakov, M. Johnson, J. Touma, and H. Özkaynak. Development and evaluation of land-use regression models using modeled air quality concentrations. In *Air Pollution Modeling and its Application XXI*, pages 717–722. Springer, 2012.

[6] D. J. Briggs, C. de Hoogh, J. Gulliver, J. Wills, P. Elliott, S. Kingham, and K. Smallbone. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of the Total Environment*, 253(1):151–167, 2000.

[7] J. Cyrys, M. Hochadel, U. Gehring, G. Hoek, V. Diegmann, B. Brunekreef, and J. Heinrich. GIS-based estimation of exposure to particulate matter and No2 in an urban area: stochastic versus dispersion modeling. *Environmental Health Perspectives*, pages 987–992, 2005.

[8] A. Champendal, M. Kanevski, and P.-E. Huguenot. Air pollution mapping using nonlinear land use regression models. In *Computational Science and Its Applications–ICCSA 2014*, pages 682–690. Springer, 2014.

[9] A. S. Sánchez, P. G. Nieto, P. R. Fernández, J. del Coz Díaz, and F. J. Iglesias-Rodríguez. Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area (spain). *Mathematical and Computer Modelling*, 54(5):1453–1466, 2011.

[10] M. Gardner and S. Dorling. Neural network modelling and prediction of hourly nox and no2 concentrations in urban air in london. *Atmospheric Environment*, 33(5):709–719, 1999.

[11] G. K. Tso and K. K. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, 2007.
[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.
[13] A. J. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
[14] R. Hosker Jr. Consequences of effluent release. *Environmental Research Laboratories*, page 147, 1975.
[15] R. Berkowicz. Ospm - a parameterised street pollution model. In *Urban Air Quality: Measurement, Modelling and Management*, pages 323–331. Springer, 2000.
[16] D. J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. van Reeuwijk, K. Smallbone, and A. Van Der Veen. Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographical Information Science*, 11(7):699–718, 1997.
[17] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric environment*, 42(33): 7561–7578, 2008.
[18] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
[19] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
[20] J. Gulliver, K. de Hoogh, D. Fecht, D. Vienneau, and D. Briggs. Comparative assessment of gis-based methods and metrics for estimating long-term exposures to air pollution. *Atmospheric Environment*, 45(39):7072–7080, 2011.
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.
[22] Weisberg, S.: Applied linear regression, vol. 528. John Wiley & Sons (2005)