# Air Quality Prediction in smart cities using Porto as a case study

Luís Maia and Nádia Soares

***Abstract*— This work focus on predicting the concentration of the air pollutant Particulate Matter 1 (PM1) from meteorological and pollution data available for the city of Porto. Two years of data containing 13 variables have been collected at 23 different locations in the city. This work presents an exploratory study of the data and the process of training a model for predicting this air pollutant indicator, using random forest. We present a comparison between different methods, specially neural networks which are the state of the art for air quality prediction, but did not behave so well in this particular dataset. An error of 17.6% was achieved on the testing set using the holdout cross validation method.**

## I. INTRODUCTION

Air pollution represents a growing problem and a serious risk to human health, particularly in large, cities were traffic and industrialization cause increasing numbers of pollutants in the air. In *Air pollution: A public health approach for Portugal* [1] was shown that in most of the northern part of the country there is an increasing concentration of nitrogen dioxide (NO2) and particulate matter (PM). This facts and the recent growth of the city lead us to believe this study can be relevant and well-timed.

In particular, PM, a complex mixture of airborne particles, are of special importance because they are associated with increased mortality and morbidity rates, as their small diameter allows them to penetrate into the alveoli as far as the bronchioles [2][3][4].

They can be directly emitted, for instance, when fuel is burnt and when dust is carried by wind, or indirectly formed, when gaseous pollutants previously emitted to air turn into particulate matter. Their concentration is influenced by place, season and weather conditions [5].

This report first introduces the problem at hand and outlines the project goals and its contributions to the field. The state of the art for air quality prediction is presented. It then describes the dataset, and the steps taken to prepare it. We move on to describe the exploratory data analysis performed, the experimental design and the model created to predict new data points. The results are presented and conclusions are drawn. Finally, we present future steps and improvements.

## II. OBJECTIVES

The aim of this study is to predict the value of particulate matters of size around $1\,\mu\mathrm{m}$ (PM1) given meteorological and pollutant variables. This problem was formulated as a prediction problem.

The first objective is to understand the data by exploring the interactions between the variables, specially, how they are correlated. We want to understand which variables have a higher influence in the quantities of PM1.

Then, we want to create a good model for predicting the quantity of this pollutant in the air. Some alternatives were explored and the chosen algorithm was Random Forests.

### A. Contributions

Most of the studies we found were done by fitting regressions and neural networks. We compare our random forests model to neural networks and others, and it has a much better performance for this particular dataset. Random forest is a robust algorithm and we argue that it can be a better choice to explore this current problem. Since it features greater interpretability than neural networks, it can help shed light into the specific variable interactions and processes that cause the quantities of pollutant particles to be high our low.

Furthermore, as far as we know, this type of study has never been done for the city of Porto. And, even though the presence of PM1 in the air has grave consequences to human health due to its small size, it typically is not a factor in this studies. Only particles of around around $2.5\,\mu\mathrm{m}$ or $10\,\mu\mathrm{m}$ tend to be studied.

## III. RELATED WORK

With the evolution of IoT and smart cities more data has become available, and machine learning techniques have shone a new light the into air quality prediction problem. Studies have been done using data from specific cities like Murcia, Spain [6] and Milan, Italy [7]. Some previous works have formulated the problem of predicting the hourly concentration of pollutants like O3, PM and SO2, as a classification problem, or as a multi-task learning (MTL), using meteorological data of previous days [8]. Deep learning has been used mostly with shallow networks, using some lagged features to handle the temporal dimension of the data [7]. These methods neglect to handle the spacial-temporal component of this datasets. A high quantity of particulate matter in a given station at time *T*, may give clues to predict that same variable at time *T+1* at a different location, depending also on factors like wind speed and direction. More complex forms of neural networks have started to be used fairly recently for this task. [9][10][11] and [12] all did experiments with artificial neural network, convolutional neural network, and a long-short-term memory to extract spatial-temporal relations.

## IV. URBANSENSE DATASET

The chosen dataset comes from a project on smart city in Porto, Portugal [13]. A set of sensors was placed at

different locations in the city and the result was a dataset of 13 variables, plus the date and location, measured at 23 locations. The variables are CO, humidity, luminosity, NO2, noise, O3, particles1 (particulate matter with size around $1\,\mu m$), particles2 (particulate matter with size around $2.5\,\mu m$), precipitation, solar radiation, temperature, wind direction and wind speed. There are more than two years of data sampled at approximately each minute. To simplify, we chose to focus on one location, *Campo 24 de Agosto*.

The dataset has several problems. Not all variables are available for all locations and some values are missing for several hours or days. There are separate files for each location and for each variable, so the first task was to unify the separate files into one dataset and clean the data.

After importing and merging the dataset for *Campo 24 de Agosto*, we had 1378550 observations, but only 10340 are complete observations. The start date is 25-06-2015 and the end date is 26-08-2017. The variables are described in Table I in the appendix.

The steps taken to prepare the dataset were: imputing missing values, standardizing the data and downsampling it to the hour. A visual representation of the statistical measures after this pre-processing can also be found in the appendix, in Figure 9.

### A. Data Preparation

*1) Missing values:* There were some values missing at random, possibly caused by glitches in the sensors. Also, there were entire sections of time with no values.

Since we used only one location, removing the missing values could chop away a big part of the dataset, and, since the data is sequential, we cannot simply remove pieces of it.

Another option considered was to use the local mean: using the mean of the previous and next instances to fill in the missing value. This was also not a viable option as we had big chunks of missing value and local mean would not be possible to calculate.

After considering all of this, we imputed them using K-nearest neighbor with k equal to 5.

*2) Standardization:* Random forest itself is not sensitive to non standardized variables, but in order to compare its results to algorithms that are, the values were scaled to the interval between zero and one.

*3) Downsampling:* The first few observations of the dataset were sampled at a constant rate of 1 minute, but as the time passed, the sample rate became very inconstant and some entries were skipped.

Merging the different variables in this condition requires some sort of resampling.

For that reason we chose to downsample the values to the hour. For that we used the mean of highest and lowest values in an hour time frame.

This step was done after imputation because with the missing values we could not know how to glue the variables together. After the entire process we ended up with a tidy dataset of 11249 observations and 13 variables. The data is now ready to be explored.

## V. EXPLORATORY DATA ANALYSIS

The distribution of the PM1 variable is shown in Figure 1.

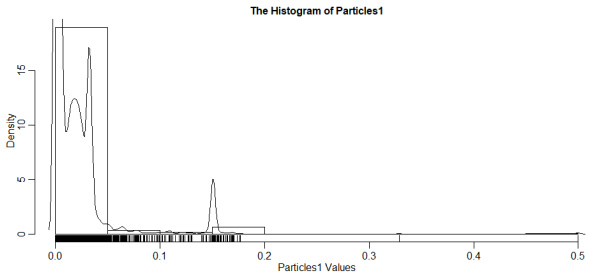


Fig. 1: PM1 distribution

Figure 2 is a plot of the PM1 over time, for one complete year, 2016. It looks like particle counts are higher during the summer months.

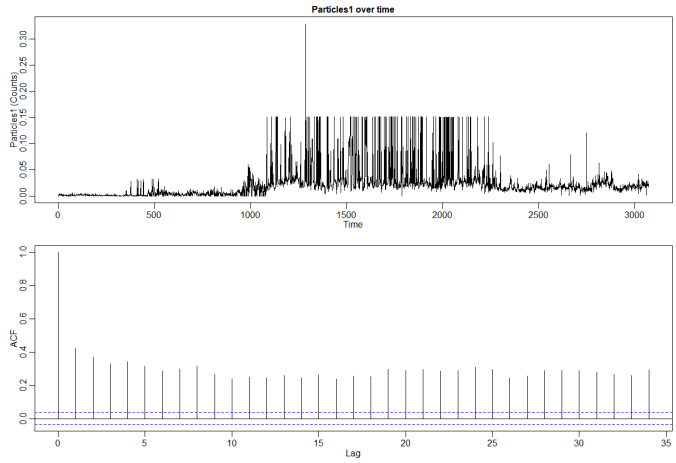


Fig. 2: a) PM1 over time for the year of 2016 and b) its Auto Correlation Function

Figure 3 shows the correlation between variables. Notice that Particles1 shows a very high correlation with Particles2.

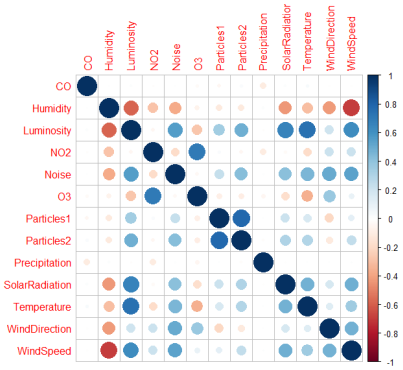


Fig. 3: Correlation for the location *Campo 24 de Agosto*.

The relationship between Particles1 and the two variables with highest correlation are shown in Figures 4 and 5.

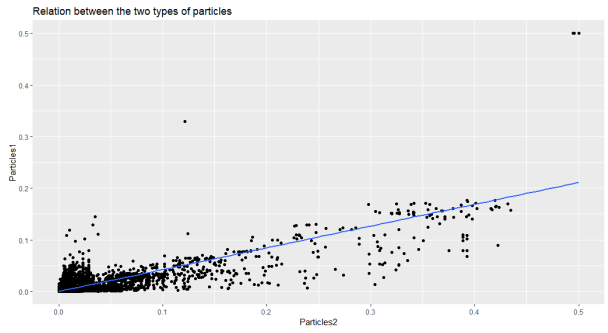


Fig. 4: Relationship between Particles1 and Particles2.

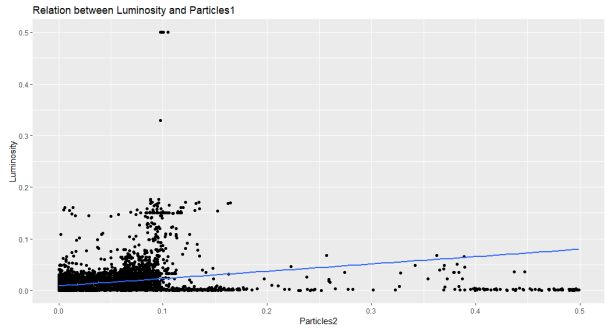


Fig. 5: Relationship between Particles1 and Luminosity.

## VI. EXPERIMENTAL DESIGN

The models created in this project were trained the Hold-out Method with 70% of the data used for the training set and the remaining 30% for the test set. There was no need to apply the nested cross validation method proposed in the previous report since no time forecasting was done.

The metrics used to evaluate the algorithms performance were the mean absolute error (MAE), the symmetric mean absolute percent error (SMAPE) and a percent error calculated from the SMAPE to make it more readable.

### A. Symmetric Mean Absolute Percent Error

We initially decided that we would use the Mean Absolute Percent Error (MAPE), but when the variables are near 0, MAPE goes to minus infinity. This is a problem since our scaled variables are sometimes near 0.

$$MAPE = \left( \frac{1}{n} \sum \frac{|Actual - Predicted|}{|Actual|} \right) \times 100$$

The SMAPE is an alternative to MAPE. Its values go from 0% to 200%, reducing the influence of these small items.

$$SMAPE = \frac{2}{n} \sum \frac{|Actual - Predicted|}{Actual + Predicted}$$

A percentage error was calculated from SMAPE to make it easier to interpreter.

$$PercentageError = \frac{SMAPE * 100}{200}$$

## VII. MODELING

We started by designing a model based on the state of the art model for this type of problem: neural networks.

### A. Neural Network

For that we chose the 5 variables with higher correlation to Particles1: Particles2, Noise, Luminosity, Temperature and Solar Radiation.

As this is a prediction problem with 5 variables, the architecture chosen for the neural network was 5 input nodes 1 hidden layer with 3 nodes and one output node.

Different activation functions were tested and the number of epoches was optimized for each of them as shown in Table I.

| Activation | MAE | SMAPE | Epochs |
|---|---|---|---|
| Rectifier | 0.006400 | 0.812 | 120 |
| Tanh | 0.006481 | 0.823 | 220 |
| TanhWithDropout | 0.012340 | 0.970 | 160 |
| Maxout | 0.006556 | 0.828 | 110 |
| MaxoutWithDropout | 0.008818 | 0.914 | 100 |
| RectifierWithDropout | 0.010255 | 1.010 | 180 |

TABLE I: Error metrics for different activation functions and the number of epoches needed for conversion.

Note that the mean of scaled Particles1 variable is 0.016400 on the training set and 0.015406 on the testing set.

The learning rate is set to be adaptative. Tests were done for various fixed learning rates but the adaptative had better results.

Initially, we wanted to focus on neural networks because adapting them to a recurrent or long-short term memory neural network would allow us to take advantage of the temporal nature of the data. They also seem to be the state of the art in air quality prediction literature, but, after testing other methods we found that the Random Forest algorithm gave much better results. We chose to train and optimize the random forest instead. A table showing the performance of the different algorithms tried is shown in the results section.

### B. Random Forest

The first model trained used the default parameters and the same set of variables picked for the neural network. It had 0.004679 MAE and 0.596 SMAPE, 30% error.

Since Random Forest does feature selection by itself, we trained a new model using all features so it could pick up the best ones.

The results improved as can be seen in Table II.

The features with higher importance were Particles2 and Luminosity which we had already chosen due to the high correlation to Particles1. But the algorithm chose features like CO, NO2 and O3 which we had not considered.

| | MAE | SMAPE | Percentage error |
|---|---|---|---|
| Training | 0.001707 | 0.329 | 16.5% |
| Testing | 0.003746 | 0.509 | 25.5% |

TABLE II: Results for random forests with all variables and default parameters

To improve its performance even further, two parameters can be tune: mtry, the number of variables randomly sampled as candidates at each split, and ntree, the number of trees to grow.

In terms of the number of trees, the default, 500 is too high as can be seen in Figure 6, showing the out-of-bag error convergence by the number of trees generated.

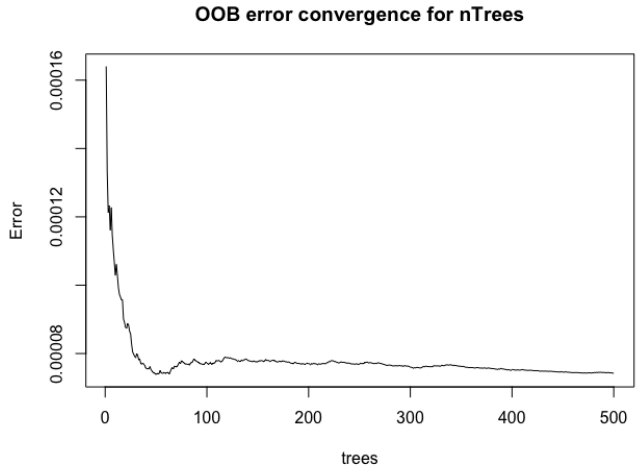


Fig. 6: Evolution of the out-of-bag error as the number of trees increases

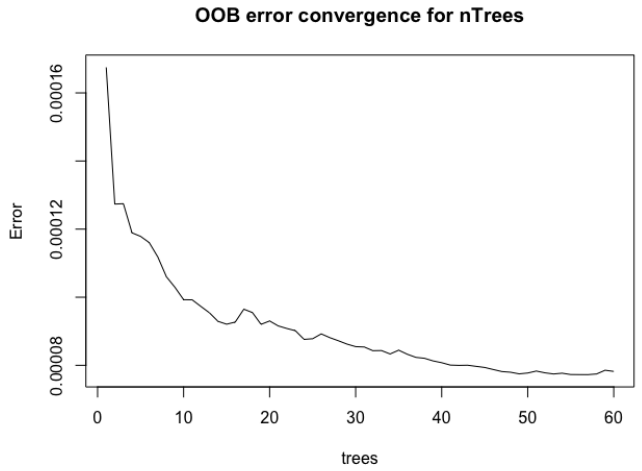


Fig. 7: Evolution of the out-of-bag error as the number of trees increases for 60 trees

The out-of-bag (OOB) error is a metric generated by the random forest algorithm internally during the run.

Decreasing the number of trees to 60 did not decrease the percentage of variance explained significantly and the error stayed the same for training and testing.

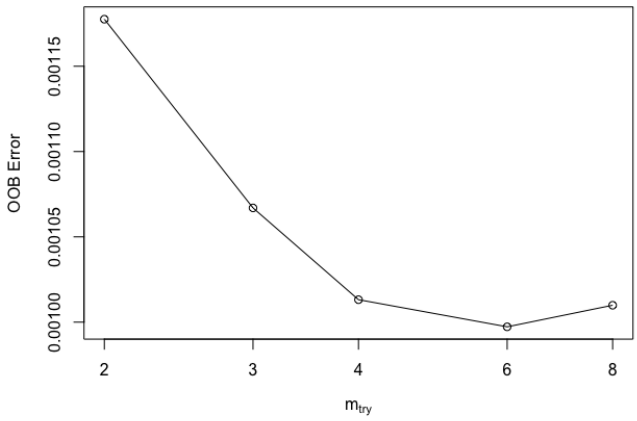


Fig. 8: Result from tuneRF showing the change in out-of-bag error with the number of variables

Using the tuneRF function from the randomForest package in R, we estimated that the optimal number of predictor variables is 6, as shown in Figure 8.

| | %IncMSE | IncNodePurity |
|---|---|---|
| CO | 16.378401 | 0.09336372 |
| Humidity | 11.309020 | 0.09282447 |
| Luminosity | 7.490452 | 0.68077522 |
| NO2 | 9.689913 | 0.05379036 |
| Noise | 8.679774 | 0.27691128 |
| O3 | 6.622210 | 0.17748574 |
| Particles2 | 42.631891 | 5.21838442 |
| Precipitation | 2.638855 | 0.09257108 |
| SolarRadiation | 7.082862 | 0.78347938 |
| Temperature | 8.363992 | 0.12679521 |
| WindDirection | 6.337508 | 0.08584922 |
| WindSpeed | 4.769444 | 0.19626314 |

TABLE III: Variable importance for random forest with mtry=6 and ntree=60

The table of variable importance using only 6 variables is shown in Table III.

This optimization improved the results even further: MAE was 0.002843 and SMAPE was 0.352, 17.6% error.

| | MAE | SMAPE | Percentage error |
|---|---|---|---|
| Training | 0.001215 | 0.216 | 10.8% |
| Testing | 0.002843 | 0.352 | 17.6% |

TABLE IV: Results for random forests with mtry=6 and ntree=60

The last attempt to improve the performance of the model was to try to apply the principal component analysis to the features.

From the original 12 preditor variables, PCA returned 10 components.

After tuning and training a new random forest, we obtained a model with mtry = 10 and ntree = 110.

Unfortunately, it did not perform better on the test set, as shown in Table V.

| | MAE | SMAPE | Percentage error |
|---|---|---|---|
| Training | 0.001931 | 0.366 | 18.3%% |
| Testing | 0.004527 | 0.597 | 29.9% |

TABLE V: Results for random forests after PCA. mtry=10 and ntree=110

## VIII. EVALUATION AND RESULTS

The results from the models discussed above are displayed in Table VI, alongside other algorithms that were trained for comparison. Notice that SVM also had a good result.

| Algorithm | MAE | SMAPE | Percentage error |
|---|---|---|---|
| NN with Rectifier | 0.006400 | 0.812 | 40.6% |
| Linear Regression | 0.007610 | 0.946 | 47.3% |
| rpartXse (Decision Tree) | 0.006114 | 0.824 | 41.2% |
| SVM | 0.005876 | 0.767 | 38.4% |
| Bagging | 0.007010 | 0.881 | 44.1% |
| RF - selected variables | 0.004679 | 0.596 | 30% |
| RF - all variables | 0.002843 | 0.352 | 17.6% |
| RF - PCA | 0.004527 | 0.597 | 29.9% |

TABLE VI: Comparison of the performance of different algorithms using the same train and test sets.

We followed the same optimization process and trained models for 3 different locations, one models for each. Table VII presents the results for each location and Figures 10 to 12 in the appendix show the predicted values versus the actual values.

| | Train: SMAPE | %error | Test: SMAPE | %error |
|---|---|---|---|---|
| 24 Agosto | 0.216 | 10.8% | 0.352 | 17.6% |
| Av.Franca | 0.476 | 23.8% | 0.580 | 29.0% |
| Candido Reis | 0.244 | 12.2% | 0.450 | 22.5% |

TABLE VII: Results of models for 3 different locations

## IX. CONCLUSIONS

In our view, the results of this project were satisfactory.

We started by exploring an algorithm which is seen as state of the art for this type of problem, neural networks. But when comparing its performance to other algorithms, we saw that random forest had much better results and decided to optimize it further. The neural network algorithm seemed very obscure in comparison to the random forests, where we can easily know which features carry the most weight in the decision. Neural networks also seemed harder to optimize, due to the high number of parameters. Random forests have other advantages like the capacity to do feature selection, which picked some variables that were not in our initial list.

A 17.6% error was achieved using a random forests with 6 features and 60 trees for *Campo 24 de Agosto*. We believe this was a good result.

## X. FUTURE WORK

Maybe a better performance could be obtained using other algorithms or tuning other parameters of the random forest. Note that SVM using the default parameters also had a good performance. Experimenting further with feature engineering could also improve the predictions.

The temporal and geographical facets of the data were not explored in the modeling at all.

Also, visualizations of the quantity of particles and other variables displayed in a map of the city could help make the message more compelling and clear. A comparison between different locations could give us an idea of how the pollutants vary across the city.

## APPENDIX

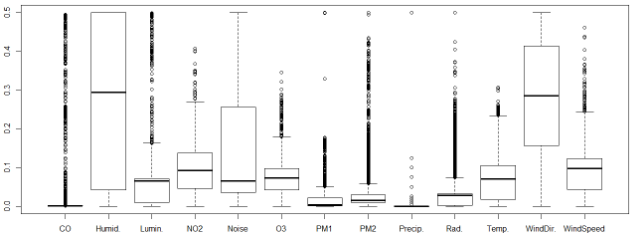


Fig. 9: Location Campo 24 Agosto - Statistical description after pre-processing

| | Min | Median | Max | Mean | SD | NA's |
|---|---|---|---|---|---|---|
| CO (Volts) | 0.0 | 1.2 | 2.5 | 1.2 | 0.2514859 | 1295059 |
| Humidity (%) | 1.0 | 60.2 | 99.9 | 48.5 | 42.373 | 1125784 |
| Luminosity (Volts) | -72193.4 | 3095.6 | 266216.0 | 12556.9 | 23590.7 | 1297381 |
| NO2 (Volts) | 0.0 | 2.0 | 2.5 | 1.9 | 0.3737143 | 1308395 |
| Noise (dB) | 0.0 | 0.0 | 129.7 | 43.3 | 52.1175 | 349863 |
| O3 (Volts) | 0.2 | 1.4 | 2.5 | 1.5 | 0.2280403 | 1308525 |
| Particles1 (Counts) | 0 | 0 | 30002 | 247 | 882.4187 | 1102909 |
| Particles2 (Counts) | 0 | 0 | 30002 | 294 | 1551.119 | 1102909 |
| Precipitation (Number of buckets - 1 bucket = 0.2794mm) | 0.0 | 0.0 | 5.6 | 0.0 | 0.04106458 | 1228768 |
| Solar Radiation (Volts) | 0.0 | 0.0 | 2.3 | 0.1 | 0.1821339 | 1261371 |
| Temperature (°C) | 1.9 | 19.5 | 46.4 | 19.1 | 5.423069 | 1103232 |
| Wind Direction (Degrees) | 0.0 | 180.0 | 270.0 | 192.5 | 98.66372 | 1266312 |
| Wind Speed (rps - 1 rps = 1.492miles/h) | 0.0 | 2.2 | 25.5 | 2.9 | 2.356852 | 1219785 |

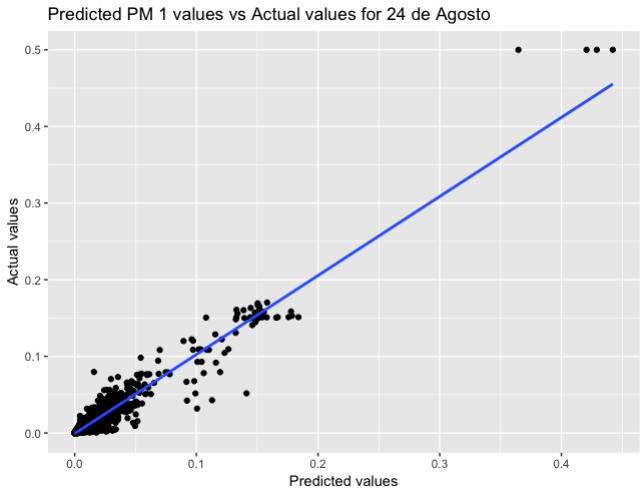


Fig. 10: Results for the 24 de Agosto location

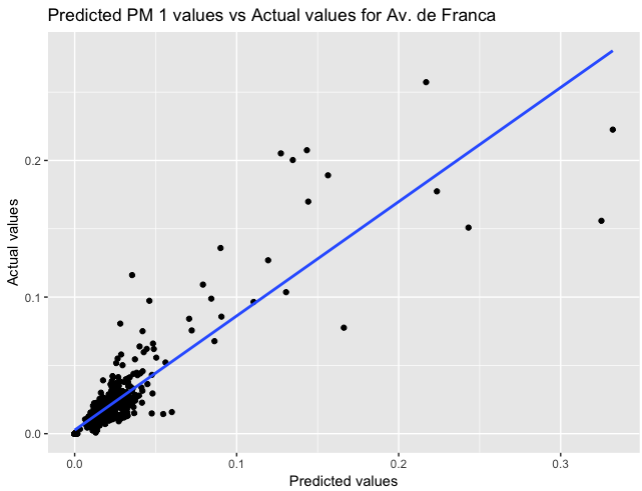


Fig. 11: Results for the Avenida de Franca location

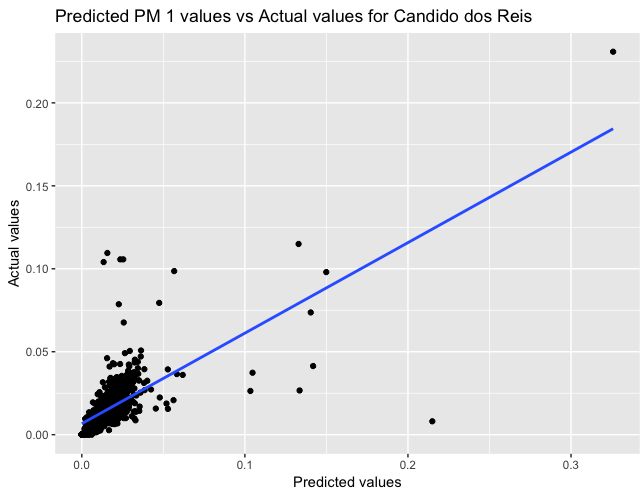


Fig. 12: Results for the Candido dos Reis location

## REFERENCES


[1] Joana Monteiro Alexandra Costa Solange Do Carmo Pereira Maria Madureira Joana Mendes Ana Teixeira João Paulo. Torres, Pedro Ferreira. Air pollution: A public health approach for Portugal. *Science of The Total Environment*, 24(643):1041–1053, 2018.

[2] Raquel Martínez-España, Andrés Bueno-Crespo, Isabel Timón, Jesús Soto, Andrés Muñoz, and José M. Cecilia. Air-pollution prediction in smart cities through machine learning methods: A case of study in Murcia, Spain. *Journal of Universal Computer Science*, 24(3):261–276, 2018.

[3] Ibrahim Kök, Mehmet Ulvi Şimşek, and Suat Özdemir. A deep learning model for air quality prediction in smart cities. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, 2018-January(December):1983–1990, 2018.

[4] Ping Wei Soh, Jia Wei Chang, and Jen Wei Huang. Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations. *IEEE Access*, 6:38186–38199, 2018.

[5] Croitoru Cristiana and Nastase Ilinca. A state of the art regarding urban air quality prediction models. *Eenviro*, 01010:1–7, 2018.

[6] Hong Zheng, Haibin Li, Xingjian Lu, and Tong Ruan. A multiple kernel learning approach for air quality prediction. *Advances in Meteorology*, 2018, 2018.

[7] Giorgio Corani. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185(2-4):513–529, 2005.

[8] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep Distributed Fusion Network for Air Quality Prediction. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pages 965–973, 2018.

[9] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie. Air Quality Prediction: Big Data and Machine Learning Approaches. *International Journal of Environmental Science and Development*, 9(1):8–16, 2018.

[10] Xiang Li, Ling Peng, Yuan Hu, Jing Shao, and Tianhe Chi. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23(22):22408–22417, 2016.

[11] Mussab Alaa, A. A. Zaidan, B. B. Zaidan, Mohammed Talal, and M. L.M. Kiah. A review of smart home applications based on Internet of Things. *Journal of Network and Computer Applications*, 97:48–65, 2017.

[12] Pasak Senawongse, Andrew R Dalby, and Zheng Rong Yang. Air Quality Prediction By Machine Learning Methods. *the University of British Columbia*, (October):1147–1152, 2015.

[13] D. Calcada, T. Moura. UrbanSense Platform http://www.slideshare.net/futurecitiesproject/2014-future-cities-conference-joel-silveirinha-the-internet-ofeverything. *Future Cities Conference*, 2014.

[14] Ali Ziat, Edouard Delasalles, Ludovic Denoyer, and Patrick Gallinari. Spatio-temporal neural networks for space-time series forecasting and relations discovery. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2017-November:705–714, 2017.

[15] Elias Kalapanidas. Applying Machine Learning Techniques in Air Quality Prediction. *Researchgate.Net*, (September 1999), 2016.

[16] World Health Organization. Health Aspects of Air Pollution with Particulate Matter , Ozone and Nitrogen Dioxide. *Report on a WHO Working Group*, (January):98, 2003.

[17] Dixian Zhu, Changjie Cai, Tianbao Yang, and Xun Zhou. A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. *Big Data and Cognitive Computing*, 2(1):5, 2018.

[18] https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9.