Wind profile prediction in an urban canyon: a machine learning approach

To cite this article: Dasaraden Mauree et al 2019 J. Phys.: Conf. Ser. 1343 012047

View the article online for updates and enhancements.
Wind profile prediction in an urban canyon: a machine learning approach

Dasaraden Mauree1, Roberto Castello1, Gianluca Mancini1,2, Tullio Nutta1,2, Tianchu Zhang1,2, Jean-Louis Scartezzini1

1Solar Energy and Building Physics Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland
2Department of Computer Science, Ecole Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland
dasaraden.mauree@gmail.com

Abstract. Resolving the wind profile in an urban canyon environment means dealing with the turbulent nature of the stream and the presence of non-negligible flux exchanges with the atmosphere inside the canopy, making any deterministic model solution computationally very intensive. In this paper, a learning-from-data method is explored, which is able to predict the wind speed in an urban canyon at different heights, given a minimal set of input features. The experimental location is provided by a street canyon located at the Swiss Federal Institute of Technology campus in Lausanne, equipped with several measuring stations to record data at high temporal resolution. Different machine learning approaches are compared in order to predict the wind speed in two directions and at different heights inside the urban canyon: an optimized Ridge Regression outperforms the Random Forest algorithm. We find particularly high accuracy in predicting the wind speed in the highest part of the canyon. None of the proposed algorithms however is able to model in an accurate way the variation of the wind speed close to the ground.

1. Introduction

Most of the efforts in predicting wind speed profiles inside the urban environment have been focusing on deterministic models based on finite volume methods. In fact most of the boundary layer theories developed [1,2] cannot be applied in the above mentioned case since many assumptions do not hold as shown by [3–5]. As a consequence fluxes exchanged between the atmosphere and the buildings inside the canopy [6], as well as the vertical profiles of meteorological variables inside the urban canopy layer, have been calculated to improve the accuracy of the prediction [7]. These methods have been shown to be successful, but often they are quite computationally intensive and they still struggle to capture the complex physical behaviour close to the ground [8]. In this study a statistical data driven approach has been considered to predict the wind speed along two directions inside an urban canyon located at the Swiss Federal Institute of Technology (EPFL) campus. Statistical approaches have already been used in the context of wind resource assessment [9], but not for this specific task. The novel approach presented here considers three different machine learning algorithms: Random Forest ensemble method, a neural network and a Ridge Regression. The goal of this paper is to determine, on the one hand, the importance of a set of physical features from a statistical point of view and, on the other one, to test the accuracy of the algorithms in predicting the wind speed in two directions at different heights. By predicting the
speeds at different heights inside the canyon, we obtain the wind velocity profile. Due to the availability of one experimental setup only, the applicability of the algorithms is limited to the urban canyon considered in the experiment. The rest of this paper is organized as follows. In Section 2, the dataset is described together with the necessary data preparation steps; the regression and feature selection algorithms are both outlined. In Section 3, the prediction results of the three different algorithms and the most important features are presented and evaluated according to the metric presented in Section 4.

2. Methodology

2.1. Data Structure

The investigation is based on almost one year (2018) of meteorological data measured inside the urban canyon at EPFL. The set consists of wind speed measurements taken by seven anemometers placed on a mast in a range of heights going from 1.5 m to 25.5 m [10]. The wind speed measured at the top of the mast is considered to be the free stream velocity of the wind, while the remaining ones are those which have to be predicted (targets). Each wind speed measurement has a magnitude and an orientation given in degrees. The vertical velocities are not going to be considered in this investigation. On top of these inputs, there is another set of measurements concerning the temperature measured at different cardinal orientations and the solar irradiance radiation. The data are taken at different frequencies: the wind speed is measured at 20 Hz resolution while the rest of the measurements is taken at 1 Hz resolution. In total a set of 20 features is extracted, including the height of each anemometer too.

2.2. Data Preparation

We perform several data manipulations to construct the dataset. Firstly, the resolution of the data has to be uniform: since the predictions were to be performed over a year of data, the choice of the time resolution to adopt is an important aspect to consider. Choosing a very coarse resolution would imply a loss of accuracy, but a noise reduction especially at low heights. However, greatly reducing the resolution could lead to a loss of information between the features and the target variables. As a consequence, we opt for a five minutes resolution. This interval can capture the meteorological variability while reducing noise and allowing for a big enough dataset to train the algorithm. Hence, we compute the mean value of each feature and of the target variables over each five minutes interval. Secondly, a feature selection is performed: the albedo is found to be extremely noisy and therefore classified as not relevant for the prediction. Also, about 10% of the measurement timestamps are empty, as a result of instruments malfunctioning. Given the large amount of data available, they have been discarded. Thirdly, the dataset is further subdivided in training, evaluation and test chunks according to the following percentages: 60%, 20%, and 20%. Then, the dataset is further split into four different subsets according to the season each data sample belongs to. By doing so, it is possible to compare yearly regression versus season wise regression performances. The output values have also been discretized: a uniform discretization of the two-dimensional output is chosen, meaning all bins of the two wind speed directions have the same length. Finally, for each input column feature we perform a normal standardization to the unity, in order to account for the different orders of magnitude of the input features.

2.3. Feature Selection methods

In order to understand which physical features are the most relevant, two selection methods are considered: a stepwise feature selection and a Random Forest embedded method. After a first set of tests, the first one is discarded because it is too computationally intensive. The chosen feature selection model is therefore Random Forest. The algorithm can in fact assign feature importance in parallel to the regression. In addition, the method takes into account the interactions between the variables providing a rank of the feature importance [11].
2.4. Regression Algorithms

Before presenting each regression algorithm, it is essential to mention the two metrics used to evaluate the algorithms performances. The mean squared error (MSE) is indicating the accuracy of the prediction, while $R^2$ is the value indicating the goodness of the regression fit with the measured data. As baseline method, a regression is performed using a non-optimized ridge regression.

2.4.1. Ridge. The first investigated method is a more refined version of the baseline. The standard ridge regression is paired with a generalized cross-validation, which is in the form of efficient Leave-One-Out cross-validation. The cross-validation is also used to optimize the regularization parameter spanning from $1^{-10}$ and $1^7$ with 200 intermediate steps. The input of the regression is augmented by a third-degree polynomial expansion. The expansion exploits all the possible combinations with the features and their value at the power of two and three, adding also a bias column filled with ones. This result in 1771 columns from the original 20 of the starting regression matrices. Optimization of the degree of expansion is performed. Increasing the degree improves the performance constantly without overfitting. Given the available computational power, third degree is the highest achievable and thus the chosen one. The presence of the optimized regularization parameter, together with a large ratio between the data points and the number of features, prevents overfitting during the regression.

2.4.2. Random Forest. In order to run the random forest prediction, the horizontal wind speed components $u_x$ and $u_y$ are uniformly discretized into several intervals according to the user-defined precision for one interval, and $u_x$ and $u_y$ are represented by the mean of each interval. Here the precision is 0.1 m/s. The number of trees in the random forest is 120 and the maximum number of layers per tree is set at 1000, due to the limit of our computational memory. It is important to mention that in order to evaluate the performance in terms of MSE, the output has to be continuous. Hence the average of each bin is taken and the MSE is extracted in this fashion. Then the Random Forest regression method is chosen because, firstly it is a natural continuum of the feature selection algorithm and secondly it has several advantages: the data do not need to be standardized and performing the regression does not require the tuning of many hyperparameters.

2.4.3. Neural Network. The last machine learning regression method considered is a Neural Network. The Neural Network is used to capture the complexity of turbulent phenomena in the data, especially in order to obtain accurate predictions at low heights where the relationship is extremely non-linear and complex. The number of layers is set to one after scanning a range of layers in the interval from 1 to 5: the number of layers and the number of neurons is chosen by evaluating the final MSE of different combinations over the validation dataset. The combination of one layer with 100 neurons allows to minimize the risk of overfitting in a problem where the physical dependence between the features and the target is already known deterministically. The distribution of the initialized weights is a uniform distribution $U(-\sqrt{k}, \sqrt{k})$ where $k = \frac{1}{n_{\text{features}}}$. To further prevent overfitting during training, the behaviour of the MSE loss function on the validation ($MSE_v$) set is analysed by setting two criteria: if the moving average measured over 10 epochs increased 3 times sequentially the training is stopped; if the $MSE_v$ rate of change over one epoch is bigger than 40%, then the training is stopped and the weights are not updated in order to prevent the overfitting.

Results

Table 1 shows the features whose relative importance is invariant under seasons. It can be noticed as, even within the summer season, the sonic temperature is the most influencing feature. In fact, the temperature is a confounding variable for the sequence of night and day and consequently it is what drives the free stream wind at the top. Beside the sonic temperature, the first most important features are identified with the height of the anemometers, and the top anemometer’s speed decomposition in x and y directions.
Table 1: Feature importance

| Feature          | Relative Importance |
|------------------|---------------------|
| Sonic Temperature [°C] | 0.11                |
| Height [m]       | 0.11                |
| $u_{top\,y}$ [m/s] | 0.07                |
| $u_{top\,x}$ [m/s] | 0.07                |
| $u_{top\,z}$ [m/s] | 0.05                |

2.5. Algorithm predictions

First, it is important to mention that predictions performed over each season, instead of over the entire year of data, led to an improvement of 25% in the MSE. Then, for each season the MSE and $R^2$ of each anemometer and speed component is derived. Among the set of results, the most relevant ones are chosen in order to compare the accuracy of the predictions. Summer is chosen to be the representative season for evaluating the results of the algorithm; the observations for the summer season are therefore applicable to the other seasons as well.

As visible in Figure 1, the algorithm constantly providing the lowest MSE is the Ridge Regression. The predictions for the lowest anemometers have lower values of MSE because the order of magnitude of the speed is much smaller compared to those ones measured by the highest anemometers.

Figure 1: Mean Square Error (left) and $R^2$ (right) on test sample data set (Summer)

Figure 2: Graphical representation of the regression fit of the testing sample for anemometer 5 (right) and anemometer 2 (left) in Summer

Figure 1 also reports the trend of the coefficient of determination $R^2$: for lower anemometers (from 1 to 3) it can be noticed that the best fit is given by the Random Forest regression, being the closest to 1. For higher anemometers the Ridge Regression is giving the best result, performing better than the other two...
methods with a $R^2$ peak value of 0.82. As it can be seen in Figure 2, the Ridge Regression algorithm is able to reproduce the trend observed in the test dataset quite accurately for high anemometers and less for low ones. Figure 3 shows the average speed predicted for each anemometer by each algorithm and it compares them with the true average values. It can be noticed as the wind speed predictions lie within the distribution of the measured wind profile inside the canyon (equivalent to one standard deviation error bars visible in Figure 3).

![Figure 3: Wind speed Profile constructed after prediction of target variables on test sample data set (Summer)](image)

3. Discussions and Conclusions

The feature selection over four different seasons leads to the same most important features: the free stream velocity at the top, the sonic temperature and the height are the most important factors to predict the wind speed at different heights. Moreover, we find that the predictions derived using the four season-wise datasets are more accurate than the predictions produced over the entire year dataset. Not only the task for each season is more specific, but also the features in each season have probably different types of dependencies which can be better captured if the tasks are performed independently. Finally, among the set of regression algorithms the ridge regression is the one giving the best performance in terms of MSE and $R^2$.

Among the three methods, the one which underperformed is the Neural Network. Its values of $R^2$ and MSE are reported in Figure 1. Compared to the other two methods, the clearest difference in performance is visible in the $R^2$ values: especially for the winter season the fit is particularly suboptimal. This can be explained by the fact that the data for the winter season are less in quantity compared to the rest of the seasons, suggesting that the higher the number of data samples, the higher is the accuracy on the prediction. However, the network has structural limitations; a systematic optimization of the architecture has not been performed even though a set of different layer configurations and neurons have been considered. We can only assert that the architecture adopted for this paper underfits the data.

It is worth noticing that Random Forest regression gives the best performances both in MSE and $R^2$ at low anemometers and it follows the average speed per anemometer very accurately as shown in Figure 3. This can be explained by considering that the bin division makes the task of predicting the wind speed easier. In fact, the variability of the data is reduced by discretization. The difference in accuracy between the anemometers close to the ground and those higher up can be explained by considering the more complex physical phenomena which govern the wind speed profile, as highlighted in Section 1. The particles motion is limited, but extremely variable; hence it is very challenging for the algorithm to learn from a set of features which are not consistently influencing the output. In addition, the direction of the wind is not normally distributed: the prediction of the $y$ component of the wind speed is consistently more accurate than the $x$ component. By carrying out a ridge regression with as single output the $y$ component of the speed, the prediction improved especially for the values of $R^2$.

To conclude, two different regression problems can be identified in this investigation: one for the first three lowest anemometers and another one for the rest of them. The dependency between the
features and the output variables can be different according to the set of anemometers chosen, as proven by regression results. Hence, further studies could investigate different regressions for two groups of different anemometers or even for each anemometer individually. In this way, even if the algorithm is not able to capture the inter-dependencies among the anemometers, it could still be able to predict the complex physical wind speed behaviour at lower anemometers. Further improvements could also be introduced by optimizing the time resolution of the data. Changing this factor means varying the wind speed vector output and hence changing its probability distribution. As a consequence, different regressions for different time resolutions should be performed and that one outputting the best MSE and $R^2$ should be chosen. Finally, by integrating the model with a set of additional features concerning the urban neighbourhood (building heights, ratio, etc.) from other canyons, the model could be applied to different urban areas, enabling a more widespread set of applications.

Acknowledgements
This research project has been financially supported by the Swiss Innovation Agency Innosuisse and is part of the Swiss Competence Center for Energy Research SCCER FEEB&D and by the HyEnergy project of the National Research Programme 75 “Big Data” (PNR75) of the Swiss National Science Foundation (SNSF). MoTUS experiment was funded by EPFL, Switzerland and by the ENAC Faculty.

References
[1] Monin A S and Obukhov A M 1954 Basic laws of turbulent mixing in the surface layer of the atmosphere Contrib. Geophys. Inst. Acad. Sci. USSR 151 163–187
[2] Foken T 2006 50 Years of the Monin–Obukhov Similarity Theory Boundary-Layer Meteorol 119 431–47
[3] Karam H A, Filho A J P, Masson V, Noilhan J and Filho E P M 2009 Formulation of a tropical town energy budget (t-TEB) scheme Theor Appl Climatol 101 109–20
[4] Rotach M W 1993 Turbulence close to a rough urban surface part I: Reynolds stress Boundary-Layer Meteorology 65 1–28
[5] Roth M 2000 Review of atmospheric turbulence over cities Quarterly Journal of the Royal Meteorological Society 126 941–990
[6] Salamanca F, Krpo A, Martilli A and Clappier A 2010 A new building energy model coupled with an urban canopy parameterization for urban climate simulations—part I. formulation, verification, and sensitivity analysis of the model Theor Appl Climatol 99 331–44
[7] Mauree D, Coccolo S, Kaempf J and Scartezzini J-L 2017 Multi-scale modelling to evaluate building energy consumption at the neighbourhood scale PLOS ONE 12 e0183437
[8] Mauree D, Blond N, Kohler M and Clappier A 2017 On the Coherence in the Boundary Layer: Development of a Canopy Interface Model Front. Earth Sci. 4
[9] Veronesi F, Grassi S, Raubal M and Hurni L 2015 Statistical Learning Approach for Wind Speed Distribution Mapping: The UK as a Case Study AGILE 2015 Lecture Notes in Geoinformation and Cartography (Springer, Cham) pp 165–80
[10] Mauree D, Deschamps L, Becquelin P, Loesch P and Scartezzini J-L 2017 Measurement of the impact of buildings on meteorological variables Building Simulation Applications, BSA 2017 273–8
[11] Paul J, Verleysen M and Dupont P Identification of Statistically Significant Features from Random Forests 12