Wind Profile Prediction in an Urban Canyon: a Machine Learning Approach

Gianluca Mancini, Tullio Nutta, Tianchu Zhang
Department of Computer Science, EPFL Lausanne, Switzerland

***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 which make any deterministic model very computationally intensive. In this paper a statistical data driven learning approach is taken: the wind speed at different heights in a urban canyon is predicted. An urban canyon located in the École Polytechnique Fédérale de Lausanne (EPFL) campus provides the optimal experiment location thanks to the presence of several meteorological measuring stations. Three different machine learning approaches are compared in order to predict wind speed in two directions at different heights inside the urban canyon: an optimized ridge regression outperformed random forest and a neural network in terms of performance. A particularly high accuracy in predicting the wind speed in the highest part of the canyon is shown. None of the algorithm proposed, however, is able to accurately capture the variation of the wind speed close to the ground.**

I. Introduction

Most of the efforts in predicting wind speed profiles inside an 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 don't hold as shown by [3], [4] and [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 [7], have been calculated to improve the accuracy of the prediction. These methods have been shown to be successful, but they are often very computationally intensive and they still struggle to capture the complex physical behaviour close to the ground. In this study a statistical data driven approach has been considered in order to predict the wind speed in two directions inside an urban canyon located in the EPFL campus. Statistical approaches have been used in the context of wind resource assessment [8], but not for this specific task. The novel approach presented in this paper compares three different machine learning algorithms: an ensemble method, namely Random Forest, a neural network and a ridge regression. Hence 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, to test the accuracy of the algorithm proposed in predicting the wind speed in two directions at different heights. By predicting the wind speeds at different heights inside the urban canyon the velocity profile will be obtained. It is important, however, to consider that the applicability of the above mentioned algorithms is limited to the urban canyon considered in the experiment. The rest of this paper is organized as follows. In the section II the data source will be described and the necessary data preparation procedures are described; the regression and feature selection algorithms are outlines. Then both the prediction results of the three different algorithms and the most important features will be presented in section III and evaluated according to the metric developed in section IV.

II. Method

A. Data Structure

The investigation is based on almost an year (2018) of meteorological data measured inside the urban canyon at EPFL. The data consists of a set 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. 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 that have to be predicted statistically (target variables). Each wind speed measurement has a magnitude and an orientation given in degrees. The vertical velocities are not going to be considered in this investigations. On top of these data, there is a set of other measurements concerning the temperature measured from different orientation, the solar radiation, irradiance, solar radiation. The data are taken at different frequencies: the wind speed measurements have 20 Hz resolution while the rest of the measurements have 1 Hz resolution. In total a set of 20 features are obtained including the height of each anemometer too. These will be used to predict the horizontal wind speed in two different direction at different heights based on changing meteorological parameters.

B. Data Preparation

Different data manipulation had to be performed for different phases of the investigations. Firstly, the resolution of the data had to be uniform: since the predictions were to be performed over a year of data, the choice of the time resolution to adopt was quite a critical one. 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 it was decided to opt for a five minutes resolution: this variation can capture the meteorological variability while reducing noise and allowing for a big enough dataset to train the algorithm. Hence the average values of each feature and of the target

variables over five minutes intervals were computed. Secondly, a preliminary feature selection was performed: the albedo was extremely noisy and it was not considered as a relevant feature that would have an influence on the prediction. Also about 10% of the time step measurement was corrupted as a result of instruments errors; because of the large amount of data, these were disregarded. Thirdly, the dataset was divided in a training, an evaluation and a testing set with the following proportions (60%, 20%, 20%). Then, the dataset was divided in four different smaller datasets according to the season each data point belonged to. By doing so, yearly regressions and season wise regression will be compared in performance. Eventually, in order to perform the feature selection by using Random Forest, the output had to be discretized: a uniform discretization of the two dimensional output was decided, meaning all bins of the two directions of speed had the same length. Finally a standard normal standardization of each input column feature was performed in order to take into account for the differences in orders of magnitudes of the features.

C. Feature Importance

In order to understand which physical features were the most relevant ones, two feature selection methods were considered: a stepwise feature selection and Random Forest. After a set of analysis, however, the first one was discarded because it was too computationally intensive. The feature selection model chosen is Random Forest. The algorithm, in fact, can assign feature importance in parallel. In addition the method takes into account the interactions between the variables providing a rank of the feature importance [9].

D. Regression Algorithms

Before explaining each regression algorithm, it is essential to mention the metrics used to evaluate the algorithms performances: the mean squared error (MSE) is an indication of the predictions' accuracy, while R^2 is the parameter indicating the fit of the regression with the data. In addition, a baseline regression was performed using a non optimized ridge regression.

1) Ridge: The first method used is a more refined version of the baseline used. The standard ridge regression is paired with a generalized cross-validation, which is a 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^{5} with 200 intermediate steps. The input of the regression is augmented by a third degree polynomial expansion. The expansion exploits all the different combination possible with the features and their value at the power of 2 and 3, adding also a bias column filled with ones. This result in 1771 columns from the original 20 of the starting regression matrix. Optimization of the degree of expansion was performed and was found that increasing the degree improved the performance constantly without overfitting. Degree 3 was the highest achievable, and therefore the chosen one, with the available RAM on the whole year dataset. 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) Random Forest: In order to run the random forest prediction, the u_x and u_y is 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 maximal number of layers per tree is chosen as 1000, due to the limit of the computational memory. It is important to mention that in order to evaluate the performance in terms of MSE, the output had to be continuous. Hence the average of each bin was taken and the MSE was found in this fashion. Then Random Forest regression method was 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 will not involve tuning many hyperparamters.

3) Neural Network: The last regression method to be considered is a Neural Network. The Neural Network was chosen with the aim of capturing the most complex turbulent phenomena in the data, especially to have accurate predictions at low heights where relationship are extremely non linear and complex. The number of layers was set to 1 since after testing a set of layers from 1 to 5: the number of layers and the number of neurons was chosen by evaluating the final MSE over the data of different combinations. The combination of one layer with 100 neurons has allowed to minimize the risk of overfitting in a problem where the physical dependence between the features and the target was already known deterministically. The distribution of the initialized weights is uniform Weights $\sim U(-\sqrt{k}, \sqrt{k})$ where $k = \frac{1}{n_{features}}$. To further prevent overfitting during training, the behaviour of the mean squared error loss function on the evaluation (mse_{ev}) set was analyzed by setting two measures: if the moving average measured over 10 epochs increased 3 times sequentially the training was stopped; if the rate of change of mse_{ev} over one epoch was bigger than 40% the training was stopped and the weights were not updated in order to prevent overfitting of the training set.

III. Results

A. Feature Importance

Table III-A shows the features whose importance is invariant of the season. The complete feature importance table is present in the section G of the Appendix. It is surprising to find out that statistically, even within the summer season, the sonic temperature is the most influencing feature. In fact the temperature is a confounding variable foe the sequence of night and day and consequently it is what drives the free stream wind at the top. The relative importance of feature is consistent for the all seasons.

For all the seasons, the first 4 most important features are: the sonic temperatures, the height of the anemometers, the top anemometer's speed decomposition in x and y directions.

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 [W/m^2]	0.05

This results of the feature importance are consistent with the physical model [7].

B. Algorithms Predictions

Before commenting on the relative performance of the algorithms, it is important to mention that the prediction performed over the seasons, instead of over one year of data, led to an improvement of the 25 % of MSE (Appendix Subsection E). For each season the MSE and R^2 of each anemometer and speed component was calculated. Among the set of results, the most relevant ones were chosen in order to compare the accuracy of the predictions. Summer was chosen as the representative season to evaluate the results of the algorithm; the observations made for Summer are also applicable to the rest of the seasons whose details can be found in the Appendix Section C. As shown by Figure 1, the

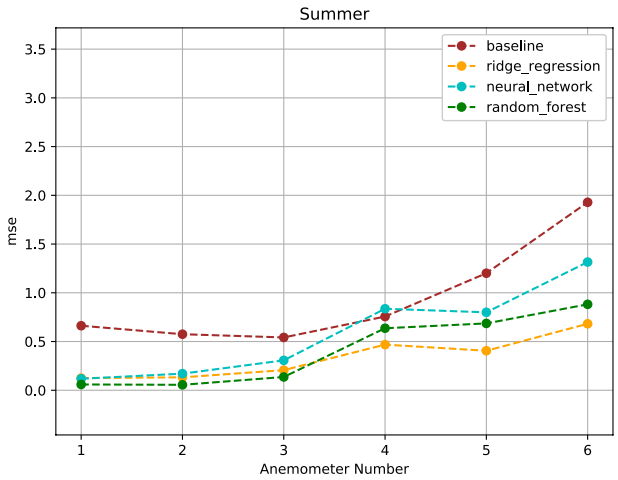

Fig. 1. Mean Square Error on test sample data set, Summer

algorithm which provides consistently the lowest MSE is the Ridge Regression. The predictions for the lowest anemometers have lower values of MSE because of the order to magnitude of the speed is much smaller compared to those ones measured by higher anemometers. In fact, as it can be seen in Figures 2 and 3, the Ridge Regression manages to reproduce the trend of the test set quite accurately at high anemometers while not at low ones. It is important to mention that none of the regression algorithm led to overfitting; in the case of the neural network this is particularly relevant. In figure **??** an example of the training behaviour of the neural network is provided: after reaching the plateau the mse_{ev} has been slowly increasing; consequently the training was stopped. Figure **??** also shows that not only the training method didn't lead to overfitting, but it is also probably underfitting: in fact the

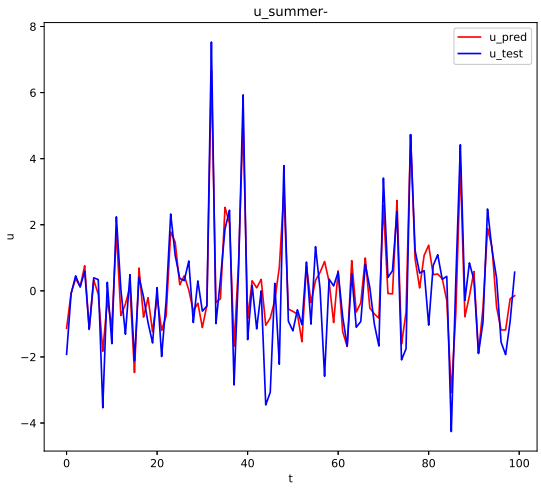

Fig. 2. Graphical representation of the regression fit of the testing sample for anemometer 5 in Summer

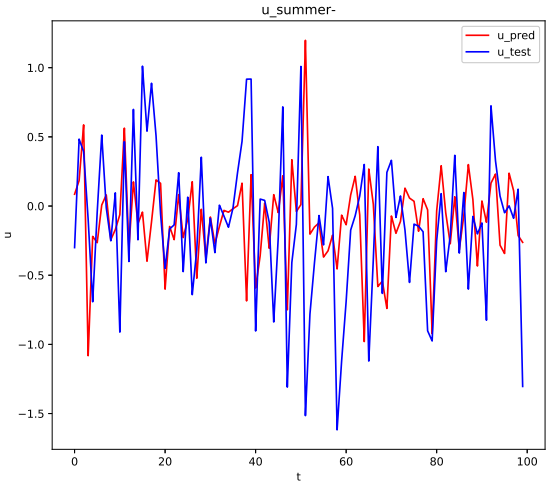

Fig. 3. Graphical representation of the regression fit of the testing sample for anemometer 2 in Summer

behavior of the MSE for the training and the evaluation set are very close to each other meaning that the training capture the global behaviour of the data without clearly showing a divergence after a certain number of epochs. Figure 4 shows the behaviour of the coefficient of determination. For lower anemometers (from 1 to 3) it can be seen that the best algorithm is the Random Forest Regression as it is the closer to 1. As in the case of MSE explained above, the performance of Random Forest Regression is boosted by the number of bins in the interval considered. For higher anemometers the Ridge Regression is giving the best result performing better than the other two methods with a peak R^2 value of 0.82. Eventually by plotting the average speed predicted by each anemometer

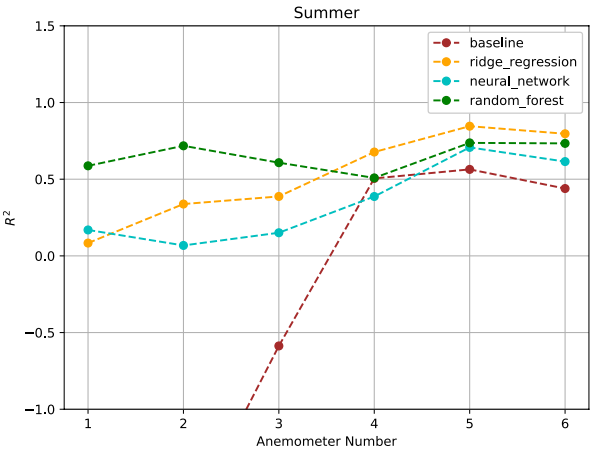

Fig. 4. R^2 on test sample data set, Summer

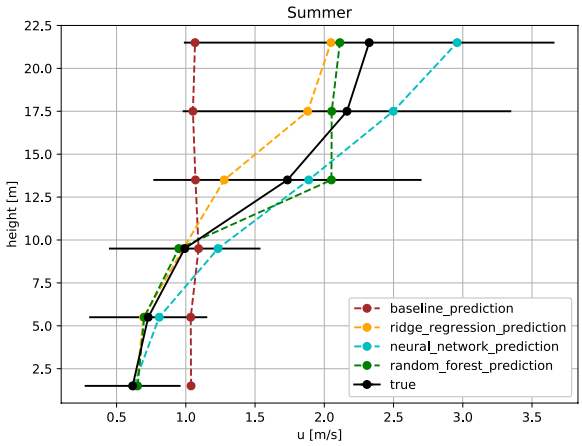

Fig. 5. Wind speed Profile constructed after prediction of target variables on test sample data set, Summer

and comparing it with the true average values, it can be seen that the predictions of the wind speeds for methods lies within the error bars of one standard deviation recreating the wind profile inside the canyon.

IV. Discussion and conclusions

From the results highlighted above important considerations can be drawn. Firstly, 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. Secondly it can be concluded that the predictions produced using four different datasets splitted season wise produces more accurate results than for predictions produced over a one year of data. Not only the task for each season is more specific, but also the features in each season probably have different dependencies which can be better captured if the tasks are performed independently. Thirdly, among the set of regression algorithms the ridge regression is that one giving the best performance in terms of MSE and R^2.

Among the three methods the one which performed the worse is the neural network. In fact its values of R^2 and MSE were exceptionally high as shown in Figure 1. The clearest difference in performance compared to the other two methods was concerning the R^2 values: especially for the winter the fit is particularly unoptimal (Appendix Section A). This can be explained by the fact that the data for the winter season were less in quantity compared to the rest of the seasons, suggesting that a greater number of data points would enhance the prediction. The limitations of the network, however, are structural: in fact a systematic optimization of the architecture wasn't performed even though a set of different layer configurations and neurons were considered. The algorithm proposed in this paper is therefore underfitting the data.

It is worth mentioning 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 5. This can be explained by considering that the bin division makes the task of predicting the 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 I. In fact 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 (Appendix Section D). By carrying out a ridge regression with as single output the y component of the speed, the prediction improved expecially for the values of R^2 (Appendix Section H).

Two different regression problems can be identified in this investigation: one for the first three lowest anemometer and another one for the rest of them. The dependency between the features and the output variables are different according to the set of anemometers chosen as proven by the regression results. Hence, further investigations could perform different regressions for two groups of different anemometers or even for each anemometer individually. In this manner, even if the algorithm would not be able to capture the inter-dependencies among the anemometers, it could be able to better predict the complex physical wind speed behaviour at lower anemometers.

Further improvements could also be performed by optimizing the time resolution of the data. Changing this factor means varying the variation of the 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, considering the results of the algorithms at relatively high heights, by integrating the model with a set of additional

parameters concerning the urban surrounding, the model could be applied to different urban areas allowing for a more widespread set of applications.

V. Acknowledgments

The team would like to thanks Doctor Castello and Doctor Mauree for the mentoring and the guidance both technical and organizational throughout the project.

References

[1] A. S. Monin and A. M. Obukhov, "Basic laws of turbulent mixing in the surface layer of the atmosphere," *Contrib. Geophys. Inst. Acad. Sci. USSR*, vol. 151, p. 163187, 1954.

[2] T. Foken, "50 years of the moninobukhov similarity theory," *Boundary Layer Meteorol*, vol. 119, p. 431447, 2006.

[3] F. A. J. P. M. V. N. J. Karam, H. A. and E. P. M. Filho, "Formulation of a tropical town energy budget (t-teb) scheme," *Theor. Appl. Climatol*, vol. 101, p. 109120, 2009.

[4] M. W. Rotach, "Turbulence close to a rough urban surface part i: reynolds stress," *Boundary Layer Meteorol*, vol. 65, p. 128, 1993.

[5] M. Roth, "Review of atmospheric turbulence over cities. q. j. r." *Boundary Layer Meteorol*, vol. 126, pp. 941–990, 2000.

[6] K. A. M. A. Salamanca, F. and A. Clappier, "A new building energy model coupled with an urban canopy parameterization for urban climate simulationspart i. formulation, verification, and sensitivity analysis of the model," *Theor. Appl. Climatol*, vol. 99, p. 331344, 2010.

[7] K. J. H. Mauree, D. and J.-L. Scartezzini, "Multi-scale modelling to improve climate data for building energy models," January 2016.

[8] M. R. F.Veronesi*, S.Grassi, "Statistical learning approach for wind resource assessment," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 836–850, 2015.

[9] M. V. Jeroome Paul and P. Dupont, "Identification of statistically significant features from random forests."