Politecnico di Milano

Scuola di Ingegneria Civile,Ambientale e Territoriale
Master in Geoinformatics

My Thesis

Bresciani Matteo

Referent professor: Brovelli Maria Antonia

June 14, 2022

Contents

1 Introduction **1**
1.1 D-Dust 1
1.2 Overview 2

2 Background and state of art **3**
2.1 Data collection 3
2.2 Feature selection 3
2.3 Feature Selection 3
2.3.1 Filter Methods 4
2.3.2 Wrapper and Embedded Methods 4
2.3.3 Multiscale Geographically Weighted Regression (MGWR) 5
2.4 Data modelling using Machine Learning 5

3 Data Collection and Pre-Processing **6**
3.1 Data preparation 6
3.2 Data cleaning 6
3.3 Feature Selection 6
3.4 Results 6

4 Case of Study (Data Modelling) **7**
4.1 Datasets description 7
4.2 Results 7
4.3 Interpretation of the results 7

5 Conclusion **8**

6 References **9**
6.1 Bibliography 9
6.2 Software used 9

Abstract

qq

Chapter 1

Introduction

Science demonstrated that deterioration of ambient air quality, due to the growing concentration of pollutant in the atmosphere, has cause a significant increment of deaths in the world. Pollutants such as particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide, cause respiratory diseases, and are important sources of mortality. Almost the entire global population (99%) breathes air that exceeds WHO air quality limits, and threatens their health.

In Europe air is getting cleaner, but persistent pollution, especially in cities, is damaging people's health. The latest reports are based on the European Environment Agency's (EEA), which shows that exposure to air pollution caused around 400,000 early deaths in the European Union (EU) in 2016.

One of the most harmful pollutants is the **particulate matter (PM)** which, according to is diamater (2.5µm or 10 µm), can get deep into your lungs or even get into your bloodstream.

Most of the particles come from chemical reactions such as sulfur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industries and automobiles.

However, a significant sources of PM are the chemical reactions generated by intensive farming. In particular this is a significant issue in the Po Valley, where intensive agricultural activity is very employed.

1.1 D-Dust

In this context, the **D-DUST project** (Data-driven moDelling of particUlate with Satellite Technology aid) aims to provide knowledge about the impact of agricultural and livestock activities on pollutants in the Po Valley (Northern Italy).

For doing that, data from ground sensor are combined with contribution provided by satellite platforms and, with the use of data science tecniques like machine-learning and geostatistical models, provide meaningfull information related to the contribution of intensive farming on pollution. The final target of the project is to provide a data-driven best-practices to policymakers, farming operators and citizens in order to minimize the production processes' effects on air quality.

1.2 Overview

Chapter 2

Background and state of art

2.1 Data collection

2.2 Feature selection

Data is in most cases incomplete and noisy. Nowadays, dealing with big amount of informations, the probability of incorrect data is higher without a proper data preparation. But only high-quality data can generate accurate models and predictions. Hence, it's crucial to process data with the best possible quality. This step is called data preprocessing, and it's one of the essential steps in data science, artificial intelligence, and machine learning.

2.3 Feature Selection

Feature Selection is an important step in data pre-processing. It consists in selecting the best subset of input variable as the most pertinent. Discarding irrelevant data is essential before applying Machine Learning algorithm in order to:

- **Reduce Overfitting**: less opportunity to make decisions based on noise;
- **Improve Accuracy**: less misleading data means modelling accuracy improves. Predictions can be greatly distorted by redundant attributes;
- **Reduce Training Time**: With less data the algorithms will train faster;

Due to the fact that there isn't a best feature selection technique, many different methods are performed. The aim of this part is to discover by experimentation which one/ones work better for this specific problem. In this part, are shown the supervised methods used, which are classified into 3 groups, based on their different approach.

2.3.1 Filter Methods

Filter-based feature selection methods adopt statistical measures to evaluate the correlation/dependence between input variables. These select features from the without machine learning algorithm. In terms of computation, they are very fast and are very suitable in order to remove duplicated, correlated, redundant variables. On the contrary, these methods do not remove multicollinearity. The filter-based feature selection methods used are the following:

- **Correlation coeffcients**: Correlation is a statistics measure that shows the strength of association between an independent variable and its target variable. The values range between -1.0 and 1.0. A coefficient of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no relationship between the movement of the two variables. In this study, I selected three types of correlation:
  - **Pearson correlation**: It's one of the most coefficients used in statistics. This measures the strength and direction of a linear relationship between two variables.
  - **Spearmanr correlation coefficient**: It's a non-parametric coefficient which, instead of Pearson, measures the monotonic relationship between two variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. Hence, since monotonic relation is less restrictive than linear relation, Spearmanr coefficent can provide further informations;
  - **Kendall Correlation**: The Kendall correlation is similar to the Spearman correlation because is non-parametric too but it measures the dependence between two variables instead their correlations;
- **Fisher Score**: F-score is one of the most used supervised feature selection methods. In my study this is implemented using *SelectKBest method*, which compute the score for each variable using the *f_regression()* function to evaluate them.
- **Variance Threshold**: It aims to remove all features with variance which doesn't meet a threshold value. Usually it removes all zero-variance features, so variables that contains useless information.

2.3.2 Wrapper and Embedded Methods

Wrapper methods, as the name suggests, wrap a machine learning model, with different subsets of input features: In this way the subsets are evaluated following the best model performance. Embedded methods instead are characterised by the benefits of both the wrapper and filter methods, by including interactions of features but also having a reasonable computational cost.

- **Exhaustive Feature Selection**: This algorithm follow the exhaustive feature selection approach with brute-force evaluation of feature subsets; the best subset with its accuracy is selected by optimizing a specified metric given an arbitrary regressor or classifier;
- **Recursive Feature Selection**: The goal of RFE is to select features by recursively considering smaller and smaller sets of features. The algorithm will return the optimal subset with its accuracy. The dimension of the subet is selected as input by the user;
- **Random Forest Importance**:

2.3.3 Multiscale Geographically Weighted Regression (MGWR)

Due to the fact that this study is related to geographic and spatial data, each pieces of information is very sensitive to the geographic distance (for the Tobler's first law of geography: *"everything is related to everything else, but near things are more related than distant things"*). MGWR is a geographic weighted regression tecnique which is an extension of another regression method which is GWR (Geographic Weighted Regression). GWR explores the potential spatial relationships and provides a measure of the spatial scale through the determination of an optimal bandwidth. MGWR instead provide an optimal bandwidth for each covariate involed in the regression. So it's therefore known as multiscale (M)GWR. So the use of MGWR method could be innovative, since multivariate models are increasingly encountered in geographical research to estimate spatially varying behaviour between a target and its predictive variables.

2.4 Data modelling using Machine Learning

Chapter 3

Data Collection and Pre-Processing

3.1 Data preparation

3.2 Data cleaning

3.3 Feature Selection

3.4 Results

Chapter 4

Case of Study (Data Modelling)

4.1 Datasets description

4.2 Results

4.3 Interpretation of the results

Chapter 5

Conclusion

Chapter 6

References

6.1 Bibliography

6.2 Software used

- **LaTeX**: used to write and to build the document [https://www.draw.io/];
- **GitHub**: used to store and manage project repository [https://github.com/];
- **GitHub Desktop**: is the official GitHub application which allows us to contribute to the project repository in an easy way [https://desktop.github.com];