Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

geoinformation engineering
ikg eth zurich

Comparison of machine learning algorithms for wind resource assessment

Author: Athina Korfiati
Supervisor: Prof. Dr. Martin Raubal

Master Thesis, HS 2015
Advisors: Dr. Fabio Veronesi, René Buffat

Introduction

The threat of global warming, the depletion of fossil fuels, and the increasing population are only some of the reasons that underline the need for new energy sources to satisfy a constantly increasing global energy demand. Among renewable energy sources, wind energy had a significant growth over the last years and is expected to hold a large share of electricity until 2050. For this reason, ***Wind Resource Assessment*** is more opportune than ever, as it is fundamental for the identification of optimal sites for wind energy investments.

The Problem: meteorological stations are sparsely located

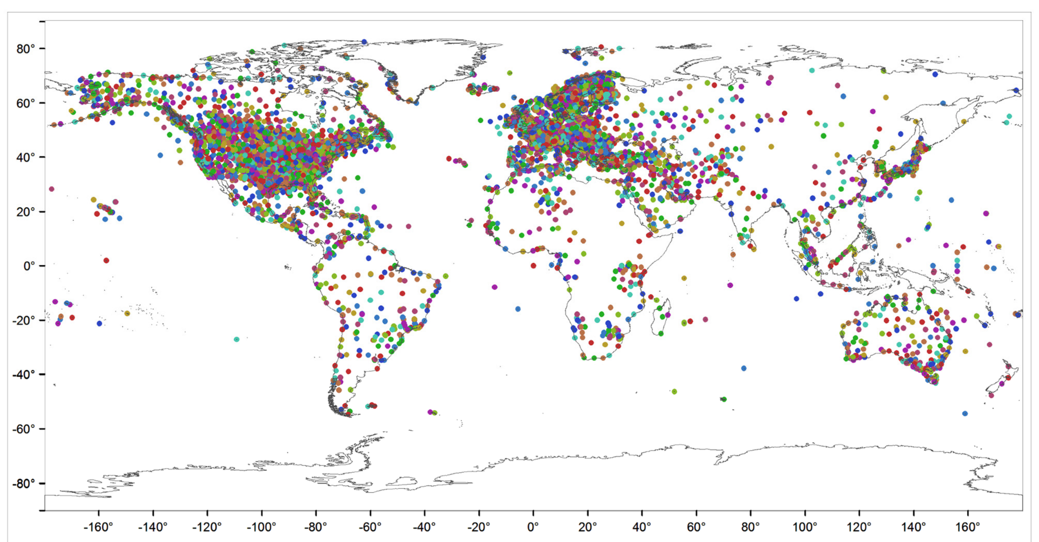
The Solution: model wind characteristics and estimate wind in locations where no data exist

Multiple machine learning algorithms have been tested. However, testing an algorithm on limited geographical regions, and thus on specific conditions of terrain or land-use, makes it difficult to draw any conclusions regarding the algorithms' accuracy. Furthermore, the use of cross-validation techniques (e.g. k-folds cross-validation) that do not take into consideration the autocorrelation that exists in environmental data, may not be the best practice in cases where autocorrelated datasets are involved.

Method overview

In total, ***eight machine learning algorithms*** were tested and their performance was measured using different cross-validation methods.

k-folds cross-validation technique is a widely used validation method for prediction problems, but it does not take into account the autocorrelation that exists in datasets (e.g. environmental data). This can be easily understood if we consider the randomness in the folds formation.



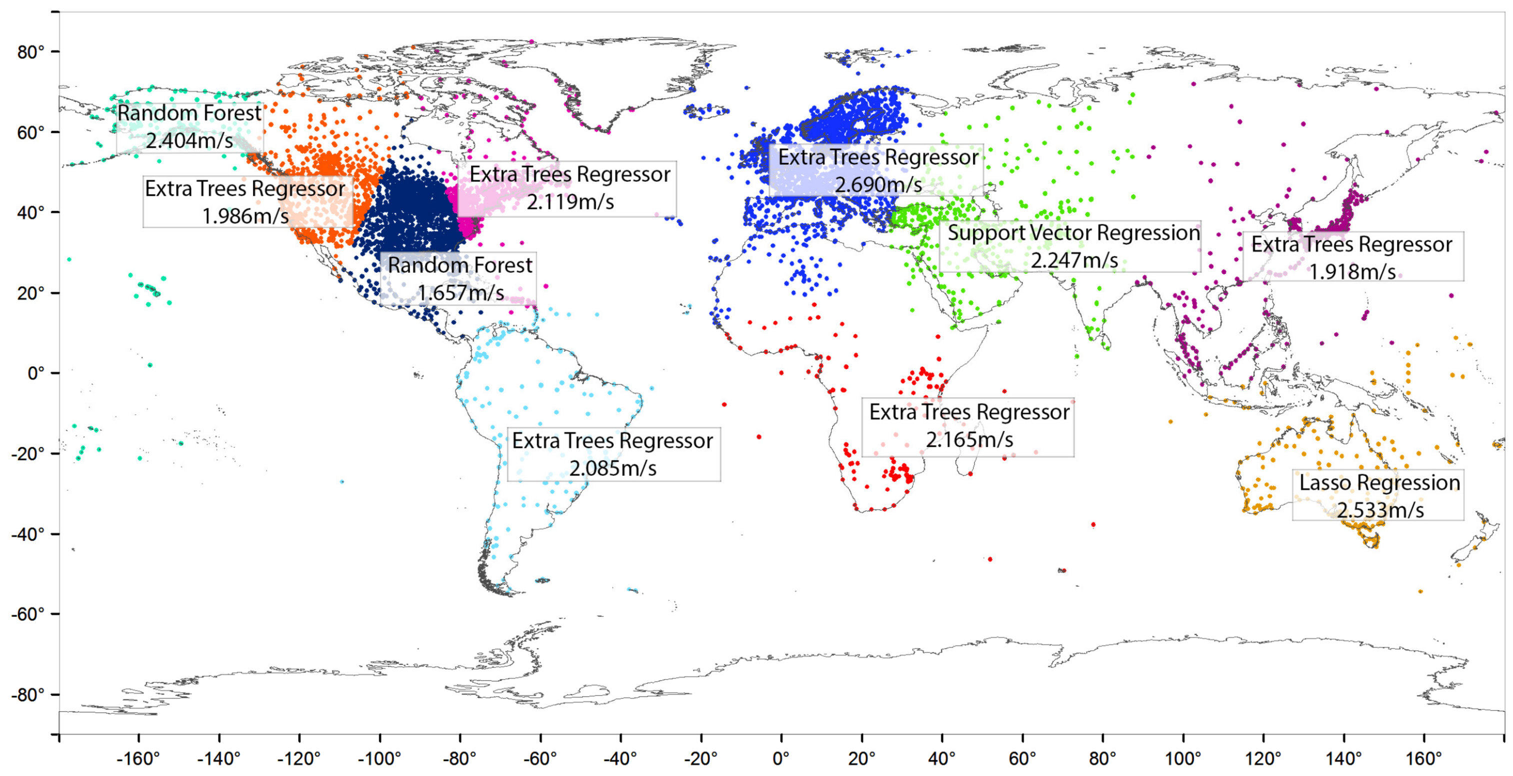
For this reason, additional validation methods are used to test the algorithms. These methods account for autocorrelation by applying geospatial rules to form the folds. ***Geographical cross-validation*** has been applied in estimation problems of other research fields, but never for wind estimation. ***Elevation and land-use cross-validation*** are special types of geographical cross-validation. This validation methods, were presented here for the first time. Having started with geographical clustering, we considered that maybe other parameters, such as elevation and land-use, might affect the wind speed's estimation accuracy. Therefore, these two additional testing methods were developed.

Results and discussion

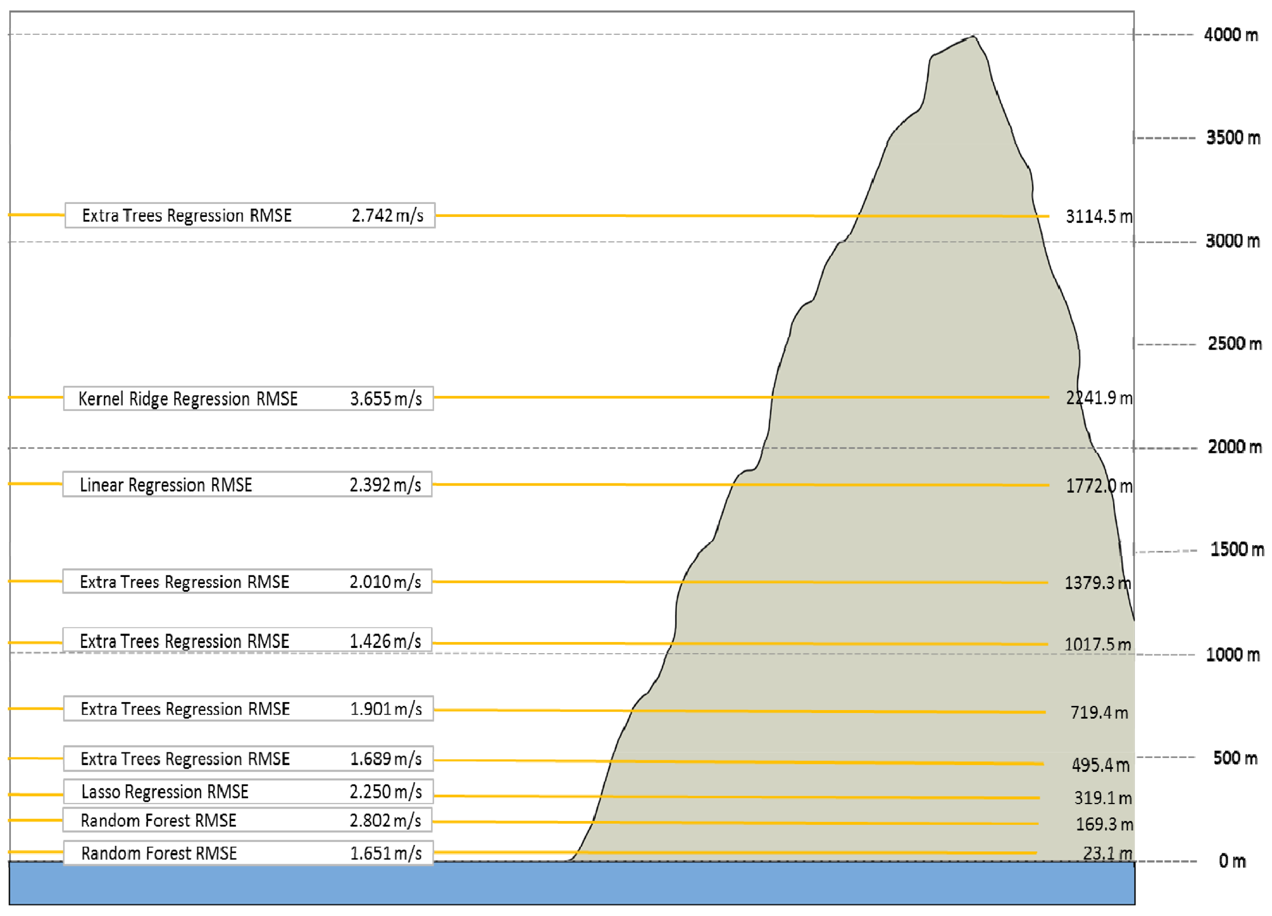
10-folds cross-validation

Algorithm	RMSE (m/s)
Random Forest	*2.058*
Extra Trees Regression	2.063
Lasso Regression	2.181
Linear Regression	2.190
Kernel Ridge Regression	2.193
k-Neighbors Regression	2.520
Decision Tree Regression	2.589
Support Vector Regression	3.105

Geographical cross-validation



Elevation cross-validation



Land-use cross-validation

Land-use	Algorithm	RMSE (m/s)
Water body	Extra Tree Regression	1.513
Forest	Extra Tree Regression	1.517
Sparse vegetation	Extra Tree Regression	2.022
Mixed	Extra Tree Regression	2.079
Urban	Random Forest	1.250
Grassland	Random Forest	1.541
Shrubland	Random Forest	2.167
Bare ground area	Linear Regression	1.542
Permanent snow	Kernel Ridge Regression	1.512
Cropland	Kernel Ridge Regression	3.465

Conclusions and future work

This Master Thesis showed that there is ***no single machine learning algorithm that can estimate the wind resource globally***. In many cases, relatively simpler methods were the best performers. Additionally, this research showed that ***k-folds cross-validation does not provide the full picture in terms of accuracy***, and should be used after consideration when autocorrelated datasets are involved. ***Autocorrelation needs to be taken into account with appropriate cross-validation techniques*** in order to find the right algorithm for every situation.

Algorithm accuracy per cross-validation method

Algorithm	k-folds cross-validation (m/s)	Geographical cross-validation (m/s)	Elevation cross-validation (m/s)	Land-use cross-validation (m/s)
Random Forest	2.058	2.235	2.386	1.893
Extra Trees Regressor	2.063	2.211	2.335	1.911
Lasso	2.181	2.476	2.332	1.980
Linear model	2.190	2.459	2.379	1.952
Kernel Ridge Regression	2.193	2.473	2.323	1.949
k-Neighbors Regressor	2.520	2.623	2.747	2.318
Decision Tree Regressor	2.589	2.629	2.957	2.462
Support Vector Regression	3.105	3.036	3.053	2.736

Some areas in which the research could continue were identified and are:

- Test ways to quantitatively explain the local accuracy differences for the various algorithms (e.g. statistical analysis of the clusters' predictors)
- Include more precise methods to determine the optimal set of parameters (e.g. grid search in Scikit-learn)
- Include more predictors (e.g. Meteosat data)
- Test ways to precisely define general application rules

IKG
Institut für Kartografie und Geoinformation