Statistical learning approach for wind resource assessment

by Veronesi, F., Grassi, S. and Raubal, M.

DOI: https://doi.org/10.1016/j.rser.2015.11.099

Veronesi, F., Grassi, S. and Raubal, M. 2016. Statistical learning approach for wind resource assessment. *Renewable and Sustainable Energy Reviews*, 56, pp.836-850.

Statistical learning approach for wind resource assessment

F. Veronesi[1*], S. Grassi[1], M. Raubal[1]

[1]Institute of Cartography and Geoinformation - ETH Zürich; Stefano-Franscini-Platz 5, 8093 Zurich

[*] Corresponding author: fveronesi@ethz.ch

Abstract

Wind resource assessment is fundamental when selecting a site for wind energy projects. Wind is influenced by several environmental factors and understanding its spatial variability is key in determining the economic viability of a site. Numerical wind flow models, which solve physical equations that govern air flows, are the industry standard for wind resource assessment. These methods have been proven over the years to be able to estimate the wind resource with a relatively high accuracy. However, measuring stations, which provide the starting data for every wind estimation, are often located at some distance from each other, in some cases tens of kilometres or more. This adds an unavoidable amount of uncertainty to the estimations, which can be difficult and time consuming to calculate with numerical wind flow models. For this reason, even though there are ways of computing the overall error of the estimations, methods based on physics fail to provide planners with detailed spatial representations of the uncertainty pattern. In this paper we introduce a statistical method for estimating the wind resource, based on statistical learning. In particular, we present an approach based on ensembles of regression trees, to estimate the wind speed and direction distributions continuously over the United Kingdom (UK), and provide planners with a detailed account of the spatial pattern of the wind map uncertainty.

Keywords: Wind speed, wind direction, statistical learning, Weibull distribution, Random Forest, Lasso.

1. Introduction

Wind energy plays a key role in reducing the level of CO_2 emissions required to mitigate the worst effects of climate change. By 2020 the UK has pledged to produce 30% of its electricity from renewable sources [1], compared to 17.8% today [2]. With the depletion of conventional sources and the increase of global warming Renewable Energy Sources (RES) have attracted the interest of investors. Among all RES, wind energy has had a substantial growth over the last five years, reaching a global installed capacity of around 370 GW (gigawatts) at the end of 2014 with an overall turnover of 277 billion Euros [3]. Moreover, according to the latest statistics [4], electricity produced from onshore wind farms is becoming cheaper than other traditional sources of electricity such as nuclear, coal, and combined gas cycle. In the United States the unsubsidized levelized cost of 1 MWh (megawatthour) produced by onshore wind is already lower or equal to all other sources of electricity.

When selecting a site for investing in a wind energy project, wind resource assessment plays a fundamental role. Meteorological stations collect climate data, but they are sparsely located and therefore do not provide the full data coverage necessary for the optimal placement of wind farms. In order to obtain an estimate of the wind characteristics in unknown locations, a way to model the wind field is required. In the last decades, multiple models have been developed for this scope and the research in the field has focused on two main directions: numerical wind flow models (i.e. methods based on physics, also referred to as physical methods) and statistical methods. Physical methods model the wind field by solving physical equation, such as the equations that govern the mass and momentum-conservation laws, or computational-fluids dynamic models. Statistical methods on the contrary, estimate the wind resource by correlating past observations with environmental data, such as elevation, slope, and temperature. Both methods have been widely used in literature, at various scales and with different level of accuracy. Below we present an extensive overview of the literature to provide the reader with a classification of wind resource assessment methods.

1.1 Numerical Wind Flow Models

These methods estimate the wind resource by solving some of the equations that govern the motion of air in the atmosphere. Numerical wind flow models can be divided by level of sophistication or complexity [5] and partly also according to the scale at which they operate. In wind resource assessment we generally refer to three main scales of operation: macro-scale (known as synoptic scale with a resolution in the order of 2’000 kilometres or larger), meso-scale (few kilometres to thousands kilometres) and micro-scale (hundreds of meters to few kilometres). Synoptic scale models study large-scale phenomena, such as large depression fronts, which are mostly driven by Coriolis force and pressure gradient. These methods will not be treated in this review.

The first level of sophistication is occupied by mass-consistent models, such as NOABL (Numerical Objective Analysis Boundary Layer), developed in the ‘70s in the US [6, 7]. These methods solve only the equation of conservation of mass, which when applied to the atmosphere states that if a wind mass is forced over a slope it must accelerate so that the same volume of air passes in any given region [5]. Mass-consistent methods are still widely used for generating both meso-scale and micro-scale wind speed maps. Of particular interest is the work carried out in the UK by the UK Energy Technology Support Unit (ETSU) for the creation of a long-term wind speed database

[8]. They started from overlapping grids of 100 km of resolution, with data collected from 56 stations for a time period of 10 years, from 1975 to 1984. They then applied NOABL to downscale the map at 1 km of resolution at three heights: 10 m, 25 m and 45 m. To the best of our knowledge nowhere in literature there is a mention of the computational time needed to create the wind map mentioned above. However, since these long-term databases are updated very infrequently, the time needed to create them is somewhat not influential in the planning process for new wind farms. For micro-scaling these data would be used as look-up tables and their estimates would just be further downscaled, thus minimizing computational time. Regarding its accuracy, the technical report from Best et al. [9], created for the MET Office (UK Meteorological Office), shows a plot of wind estimations against weather observations from which the overall deviation of the estimates seems to be around 2 - 5 m/s. Moreover, another report from the MET Office [10] mentioned the bias of the estimates (i.e. the mean of the residuals' distribution) from this method as equal to 1 m/s.

The second level of sophistication is occupied by models, developed in the '80s and '90s, to include not only mass-conservation, but also momentum-conservation. These models are based on the theory advanced by Jackson and Hunt [11] and work by solving a linearized form of the Navier-Stokes equation governing fluid flows. Because of this characteristic these models are often referred to as linear wind flow models. Probably the most famous linear model is WaSP (Wind Atlas Analysis and Application Program [12]), developed by Risoe National Laboratory of Denmark and used to create the European Wind Atlas in 1989 [13]. The Jackson-Hunt theory assumes that topography causes small perturbations in an otherwise constant wind flow, this allows the equations to be solved efficiently [5]. WaSP incorporates techniques to account for obstacles and roughness changes, even though it is not equipped to handle complex terrains [5]. Despite its known limitations, WaSP has been and remains very popular in the industry and has been used to generate various wind speed maps globally [14 – 17]. Regarding the scale of analysis, WaSP can be used for both meso- and micro-scale modelling. In the late '90s for example, it was coupled with the Karlsruhe Atmospheric Mesoscale Model (KAMM) [18], to account for topography and create the first example of meso-micro scale model of the wind resource [19].

In alternative to linear models, the next level of sophistication consists of methods able to solve the full spectrum of equations of computational fluid dynamics (CFD) applied to air flows. These models take into account mass and momentum conservation, plus the effect of turbulence created by the interaction between wind and complex terrains. Examples of these model are based on Reynolds Average Navier Stokes (RANS) turbulence models [20, 21], and the Large-Eddy-Simulation (LES) model [22 – 25]. Additional information are provided in the comprehensive review of these models applied to fine-scale computation of wind flows carried out by Ayotte [26].

The final level of sophistication is occupied by Mesoscale Numerical Weather Prediction (NWP) Models [5]. These methods have been developed for weather forecasting; they include the full sets CFD equations, but they also include schemes to take into account: solar and infrared radiations, a soil model, clouds microphysics and convection. Examples of such models are: the Regional Atmospheric Modeling System (RAMS, http://rams.atmos.colostate.edu/rams-description.html), Skiron (http://forecast.uoa.gr/index.php), Weather Research and Forecasting (WRF, http://www.wrf-model.org/index.php), MM5 (http://www2.mmm.ucar.edu/mm5/overview.html), Consortium for Small scale Modeling COSMO (http://cosmo-model.cscs.ch/). These methods were developed to forecast weather patterns, based on the current situation. They are applied at the local and global scale, starting from direct weather observations from stations, radiosondes or satellite data. Due to enormous amount of equations to solve simultaneously these methods require substantial amount of computational resources to be used successfully, and cannot be used for micro-scale modelling with the current generation of supercomputers. For this reason, numerical approximations is often applied. NWP divide the atmosphere in 3D volumes (generally cubes) whose centres are used to define a three-dimensional grid. At each point the NWP solves weather parameter equations. For global models a resolution of 40 km and 40 vertical layers is often applied [27, 28]. For micro-scale modelling, on the contrary, these methods are generally coupled with other, faster to execute, algorithms. Examples include AWS Truepower's MesoMap and SiteWind systems [29], 3TIER's FullView system (http://www.vaisala.com/), the Risø National Laboratory's KAMM–WAsP system [19], and Environment Canada's AnemoScope system [30]. Al-Yahyai et al. [31] present a review of the use of NWP data for wind energy resource assessment, with particular attention on the accuracy of the methods. The wind data derived from NWP models are biased and in most of the studies considered, they underestimate about 5% the wind speed, in particular close to the surface. Moreover, according to their review, NWP models still present limitations due to: the simplifications in physics, the uncertainties in the initial state, in the lateral boundary conditions and in the surface characteristics. The study concluded that the overall wind speed bias ranged between 0.25 m/s and 2.5 m/s as function of the terrain's complexity.

These models have been used to simulate the wind field in a specific location in Spain and their performance compared in order to generate a benchmarking of wind flow models [32]. The conclusion of the study is that the downscaling at fine resolution of the global meteorological and topographical datasets is possible but the computational cost remains very high for engineering applications. The challenge is to find the trade-off between the physical downscaling at a specific resolution sufficient to capture the most relevant aspects of the local wind. The author suggests the linearization of the model to a finer grid to take into account the local spatial elements of the topography (such as hills generating the speedup effect). For this purpose, the study points out that CFD model can help in taking into account spatial elements such as hills or tree canopies for a more complete wind resource assessment including the wind direction’s rose, wind speed and atmospheric stability. The impact of stability and forest canopies on complex terrain flows was demonstrated with the Alaiz test case used to benchmark the models described in the study. An important final outcome of the study is that “advance models require systematic model evaluation processes in order to assess their impact on the chain of uncertainties of the wind resource assessment process. There is a lack of high fidelity experiments due to the difficulties of meeting the required spatial and temporal coverage of the relevant physical scales of the modelling system at a reasonable cost” [32].

1.1.1 Evaluation of Uncertainty in Numerical Wind Flow Models

The methods mentioned above have been proven over the years to be able to estimate the wind resource with a relatively high accuracy. However, weather stations, which provide the input data for wind resource assessment, are often sparsely located and do not offer a complete data coverage. This adds an unavoidable amount of uncertainty to the estimations, which physical methods are not always able to describe in details, as pointed out by Rodrigo et al. [32]. The error can be measured by comparing values estimated from the model, with real values observed at weather stations or in areas where we have a measuring mast specifically positioned for this analysis. This is the methods used for example in Gasset et al. [33] to test the methods used in the Canadian Wind Energy Atlas, based on NWP models. They compared their estimations with 10 measuring masts. They reported RMSEs (Root Mean Square Error) of 0.74 and 0.82, depending on the resolution of the land cover data they used (higher resolution equals better results). They also did not provide a clear indication of the computational time required for the two analyses.

Beaucage et al. [34] compared several types of physical methods on four different relatively small sites (maximum size 17x17 km) at a resolution of 50 m. They started from a minimum of four to a maximum of nine measuring towers. They achieved the following averaged RMSE 0.74 for CFD, 0.62 for WAsP, and 0.44 for NWP. However, the authors report that to estimate wind speed in these relatively small sites, they had to run the model for a minimum of 2.5 hours (for CFD) to a maximum of 864 hours (for NWP). Janjai et al. [35] used an atmospheric mesoscale model, mapping wind speed in Thailand. They validated their model by comparing their estimates with the weather observations, and computed RMSE value between a minimum 0.61 to a maximum of 1.34 m/s.

A more accurate representation of the uncertainty is provided by Weekes and Tomlin [36], who investigated the accuracy of a weather forecasting model in UK. They presented a method to calculate the uncertainty that arises from errors in the input parameters, not due to the model assumptions and simplifications. To achieve this, they employed a method based on quasi-random sampling and simulation. They basically simulated numerous outcomes of the model based on different combinations of input parameters and they report a range of uncertainty of around 35%. In another example [37], the same authors demonstrated that the error distribution is highly dependent upon the local terrain features of the estimation site. For example, for rural areas they can obtain the best accuracy with a minimum error of 0.44 m/s, while for more challenging environment, such as coastal areas, they report errors around 1 m/s. Their process is also a step forward in terms of computational time. They started their investigation using wind estimates produced on a 1 Km resolution using the NOABL models adapted to the UK by the MET Office (describe in section 1.1). From this they employed a downscaling technique based on terrain complexity and land-use, which is fast and can decrease the error in the NOABL data down to levels around 0.5 m/s.

Another method to assess the uncertainty of wind flow model due to the speedup effect related to the topography has been proposed by Clerc et al. [38]. The wind flow model used in the study is a combination of the well-established orography model MS3DJH/3R [39] and an empirical roughness and obstacle models [40, 41]. The model has been applied to two wind farms with different layouts. It couples the measurements of the masts and the speedup effect to adjust the estimate of the AEP (Annual Energy Production) of the WTs of the wind farms. The empirical functions provide robust estimates of uncertainty and correlation of errors both for sites with flat terrain and with forested and highly complex terrain. Nevertheless, the model is relative to the region of the wind farm and can be considered as an alternative to micro siting rather than to a mesoscale model. The uncertainties of the latter are the input reference wind speed, the regional and local aerodynamic parameters, the blending height and the Weibull shape factor required to construct a distribution of wind speeds.

1.2 Statistical Methods

Statistical methods have also been used for wind resource assessment [42, 43, 44, 45, 46, 47, 48, 49]. In this case the wind resource is generally estimated based on the spatial correlation between measured data and environmental predictors, e.g. topography and land-use. The first attempt in this direction was the study by Lorenc [42], which presented a three dimensional interpolation for multivariate geo-potential height, thickness and wind speed. Three different optimum interpolation techniques have been selected in the study to pre-process the observations, to check the data and to produce the grid-point values. The radiosonde and surface observations spread worldwide have been collected in order to extrapolate the wind speed at different elevations using physical relationship. In this study the authors report validation errors generally above 2 m/s.

In Luo et al. [43] seven spatial interpolation techniques such as trend surface analysis (TSA), inverse distance weighting (IDW), local polynomial (LP), thin plate spline (TPS), ordinary kriging, universal kriging and ordinary co-kriging were applied to generate daily mean wind speed surfaces of the UK. A set of wind speed observations between 1998 and 2002 of 189 weather stations collecting hourly values has been used. The predicted mean wind speed surface for a specific day has been generated on a regular grid of 5km resolution across the England and Wales land surface area. A leave-one-out cross-validation process was used in this study to assess the uncertainty of the map. This validation loops through the data and at each iteration excludes one weather observation from the dataset used to train the algorithm. Then the excluded observation is compared to what the algorithm estimates for the same location, and the error is measured in terms of residual. By applying this process, the authors report a RSME of 1.47m/s and a Mean Error (ME) of –0.01m/s using the co-kriging method, which was the best performer. The other methods show a RMSE higher than 1.61m/s and a ME comprised between –0.09m/s and 0.01m/s.

The MET Office also used geostatistical interpolation (i.e. “inverse-distance weighted interpolation of residuals from a multiple regression model”) to produce a long-term wind database (Met Office UK small and medium wind database) in the UK [44]. Here they gathered data from 230 weather stations belonging to the National Climate Information Centre (NCIC), for a time period ranging from 1981 to 2010 (for the latest version of the database dated 2013). Then they created a detailed wind speed database at 1 km of horizontal resolution, for heights ranging from 10 m to 45 m above ground. According to a recent MET Office report [10] the bias of this database is 0.4 m/s, calculated excluding 10% of the observations from the weather stations and re-predicting them (i.e. a 10-fold cross validation). This makes this database better than the one created with NOABL, which is why the MET Office suggests using this one instead.

Another example of kriging interpolation applied to wind resource assessment is provided by Cellura et al. [45]. In this work they created a wind map of Sicily using two different interpolation methods: inverse distance weighting (IDW) and universal Kriging. The measurements at 10m agl of hourly mean and maximum wind speed and direction of 29 weather stations in Sicily over a period of 3 years have been used as input data. Their validation showed that the universal kriging method is not adequate for the wind mapping as only 24% of the observations was well estimated.

These results were then used for comparison with a new method they developed called network residual kriging (NNRK) [46]. In this method a multi-layer perceptron is first used to estimate the wind resource correlating it with environmental predictors. The second step is a kriging interpolation of the residuals to further increase the accuracy of the predictions. The wind map generated at 10m agl (above ground level) was extrapolated to 50m agl using the CORINE land cover map. The extrapolated map has been compared to the Italian wind atlas generated with a mass-consistent model. The comparison shows that, although the spatial trend is similar, the NNRK underestimates the wind speed.

The wind field over the alpine region of Switzerland has been generated using a multiple kernel learning regression model [47]. This model finds correlation patterns between the wind resource and the environmental predictors, and then uses these patterns to estimate the wind resource in regions were only spatial data of the predictors are available.

The mean wind speed data at 10m agl of 148 weather stations over a 20 years period have been used. The wind speed has been extrapolated to 50m agl with a logarithmic profile using the roughness values according to the land cover category. The model shows test errors varying from 0.98m /s to 1.27m/s of the mean wind speed. A similar approach has been also presented by Douak et al. [48] used a statistical learning approach named Kernel ridge regression to estimate wind speed starting from 100 training samples, and obtained a minimum error of 1.4 m/s.

1.3 Comparison between numerical wind flow models and statistical methods

With the reviewed research we can draw some general conclusions by comparing the two type of methodologies: namely numerical wind flow models and statistical methods. Since we did not find any benchmarking experiment that specifically compare these two methods on the same dataset and with the same validation techniques, the only way we have to obtain information about their general accuracy is comparing the validation results presented in literature. However, we need to point out that validation results from different studies, performed in different areas and starting from different data, are difficult to be compared directly. This is because the type of data used, the topography of the terrain and the land-cover of the study areas have a great impact on the performances of every wind resource assessment model. Moreover, most of the studies we reviewed are not consistent in providing the measure of uncertainty of their work. In some cases, the validation results are not presented, and if they are presented there are studies in which the error is presented in percentage. These discrepancies do not allow us to properly compare our results with them and therefore we do not cite these studies. Moreover, the validation results are generally presented using the Root Mean Squared Error (RMSE), calculated as the square root of the average squared residuals. This index is particularly problematic as suggested by Willmott and Matsuura [50] and Fekete [51], who suggest that RMSE should be avoided. However, for the purpose of this study we are only interested in knowing average values of RMSE for different methods to see if our results are comparable with literature studies. For this reason, even though RMSE is not the best choice, it may still provide us with a way to compare our results with previous studies.

From an accuracy perspective, we can say that numerical wind flow models seem to provide better prediction for increasing level of sophistication. From our review NWP reported average errors around 0.5 m/s, while lower level of sophistication had accuracies around 0.6 – 0.7 m/s. These results cannot be generalized, as suggested by Brower [5], but in reviews that used several methods in parallel these values seem to hold. Regarding statistical methods, the studies we reviewed always report RMSE around or above 1 m/s. Therefore, it appears that numerical wind flow models are in fact more accurate in estimating the wind resource. One possible drawback is that the level of sophistication is directly proportional to the time and amount of resources needed to complete the estimation process. For this reason, an accuracy level of around 0.5 m/s, for large areas, could translate in months of analysis carried out on supercomputers. Beaucage et al. [34] for example in their study report that to achieve a RMSE of 0.44 m/s, the NWP model had to run for a total of 864 hours (for an area of 17 x 17 Km, at 50 m of resolution). A way to solve this issue is to use long-term databases and downscaling techniques. Weekes and Tomlin [37] developed a method to downscale meso-scale estimates from a mass-consistent model. The authors suggest this method is relatively quick and efficient to run, because this method was based on the statistical downscaling of long-term average wind speed data. Even though a time figure is not provided, we can assume that it would probably be similar to the time needed for statistical wind resource assessment.

Statistical methods are certainly more time and computationally efficient, compared to numerical wind flow models, but from literature we can conclude that they are generally less accurate. Moreover, while numerical wind flow models can estimate the wind resource calibrating the model with very few weather stations, Gasset et al. [33] for example used 10 masts; statistical and geostatistical methods, since are based on correlating wind data with environmental predictors, require large amount of data. For example, ordinary kriging needs at least 100 observations to compute the variogram with the method of moments [52]. Moreover, if the dataset does not cover all the types of terrain present in the study area, chances are the statistical method would not perform well. For example, a known limitation of random forest is that it cannot estimate values of the variable outside the range of the training data. For this reason, if we do not have samples in a particular type of terrain, the estimates there would have high uncertainties.

Another important different between numerical wind flow models and statistical algorithms is the ability of the latter to assess their own accuracy. A classic example is the ability of ordinary kriging of providing an estimate of the variable of interest plus a measure of the variance of the estimates. This variance is calculated from the variogram according to the location of the nearest data point and provide the user with a local measure of the reliability of the map. This is something that numerical wind flow models cannot provide directly. Weekes and Tomlin [36] in their study talk about site-specific uncertainty since they provide an assessment of the accuracy of their methodology for various land-use type over the study region. For achieving this the authors compared the estimates provided by their model with direct observations of wind speed for each of these locations. They report that for rural areas they can obtain the best accuracy with a minimum error of 0.44 m/s, while for more challenging environments, such as coastal areas, they report errors around 1 m/s. However, we think that these results are difficult to generalize for the entire study area. In other words, we think it is difficult to argue that the range of uncertainty obtained for rural areas represents the full range of uncertainty for each area in the country with the same land-use. That is probably one of the reasons why a map of the spatial pattern of the uncertainty is never presented in literature, to the best of our knowledge. This is where statistical methods may have an advantage over numerical wind flow models. These algorithms can assess their own accuracy directly, without the need for any additional operation. For this reason, we can present the user with a detailed map of the local reliability of the estimates we are presenting. This is what we are referring to when we talk about local uncertainty in this article.

The main purpose of this research is to demonstrate empirically that with the advancements in remote sensing that provide us with free access to numerous environmental raster data, we can potentially increase the accuracy of statistical methods. Moreover, there are now algorithms that can take advantage of all these environmental data to create models that can potentially be very accurate for wind resource assessments. These algorithms belong to the class referred to as statistical learning, in which the algorithm is trained based on correlation between the variable and environmental raster data. This way we can achieve results comparable with physical methods in terms of accuracy, while reducing the computational time to a minimum and having access to a precise local uncertainty estimation. To demonstrate this we created a framework to support planners during the feasibility study in order to identify locations suitable for wind energy projects, without the need of additional time-consuming wind measurements campaigns. Additionally, we created a map that, alongside the wind speed and direction distributions, shows the spatial distribution of the uncertainties of the prediction. This would provide planners with more detailed information to refine the estimate of the future energy production of selected sites and to rank them as a function of their economic risk related to their uncertainties. In general, the delivered map would provide a tool for a detailed spatial evaluation of the investment risk related to the exploitable wind resource.

2. Materials and Methods

2.1 Study Area and Dataset

This research was conducted in the United Kingdom over a total area of 244,119 km^2 at 1 km resolution. Wind data were obtained from 188 stations across the United Kingdom. The data are part of the MIDAS (Met Office Integrated Data Archive System) Wind Data Archive and are freely available for research purposes from their Website [53]. The locations of the 188 weather stations are referred to as training locations and visualised as red points in Fig. 1. This dataset is composed of long-term wind speed measurements for a time period between 2009 and 2013, taken at hourly intervals.

As Luo et al. [43] pointed out, MIDAS stations do not record the same amount of data every day. The total number of stations we downloaded from the MET office contained 580 weather stations; of these, the large majority records only one value each day. Others record few data during the course of the day and cannot be considered valuable for our research. If we consider only stations with at least 20 hours per day we obtain a total of 212 stations. However, in many cases these recording are unusable since they contain numerous NAs. So we needed to add an additional filtering step to check the reliability of the data and remove stations where too many NAs were clustered in particular time frames or seasons. We did that by checking the time series data for any significant gaps and remove the stations in which the sampling frequency was unacceptable for prolonged periods of time. Of the 212 remaining stations, only the 188 we finally used matched these criteria.

Several covariates, or predictors, were used to perform the estimation. In particular, the Aster DEM (30 m resolution) provided by NASA [54] was used for elevation data and to create DEM derivatives, such as slope, aspect, and roughness, computed in SAGA GIS [55]. The land-use raster data (100 m resolution) were provided by the CORINE project [56]. In addition, we used raster maps at 5 km resolution from the MET office [57] with meteorological data: mean annual temperature, maximum and minimum temperature, mean wind speed, mean atmospheric pressure, air frost, cloud cover, rainfall, and relative humidity.

2.2 Wind Speed and Direction

Wind speed is usually measured by meteorological stations and towers, which belong to national meteorological services, and are placed at airports or in proximity of wind farms. The wind speed characteristics are usually (but not only) measured at 10 m above ground level. In order to assess whether the wind speed at a given location is economically viable, statistical analyses of wind data are carried out to quantify its probability distribution. The probability distribution describes the likelihood that a given value will occur, therefore the longer the data collection the more reliable the probability distribution. For wind speed this distribution is generally described by the Weibull distribution. This is not always the case, and sometimes the Weibull is not a good approximation of the wind speed distribution, even though in a majority of cases this approximation holds. The Weibull distribution, which is a special case of the generalised gamma distribution [58], is a two-parameter continuous probability distribution [59] that is defined by the following density function:

$$f(x;C;k) = \begin{cases} \frac{k}{C}\left(\frac{x}{C}\right)^{k-1} \quad e^{-(x/C)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (1)$$

where *x* is the wind speed (in m/s), *k* is the shape parameter and *C* is the scale parameter. In this equation the parameter *x* refers to long term wind speed measurements. In fact, we need at least five years of data to accurately fit a Weibull distribution.

Fig. 2a and 2b present the influence of *C* and *k* on the density function, while Fig. 2c shows an example of a typical wind distribution with the fitted Weibull density function. These plots were creating using computer generated data, just to show readers how a Weibull distribution looks like and its sensitivity to changes in the two parameters, shape and scale. Generally speaking, a variation of the *C* parameter, keeping *k* constant, directly affects the mean wind speed that increases proportionally with *C*. In contrast, a variation of *k*, keeping *C* constant, produces a decrement in the dispersion of the measurements around their mean. These two parameters uniquely identify a Weibull distribution, and therefore were used to estimate the wind speed spatially as a function of the environmental predictors. Basically for each location in our training set we fitted a Weibull to the wind distribution and calculated the two parameters. These would become our variables of interest for estimating the wind speed distribution in locations where no observations are available.

For wind direction the approach followed with speed is not possible. We tested the use of circular distributions to describe the wind directions observed at weather stations. However, the parameters of circular distributions were not as correlated with the environmental predictors as the parameters of the Weibull distribution. If this correlation is lacking, statistical methods will provide poor mapping accuracy. For this reason, wind direction was mapped using the frequencies of occurrence of selected direction intervals. Basically, for each training location we estimated the histogram of the wind direction, using 18 bins (meaning intervals of 20°). For each bin we recorded the frequency of occurrence and then we used the 18 frequencies as variables for the mapping algorithm. This way we estimated in space each of the 18 frequencies as a function of the environmental predictors.

2.3 Statistical Learning Approach

Statistical learning is a branch of statistics aimed at modelling and understanding complex datasets [60]. Generally speaking, statistical learning aims at estimating a target variable, based on a set of inputs, or predictors. In mathematical terms this can be expressed as follows:

$$Y = f(X) + \epsilon \qquad (2)$$

where *Y* denotes the target variable, *X* denotes the predictors (in this case environmental data available in raster format), and *ε* is a random error component, which depends on various factors such as measurements errors, and it is the irreducible part of the model [60]. In this research Eq.2 was solved for a total of 20 times, each with different variables but with the same set of predictors. First we solved it using as variables shape and scale, and then using the 18 frequency values of the histogram bins. The local wind speed distribution is influenced by many environmental data, such as topography [61] and land-use [62, 63]. Since we assume a correlation between the variable and the predictors, we can try to estimate our target variables as a function of the environmental predictors. The problem with working with many variables, as we do in this work, is that they present different levels of correlation with the environmental predictors. This may lead to a decrease in the accuracy of the estimation, since there may be predictors that if included in the model may lead to erroneous estimates. Moreover, having too many predictors cause the model to not be easy to interpret, since it is difficult to identify the relative importance of each covariate [60]. For this reason, the first step we took involved the use of Lasso [64], a technique to filter the predictors and keep only the most important to our analysis. This method is based on a linear model. It was possible to use it because from a preliminary correlation analysis we determined that many predictors were linearly correlated with our variables. The Lasso works by fitting a linear model to solve Eq. 2, with the difference that it does not use the residual sum of squares (RSS) to calculate its coefficients. Instead, it multiplies the standard residual sum of squares by a penalty that shrinks the value of some coefficients toward zero, thus excluding them from the estimation. This allowed us to work only with the most correlated coefficients, hence speeding up the computations and making the results more interpretable.

A crucial objective of this study was the estimation of the site-specific uncertainty. For this reason, for estimating wind speed we needed an algorithm capable of assessing its own accuracy. We selected Random Forest (RF; [65]) as our mapping tool, which is based on ensembles of regression trees. These types of algorithms create regression trees by testing different predictors in order to find the one split that minimises the RSS in the resulting subsets, or tree leafs [66]. A regression tree can be viewed as a series of “if-then” rules that are used to define classes of probabilities. The prediction in locations where no wind data are available is performed by running the predictors through the tree in order to define the most probable value for that particular location.

Regression trees have the advantage of being very easy to explain and interpret. The regression tree can be graphically displayed and this is a great advantage compare to other methods. However, regression trees generally have a predictive power lower than other approaches [60]. RF solves this by using bagging and decorrelating the single trees. Random Forest instead of using the entire dataset to build one single regression tree, uses the bootstrap to build numerous trees, starting from the same training dataset. Bootstrapping is a statistical resampling method, which takes a dataset with *n* observations and resamples it randomly with replacements, meaning that an observation can occur more than once in bootstraping samples. This produces a series of samples, of length *n*, to which the algorithm can fit a regression tree. This way RF can fit numerous regression trees to the same dataset; this procedure, technically referred to as bagging, reduces the variance of the method and therefore increases its accuracy [60].

Random Forest has another advantage compared to pure bagging, it subsets the predictors at each split of the tree thus decorrelating the ensemble. This is a strong advantage that leads to higher accuracy, since it avoids problems with multicolinearity. The reason is simple, suppose we have one predictor that is strongly correlated with wind speed. If we allow the algorithm to choose among all predictors, most or all the trees will use this strong predictor. As a consequence all the trees will be highly correlated and this does not lead to a substantial reduction of variance [61]. Random forest overcomes these problems and reaches a higher estimation accuracy.

Random Forest has been widely used in research for different purposes, such as digital soil mapping [67, 68], ecology [69], geomorphology [70], and remote sensing [71, 72]. Its popularity is due to the fact that it can generate reliable estimates and it is robust against noise in the predictors [67], which is a crucial aspect when dealing with environmental covariates. For these reasons, RF is the perfect tool for wind speed mapping. In each predicted location, RF produces a set of estimations, depending upon the size of the forest. All the predicted values can be used to determine the variance of the Weibull parameters, from which we can determine the local uncertainty of the wind distribution.

We used the Mean Absolute Deviation (AD) to compute the spread of each estimated variable around the arithmetical average. AD is calculated according to the formula proposed by Hoaglin et al. [73]:

$$AD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad (3)$$

where n is the number of wind mean speed values computed from each bootstrapping set, x_i is the i^{th} value of the wind mean speed in a set of values from 1 to n, and $\bar{x}$ is the sample arithmetical average. This parameter is more robust compared to the standard deviation, particularly for long-tailed distributions [73]. The mean absolute deviation was also used to produce all the uncertainty maps shown in this manuscript.

2.4 Uncertainty Estimation

For wind speed mapping, as mentioned in section 2.3, RF produces a set of estimations of shape and scale for each predicted location. The mean values of the distributions of shape and scale can be used to create the Weibull distribution of the wind speed for each unknown location. We could simply calculate the deviation around the mean of the Weibull distribution to obtain a measure of the average fluctuations of the wind resource. However, standard deviation *per se* does not provide us with a measure of the local accuracy of the RF estimates. For shape and scale, measuring the deviation around the mean of their distributions, since they are directly estimated by RF, provides us with a measure of uncertainty. The higher the variance of these distributions, the higher is the local uncertainty. But if we just take the mean vales of these two parameters to fit a Weibull, we are not estimating how this uncertainty is propagated from the estimated parameters to the wind speed distribution. For doing so we need to rely once again on bootstrapping simulation. The bootstrap allows us to randomly simulate hundreds of distributions of shape and scale, with repetitions; for each new pair of distributions we can compute their means and use them to fit a new Weibull, which would be slightly different from the other. Each simulated Weibull would have a mean wind speed and the amount of differences between all the simulated wind mean speeds would be proportional to the amount of uncertainty we have in that particular location for the RF estimates. If we calculate the spread of all the wind mean speed values calculated after each simulation we obtain the standard error of the mean, which tells us how confident we are that the wind mean speed we present on the map is the real mean of the wind distribution. This is what is generally referred to as confidence interval, and allows us to provide a measure of the site-specific, or local, accuracy of the map. The smaller the error is the higher is our confidence that the mean we calculated is the real mean value of the dataset. With this method we are not only able to compute the confidence interval of the mean wind speed, but we are able to compute it for the entire wind speed distribution. This is a crucial information for planners, since it allows them to calculate the potential range of variation of the energy output based on the map uncertainty. This is a level of details we can reach only by assessing the error propagation from the RF estimates to the wind resource.

For wind direction there was no need to use an additional simulation to determine its uncertainty, because we estimated the frequency of each bin of the direction's histogram (with bins of 20°). Since RF produces a measure of its own uncertainty by providing us with a set of estimates, we can use it to determine the amount of error we have for each bin of the histogram. In this case there is no propagation to compute because what we obtain from RF is what we are presenting on the map, i.e. the wind rose of the bins estimated from RF. For this reason, as a measure of the map error, we present the average deviation of all the frequency bins estimated by RF. The bigger this number is, the less reliable are the RF estimates for wind direction in that particular location on the map.

2.5 Validation

A crucial part of every modelling experiment is the estimation of the overall error of the model. This process allows us to compare our model with real values, i.e. weather observation, and compute its accuracy. Statistical learning works by using the data we collected from the weather stations to create a function to solve Eq. 2. For wind speed for example, the algorithm correlates the Weibull parameters with the environmental predictors in order to define *f(X)* that estimates *Y* as closely as possible. This is referred to as the training process of the algorithm. The crucial aspect of every statistical learning exercise is trying to estimate the accuracy of *f(X)* when used to estimate locations where we do not have any information, i.e. locations that were not part of the training set. This is referred to as assessing the test error. We can compute the test error by using cross-validation, which is a process in which we split the training set and then we use part of it to estimate the remaining. To clarify, the training includes all the data we extracted from the 188 weather stations, such as shape, scale and the frequency of the direction histogram. In addition, the training set includes all the values of the environmental predictors extracted from the raster data in the location of the weather stations. In this study we performed a 5-folds cross-validation, in which this training dataset is divided into five parts (or folds). Four of these folds are used to train the statistical learning algorithm, while the remaining fold (around 20% of the dataset) is used to test its accuracy, i.e. calculate the test error. This same process is then repeated until each fold is used for testing. Because the folds are chosen at random, we decided to repeat the cross-validation process 100 times, in order to have a more reliable estimation of the test error. The comparison between observed and predicted values was performed using RMSE, since it is widely used in other papers to present the same result.

For the wind speed we trained the Lasso and RF on four folds, then we predicted the shape and scale Weibull parameters in the remaining fold. For each test location we then used the two Weibull parameters to simulate a wind speed distribution and then compared the estimated mean speed with the observed values, calculating the RMSE. For the wind direction, the approach is similar, for each location in the test set we estimated the histogram's frequencies. Subsequently we use the estimated distribution to calculate the mean wind direction and we compared it with the mean direction observed in the data, computing the RMSE.

2.6 Software

This experiment was performed entirely using the programming language R [74]. For fitting the Weibull distribution to the weather observations we used the package fitdistrplus [75], which uses maximum likelihood. For the Lasso we used the function glmnet, available in the package glmnet [76], which includes a cross-validation function that is used to optimize the variable selection process. For RF we used the function randomForest, available in the package randomForest [77]. This function has some settings that can be modified by the user to obtain more reliable estimates. The first is the number of trees to create. Since RF is based on bootstrapping, which is a random process, we need to fit numerous trees to obtain accurate estimates. According to Grimm et al. [67] fitting 1000 trees should be sufficient to achieve stable estimates, and this is what we did here. Another parameter that needs to be optimized is the number of predictors to exclude each time a regression tree is fitted. In this case the package randomForest present a function, tuneRF, which allows to select a range of values and test their accuracy using the internal out-of-bag validation available in RF [67].

3. Results and Discussion

As mentioned in the introduction, the proposed method performs the spatial prediction of the wind resource using a statistical learning approach that allows estimating both wind speed and direction distributions with the corresponding local uncertainty. For estimating the wind speed, we fitted a Weibull distribution to each weather station of our dataset, shown in Fig. 1. The mapping process was performed using the method described in section 2.3. The RF estimations of the scale factor *C* and the shape factor *k* are shown in Fig. 3A and 3B, while the uncertainties, as the AD of the set of estimations produced by RF for each location, are depicted in Figure 3C and 3D. It is interesting to notice that the Weibull parameters present almost opposite spatial patterns: where shape is low, scale present high values and vice versa. The uncertainty patterns follow the same rules: where shape presents low uncertainty, scale presents higher values.

As mentioned, RF uses the bootstrap to resample the dataset and create a series of trees for each location in the prediction grid. These parameters and their related uncertainties were then used to fit a Weibull distribution for each location on a 1 Km grid. Once we have the full wind distribution we can compute its mean and AD to provide planners with the average wind speed and a measure of the fluctuation of the wind resource. These indexes are presented in Fig. 4A and 4C. For wind direction, RF estimated the frequency of bins of 20°. From the RF estimates we can create a circular distribution of the wind speed, and from it calculate the mean wind direction and its AD, which are presented in Fig. 4B and 4C.

An advantage of statistical methods compared numerical wind flow models is that the former have ways to provide end users with a measure of the site-specific uncertainty of the estimated wind resource. As mentioned in section 2.4, we developed an approach to provide practitioners with a measure of the local uncertainty of the map. This information is crucial because it allows us to pin point areas where the map is less reliable and therefore the estimates of wind speed and direction that we provide should be used carefully, and maybe more data would be required to increase the local accuracy. The error maps are presented in Fig. 5. The spatial pattern of the error is similar for both wind speed and direction. In general, the error is larger in areas where we have a relatively low data density, such as in Scotland and Wales. Statistical methods rely heavily on data coverage to increase the accuracy of the map. If the data we have do not cover the full spectrum of terrains that we are going to encounter in the mapping area, the accuracy of the method will be negatively affected. In this case, in mountainous areas of Scotland and Wales the data coverage is scarce, and this increases the uncertainty of our estimations. The advantage of this method though is that it can clearly identify these problematic locations so that planners can take this into consideration. The confidence interval of the wind mean speed is presented in Fig. 5A. Uncertainty values are generally very low, with a mean value of 0.07 m/s, and a maximum value of 0.20 m/s. These values may seem negligible but if we put them into context we realise that a 0.20 m/s error in just one site means that in that area we have an average error that is almost one third of the average cross-validation error. This is definitely a very informative value to have during a feasibility study, since it allows us to determine that in such areas the likelihood of the estimated wind speed to be accurate is low and therefore an additional investigation of the wind resource is necessary before planning wind facilities here. This is a crucial information that is not directly available when physical methods are used. For example, Weekes and Tomlin [36] in their research mentioned the fact that the average error they obtained was lower in rural areas, and substantially higher in coastal areas. However, in their case they did not have a way to define the spatial pattern of this error, therefore the only thing their results may suggest to planners was caution in coastal areas over the whole country. With a detailed uncertainty estimation we are able to identify the coastal areas where the estimation present lower accuracy values and distinguish those to coastal areas were, on the contrary, the accuracy is high. For example, from Fig. 5A we can easily see that in England no coastal area present uncertainties higher than 0.1 m/s, while in North Wales and Scotland we have areas that present values of uncertainty in the highest range.

The approach described in section 2.5 was used to provide a measure of validation for the mapping process. The results indicate that for wind speed the method used in this research has a RMSE of 0.70 m/s, a Mean Absolute Error (MSE) of 0.5 m/s, and a bias of -0.01 m/s. This is the overall error of the map in comparison to real weather observations. This value can be compared to previous works that employed both physical and statistical methods to map the wind resource. Even though we cannot directly correlate the results achieved in literature with ours, because of different settings and because RMSE is not the best index for comparisons [50, 51], we can still compare our findings with the literature in order to assess where our work fits into the state of the art of wind resource assessment. According to the review of previous research we presented in section 1, the value of RMSE we obtained is, to the best of our knowledge, lower than any other previous test involving statistical or geostatistical algorithms. The lowest value recorded in literature was the 0.98 m/s, achieved by Foresti et al. [47] in Switzerland, which is probably more challenging to map than Britain because of the different topography. However, Luo et al. [43] used the same identical dataset to compare several geostatistical algorithms and reported a minimum RMSE value of 1.47 m/s, double what we obtained here. Moreover, our results can also be directly compared to the accuracy of the Met Office UK small and medium wind database, which report a bias of 0.4 m/s [10]. We were able to achieve an accuracy much lower than theirs, and with a validation process that excluded 20% of the observations. This means that the long-term wind map we generated can potentially replace the one in use today.

This improvement is certainly due to the amount of environmental covariates we used in this research. Nowadays, remotely sensed covariates are publicly available and characterized by a very good spatial resolution. This revolution in open data has the potential of increasing the reliability of estimation methods based on geostatistics, statistical learning and machine learning. However, these class of algorithms will always be as good as the dataset used for training. If the training dataset does not properly cover all the characteristics of the area under study, the results will always have higher margins of error for certain zones, which is exactly what happens here for the Scottish Highlands. Physical methods can potentially overcome these limitations, since they require relatively few sampling stations to perform the wind resource assessment. These methods are the industry standard because they are thought to be the best way to assess the wind resource over large regions. However generally speaking, numerical wind flow models requires a substantially higher amount of computing resources and time, compared to statistical algorithms, to produce a wind speed map at national scale. In this research, the training process of the statistical model and the estimation, at 1km of resolution, over an area of 244'119 Km^2 took 1.81 minutes. The uncertainty estimation took longer (around 1 day), because the bootstrap needs to be run for a sufficient number of times to obtain meaningful results (in this case 500 times), but the whole process took in total just a small fraction of the time reported in literature for wind mapping with deterministic methods.

Regarding wind direction, the same cross-validation approach was adopted to test the accuracy of the RF estimations. For each test location we estimated the full direction histogram and then compared the observed and the estimated mean wind direction. The result is a RMSE of 9.5°. We did not find any mention to an accuracy estimation for wind direction estimates in our review of previous research, therefore we do not have ways to compare our results with literature. However, for this parameter we used the same robust validation methods we employed for wind speed and the results indicate that the error is below the threshold of 20° we used to divide the direction histogram. These information allows us to have a relatively high level of confidence regarding the reliability of the wind direction map.

In Fig. 5, we present two maps to provide a quick way of visualizing the spatial pattern of the wind speed and direction uncertainties. However, since we estimated the full distribution of both wind speed and direction, showing one single index for each location is extremely simplistic, compared to the amount of information that we can generate from each predicted location in the wind map. In fact, with the approach described in section 2.4 we can determine precisely the propagation of the estimation error to the full distribution of both wind speed and direction. An example of the kind of output we can produce for each single location at 1 Km of resolution is presented in Fig. 6. One of the main issues when selecting a site for a wind energy project is a spatial assessment over large regions with regard to the investment risk, which includes the wind resources available, the financial risk, and the uncertainties related to construction and operations [78]. If one excludes the risks related to financial aspects, in addition to construction and maintenance failures, which are not addressed in this work, then an optimal pre-feasibility assessment of wind resources is critical and fundamental because it is subject to various uncertainties and can thus significantly impact the success of a project [79]. At small scale, the estimate of the uncertainties of the wind characteristics can be integrated into a GIS based model when quantifying the uncertainties in wind energy production and the impact of wake effect within a wind farms [80]. These uncertainties can be directly plugged into equations to estimate the potential energy output of wind farms located in these areas. This would allow planners to precisely estimate the amount of electricity available to produce plus the confidence intervals of their results.

4. Conclusions

In this work we presented an approach based on statistical learning to spatially estimate wind speed and direction distributions. The validation results demonstrate that this method produces better results compared to any previous example of statistical methods applied to wind resource assessment, and comparable with studies that used numerical wind flow models. Moreover, since this method is capable of assessing its own accuracy, we were able to create a map of the local estimation uncertainty, which is something that only statistical methods are able to provide. However, this aspect is crucial for planning wind farms, because in areas not properly covered by wind speed measurements the site-specific accuracy of the map may well be relatively low and this means that the wind distribution may be subject to more fluctuations than reported on the map. With this method we are able to pin point these problematic locations and consequently warn planners that more data are required for obtaining an accurate assessment. Moreover, the map uncertainty allows planners to precisely propagate the wind resource error to the power output for a potential site in order to understand its impact on future revenues. We did not cover this aspect in this paper because we do not have access to power output data from existing wind farms, which would be required to validate our results. However, this is something we are planning to explore in the future.

The availability of numerous remotely sensed environmental covariates allowed statistical methods to reach a point in which their accuracy level is comparable with much more sophisticated weather forecasting algorithms. However, statistical methods present several advantages compared to physical estimators. They are faster and computationally more efficient, and they can also assess precisely the site-specific uncertainty in each estimated location of the map. Moreover, statistical methods have clear advantages over CFD algorithms, since they require less computer power and time to execute. Based on the promising results of this study, we will explore new ways of combining both physical and statistical methods into hybrid algorithms in the future.

Acknowledgements

The authors would like to thank the UK Meteorological office for providing the wind speed data for this research and some of the covariates. Other data providers we would like to thank are: NASA for the Aster DTM and the EU for the land-cover raster.

References

[1] Clark P. Green energy auction sets stage for cheaper solar and wind power. 2015, Financial Times.

[2] Department of Energy & Climate Change. Energy trends and prices statistical. 2014 - https://www.gov.uk/government/news/energy-trends-and-prices-statistical-release-25-june-2015 – Last accessed: 27.02.2015

[3] Council, G.E Global wind report - Annual market update 2014 - http://www.gwec.net/wp-content/uploads/2015/03/GWEC_Global_Wind_2014_Report_LR.pdf - Last accessed: 16.11.2015

[4] Lazard. Lazard's Levelized Cost of Energy Analysis – Version 8. www.lazard.com/PDF/Levelized Cost of Energy - Version 8.0.pdf – Last accessed: 27.02.2015

[5] Brower, M. Wind resource assessment: a practical guide to developing a wind project. John Wiley & Sons; 2012.

[6] Traci, RM, Phillips, GT, and Patnaik, PC. Developing a site selection methodology for wind energy conversion systems. Final report, 15 June 1977-15 September 1978 (No. DOE/ET/20280-3). Science Applications, Inc., La Jolla, CA (USA).

[7] Phillips, GT. Preliminary user's guide for the NOABL objective analysis code. Special report, 15 June 1977-15 June 1978 (No. DOE/ET/20280-T1). Science Applications, Inc., La Jolla, CA (USA).

[8] Burch, SF, and Ravenscroft, F. Computer modelling of the UK wind energy resource: Overview report. Energy Technology Support Unit Report WN7055, UK Department for Business Enterprise and Regulatory Reform; 1992.

[9] Best M, Brown A, Clark P, Hollis D, Middleton D, Rooney G, Thomson D and Wilson C. Small-scale wind energy Technical Report. MET Office; 2008.

[10] Technical bulletin - UK Wind Map - site search for small and medium wind - http://www.metoffice.gov.uk/media/pdf/l/7/14_0058_Site_search_for_sml_med_wind_projects.pdf - Last accessed: 16.11.2015

[11] Jackson PS, Hunt JCR. Turbulent wind flow over a low hill. Quarterly Journal of the Royal Meteorological Society 1975; 101(430):929-955

[12] Bowen AJ and Mortensen NG. Exploring the limits of WAsP: the Wind Atlas Analysis and Application Program, in European Union Wind Energy Conference, 1996. Göteborg, Sweden p. 584-587.

[13] Troen I. and Petersen EL. European wind atlas. 1989.

[14] Bilgili M, Şahin B and Kahraman A. Wind energy potential in Antakya and Iskenderun regions, Turkey. Renewable Energy, 2004; 29: p. 1733-1745.

[15] de Araujo Lima L and Bezerra Filho CR. Wind energy assessment and wind farm simulation in Triunfo - Pernambuco, Brazil. Renewable Energy, 2010; 35: p. 2705-2713.

[16] Himri Y, Himri S. and Boudghene Stambouli A. Assessing the wind energy potential projects in Algeria, in Renewable and Sustainable Energy Reviews 2009; p. 2187-2191.

[17] Radics K and Bartholy J. Estimating and modelling the wind resource of Hungary, in Renewable and Sustainable Energy Reviews 2008; p. 874-882.

[18] Adrian, G. Synthetic wind climatology evaluated by the non-hydrostatic numerical mesoscale model KAMM. Environmental Meteorology. Springer Netherlands, 1988; p. 397-411.

[19] Frank HP, Rathmann O, Mortensen N, Landberg L. The Numerical Wind Atlas, the KAMM/WAsP Method. Riso-R-1252 report from the Risoe National Laboratory, Roskilde, Denmark; 2001. p. 59.

[20] Palma JMLM, et al. Linear and nonlinear models in wind resource assessment and wind turbine micro-siting in complex terrain, in Journal of Wind Engineering and Industrial Aerodynamics 2008; p. 2308-2326.

[21] Undheim O, Andersson HI and Berge E. Non-linear, microscale modelling of the flow over Askervein hill. Boundary-Layer Meteorology, 2006; 120: p. 477-495.

[22] Bechmann A and Sørensen NN. Hybrid RANS/LES applied to complex terrain. Wind Energy, 2011; 14: p. 225-237.

[23] Fröhlich J and Terzi DV. Hybrid LES/RANS methods for the simulation of turbulent flows. Progress in Aerospace Sciences, 2008; 44: p. 349-377.

[24] Porté-Agel F, et al. Large-eddy simulation of atmospheric boundary layer flow through wind turbines and wind farms. Journal of Wind Engineering and Industrial Aerodynamics, 2011; 99: p. 154-168.

[25] Silva Lopes A, Palma JMLM, and Castro FA. Simulation of the Askervein flow. Part 2: Large-eddy simulations. Boundary-Layer Meteorology, 2007; 125: p. 85-108.

[26] Ayotte KW. Computational modelling for wind energy assessment. Journal of Wind Engineering and Industrial Aerodynamics, 2008; 96: p. 1571-1590.

[27] DWD Deutschen Wetterdienstes. Annual report. 2009.

[28] National Center for Atmospheric Research. Annual Report. 2008.

[29] Brower M. Validation of the WindMap Program and Development of MesoMap. In: Proceeding from AWEA’s WindPower conference; Washington (DC); 1999.

[30] Yu W, Benoit R, Girard C, Glazer A, Lemarquis D, Salmon JR, Pinard J-P. Wind Energy Simulation Toolkit (WEST): a wind mapping system for use by the wind-energy industry. Wind Eng 2006;30:15–33.

[31] Al-Yahyai S, Charabi Y, and Gastli A. Review of the use of numerical weather prediction (NWP) models for wind energy assessment, in Renewable and Sustainable Energy Reviews 2010; p. 3192-3198.

[32] Rodrigo JS, et al. Benchmarking of wind resource assessment flow models. The Alaiz complex terrain test case. 2013.

[33] Gasset N, Landry M, Gagnon Y. A comparison of wind flow models for wind resource assessment in wind energy applications. Energies 2012; 5(11):4288-4322

[34] Beaucage P, Brower MC, Tensen J. Evaluation of four numerical wind flow models for wind resource mapping. Wind Energy 2014; 17(2):197-208

[35] Janjai S, Masiri I, Promsen W, Pattarapanitchai S, Pankaew P, Laksanaboonsong J., Bischoff-Gauss I, and Kalthoff N. Evaluation of wind energy potential over Thailand by using an atmospheric mesoscale model and a GIS approach. Journal of Wind Engineering and Industrial Aerodynamics 2014; 129: 1-10.

[36] Weekes SM, and Tomlin, AS. Low-cost wind resource assessment for small-scale turbine installations using site pre-screening and short-term wind measurements. IET Renewable Power Generation 2014; *8*(4), 349-358.

[37] Weekes SM, and Tomlin AS. Evaluation of a semi-empirical model for predicting the wind energy resource relevant to small-scale wind turbines. Renewable Energy 2013; 50 (2013): 280-288.

[38] Clerc, A., et al., A systematic method for quantifying wind flow modelling uncertainty in wind resource assessment. Journal of Wind Engineering and Industrial Aerodynamics, 2012. 111: p. 85-94.

[39] Walmsley, J.L., P.A. Taylor, and T. Keith, A simple model of neutrally stratified boundary-layer flow over complex terrain with surface roughness modulations (MS3DJH/3R). Boundary-Layer Meteorology, 1986. 36: p. 157-186.

[40] Kaimal, J.C. and J.J. Finnigan, Atmospheric boundary layer flows: their structure and measurement. Oxford University Press Oxford, 1994. 289pp: p. 289.

[41] Perera, M.D.A.E.S., Shelter behind two-dimensional solid and porous fences, in Journal of Wind Engineering and Industrial Aerodynamics1981. p. 93-104.

[42] Lorenc, A.C., A Global Three-Dimensional Multivariate Statistical Interpolation Scheme, in Monthly Weather Review1981. p. 701-721.

[43] Luo, W., M.C. Taylor, and S.R. Parker, A. comparison of spatial interpolation methods to estimate continuous wind speed surfaces using irregularly distributed data from England and Wales. International Journal of Climatology, 2008. 28: p. 947-959.

[44] Perry, M., and Hollis, D. The development of a new set of long-term climate averages for the UK. International Journal of Climatology 2005, 25(8), p. 1023-1039.

[45] Cellura M, Cirrincione G, Marvuglia A et al (2008a) Wind speed spatial estimation for energy planning in Sicily: Introduction and statistical analysis. *Renewable Energy* **33**:1237-1250

[46] Cellura M, Cirrincione G, Marvuglia A et al (2008b) Wind speed spatial estimation for energy planning in Sicily: A neural kriging application. *Renewable Energy* **33**:1251-1266

[47] Foresti L, Tuia D, Kanevski M, Pozdnoukhov A (2011) Learning wind fields with multiple kernels. *Stochastic Environmental Research and Risk Assessment* **25**(1):51-66

[48] Douak F., Melgani F., and Benoudjit N. Kernel ridge regression with active learning for wind speed prediction. Applied Energy 103 (2013): 328-340.

[49] Veronesi, F., Grassi, S., Raubal, M., & Hurni, L. Statistical Learning Approach for Wind Speed Distribution Mapping: The UK as a Case Study. In AGILE 2015: 165-180. Springer International Publishing.

[50] Willmott, Cort J., and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30.1 (2005): 79.

[51] Fekete, Balázs M., et al. "Uncertainties in precipitation and their impacts on runoff estimates." Journal of Climate 17.2 (2004): 294-304.

[52] Webster, Richard, and Margaret A. Oliver. "Sample adequately to estimate variograms of soil properties." Journal of soil science 43.1 (1992): 177-192.

[53] Met Office (2012) Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current). NCAS British Atmospheric Data Centre: http://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0 - Last accessed: 05.03.2015

[54] Center N.L.P.D.A.A. (2011) ASTER L1B. USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota

[55] Conrad O (2007) SAGA - Entwurf, Funktionsumfang und Anwendung eines Systems für Automatisierte Geowissenschaftliche Analysen. Mathematisch-Naturwissenschaftlichen Fakultäten vol. PhD. University of Göttingen

[56] EEA Corine Land Cover (2006) http://www.eea.europa.eu/publications/COR0-landcover - Last accessed: 05.03.2015

[57] Jenkins GJ, Perry MC, Prior MJ (2008) The climate of the United Kingdom and recent trends. Met Office Hadley Centre, Exeter, UK

[58] Agarwal SK, Kalla SL (1996) A generalized gamma distribution and its application in reliability. Communications in Statistics - Theory and Methods **25**:201-210.

[59] Manwell JF, McGowan JG, Rogers AL (2009) Wind Characteristics and Resources. In: *Wind Energy Explained – Theory, Design and Application*, pages 43-45. 2nd Ed. John Wiley & Sons Ltd

[60] James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*. Springer: New York

[61] Ray ML, Rogers AL, McGowan JG (2006) Analysis of Wind Shear Models and Trends in Different Terrain. In: Proceedings American Wind Energy Association Windpower

[62] Schmidli J, Billings B, Chow FK et al (2010) Intercomparison of Mesoscale Model Simulations of the Daytime Valley Wind System. *Monthly Weather Review* **139**:1389-1409

[63] Rogers AL, Manwell JF, Ellis AF (2005) Wind Shear over Forested Areas. In: Proceedings of the 43rd American Institute of Aeronautics and Astronautics Aerospace, Science Meeting

[64] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[65] Breiman L (2001) Random Forests. *Machine Learning* **45**:5-32

[66] Hansen M, Dubayah R, Defries R (1996) Classification trees: an alternative to traditional land cover classifiers. *International Journal of Remote Sensing* **17**:1075-1081

[67] Grimm R, Behrens T, Märker M et al (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. *Geoderma* **146**:102-113

[68] Wiesmeier M, Barthold F, Blank B et al (2011) Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and soil* **340**:7-24

[69] Cutler DR, Edwards Jr TC, Beard KH et al (2007) Random forests for classification in ecology. Ecology 88:2783-2792

[70] Veronesi, F., & Hurni, L. (2014). Random Forest with semantic tie points for classifying landforms and creating rigorous shaded relief representations. *Geomorphology*, 224, 152-160.

[71] Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. *Pattern Recognition Letters* **27**:294-300

[72] Chan JCW, Paelinckx D (2008) Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment* **112**:2999-3011

[73] Hoaglin DC, Mosteller F, Tukey JW (1983) *Understanding robust and exploratory data analysis*, Vol. 3. Wiley, New York.

[74] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical computing, Vienna, Austria.

[75] Marie Laure Delignette-Muller, Christophe Dutang (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34.

[76] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

[77] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

[78] Pinson P (2006) Estimation of the uncertainty in wind power forecasting. Thesis/Dissertation.

[79] Grassi S, Chokani N, Abhari R (2012) Large scale technical and economic assessment of wind energy potential with a GIS tool: case study Iowa. Energy Policy 45:58-73

[80] Grassi S, Junghans S, Raubal M (2014) Assessment of the wake effect on the energy production of onshore wind farms using GIS. Applied Energy 07/2014; DOI:10.1016/j.apenergy.2014.05.066

List of Captions

Fig. 1 Location of the MIDAS weather stations. In red are indicated the locations of the weather stations we used for our study.

Fig. 2 a) Weibull distributions as a function of *C* (constant *k*, source: [43]); b) Weibull distributions as a function of *k* (constant *C*, source: [43]); c) Example of Weibull distribution (red line) interpolating the histograms (blue bars) describing the wind speed distribution with bins of 1m/s.

This function is generally used in the literature to describe wind distributions [44; 45; 46], and it can be described by only relying on two parameters: shape and scale. Thus, if continuous spatial estimates of k and C are available, the probability distribution of wind speed in any given location can be predicted.

Fig. 3 A) Spatial distribution of the shape factor *C*; B) spatial distribution of the scale factor *k*; C) Uncertainty in terms of AD of the shape factor; D) Uncertainty in terms of AD of the scale factor.

Fig. 4 A) Map of the mean wind speed over the investigated time frame in m/s; B) map of the men wind direction in degrees; C) Wind speed deviation measured as mean absolute deviation (AD); D) Wind direction deviation as AD.

Fig. 5 Map of the reliability of the estimation methods. A) Wind speed standard error in m/s; B) Wind direction uncertainty.

Fig. 6 Site-specific wind speed and distribution uncertainty. By computing the full distributions of wind speed and direction, plus the site-specific uncertainty, we are able to create plots similar to these for each estimated point of the map. These plots represent the frequency distributions of wind speed and direction.