FORECASTING AND SPATIALLY INTERPOLATING WIND SPEED AND DIRECTION ON EARTH, DIMENSION C-137

David Landay,[*] David Hinckley Jr.,[†] and Andy Metcalf[‡]

This work aims to predict the wind speed and direction for a given region using machine learning techniques. This task is divided into three main efforts. First analysis of a region fixed in time, attempting to predict values within that region given a subset of its constituent data. Second, regressing over time to predict the values at a single station in space. Third, attempting classification methodologies on the first of the aforementioned efforts seeing if they fare any better. After an extensive effort to make the data representative and meaningful, the main trend along all of the fronts was that no method achieved better results than simply guessing the mean.

INTRODUCTION

Utilizing NOAA's *Integrated Surface Database* (ISD), an expansive dataset containing hourly meteorological observations spanning from 1901 to present, we plan to increase the spatial resolution of our understanding of the data as well as potentially temporally extrapolate to future readings.

LITERATURE REVIEW

Many papers from the past decade document methods to forecast short-term horizons (day-after or shorter) for wind speeds given temporal and spatial data. The correlation between wind speed and turbine power output makes the study of this type of forecasting an essential asset for wind farm optimization. There are several papers that apply SVM, SVR, and neural networks to approach the issues surrounding wind speed predictions (*Salcedo-Sanz et al.,*[2010],[1] *Mohandes et al.*[2004[2]]), and gain insight into the improvement of wind energy systems. Short-term predictions studies that we looked at utilized both forecast models and surrounding wind speeds as input for their various machine learning algorithms.

With regard to understanding wind speed over a region in space *Bilgili et al.,*[2007][3] apply artificial neural networks to predict wind speeds of target stations, given hourly reported data from known stations across Turkey. Long-term, hourly, data collected by the *Turkish State Meteorological Service* (TSMS), spanning 1992 to 2001 was used in this particular study. Data from 1992 to 1999 was isolated for training the ANN, and data reported from 2000 to 2001 was used to test the accuracy of the model generated by the network. Said network consisted of a single input layer, two hidden layers, and one output layer, where the mean monthly wind speeds of reference stations and corresponding month were used as input values in order to predict the mean wind speeds for a

[*] Complex Systems and Data Science Graduate Student, University of Vermont

[†] Mechanical Engineering Graduate Student, University of Vermont

[‡] Complex Systems and Data Science Graduate Student, University of Vermont

pre-selected target station. A Resilient Propagation (RP) learning algorithm was implemented in the network; a logistic sigmoid function was applied to the hidden layers and a linear transfer function was applied to the output layer as activation functions.

The take away is that the researchers here achieved an accurate prediction of wind speed at target stations by comparing regional mean wind speeds at neighboring reference stations.

DATA

NOAA's Integrated Surface Database (ISD) is a large dataset consisting of hourly meteorological observations from over 35,000 stations worldwide with some stretching back as far as 1901. The ISD merges numerous hourly surface-based datasets from around the world into one comprehensive data source that is updated daily with observations that have been put into a consistent format. The majority of these observations are land based, however a significant number are located at sea. These meteorological observations are primarily derived from automated observing systems, synoptic observations, airports, marine buoys and various other military and civilian stations that include both manual and automated observations.

Although robust, ISD station locations are not evenly distributed across the planet. Figure 1 shows the global distribution of stations that have contributed to the ISD at any time during the ISD's reporting history from 1901 to present. As one might imagine given the economic history of 20th century, stations are most heavily concentrated in North America and Europe; a distribution that has remained consistent throughout the ISDs reporting timespan. Figure 2 shows how the number of stations reporting to the ISD has changed over time. There is a substantial increase in the dataset's volume in both the 1950's and the 1970's. Interestingly, the gap in coverage in the late 60's and early 70's was caused by a disruption due to the transition from manual to automated data entry. At present, the ISD has approximately 11,000 stations contributing hourly data on a daily basis.

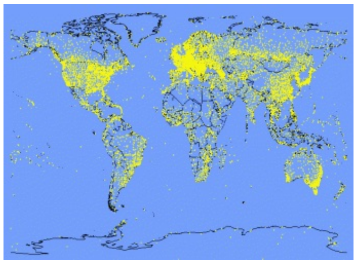

Figure 1. Station Distribution

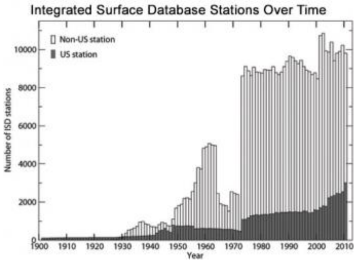

Figure 2. # Stations vs. Time

The ISD reports both daily and hourly summaries on a significant number of meteorological parameters. The most common parameters include wind speed and direction, wind gust, temperature, dew point, cloud data, sea level pressure, altimeter setting, station pressure, visibility, precipitation amounts over various time periods and snow depth. For our purposes, we will extract the features best correlate with wind magnitude and direction.

The ISD implements its own quality control (QC) measures; many times supplementing the station level QC that is already in place. Observation level accuracy is checked by 54 QC algorithms that look for proper data format in each field, extreme values and limits, as well as consistency between parameters. The entire dataset is run through these QC algorithms, with the parameters listed above being the most thoroughly checked.

Approximately 600 gigabytes of uncompressed ISD text files, with data ranging from 1901 to 2017, were uploaded to the Vermont Advanced Computing Core (VACC). This data was then parsed into a readable format by a Java script provided by NOAA and comprises the data that we used for this project.

Data Preprocessing

Although the data appears bountiful, it is not without its issues. While the measurements are supposed to be hourly, there are many instances of missing hours and measurements taken outside of the regular intervals. At this frequency, wind data also has a good amount of natural variation; we must also ask the question, does the wind of a given hour have sufficient physical meaning to warrant the effort of prediction? Another issue that arises from this being a standardized effort across time for stations on a world scale, many potential feature fields are listed but many do not actually contain data. Worse yet, many entries are only partially completed; some stations even change the fields they report moment to moment. The only mass trend is that geographic location of the station is always present and wind speed and direction measurements are mostly present. One last issue of note, when a wind speed of 0 m/s was reported the associated direction was not given a value. This makes sense because if wind is not moving, it has no direction of movement. Unfortunately, a wind speed of 0 m/s isn't technically possible, in much the same way as 10^{-1000} is not 0. It means that a barrier to measurement was not exceeded. This is an issue because each point must have a direction to be of use. There are two main ways to address this; neither of which is without folly. The first is to simply assign it some default value, chosen as North (0°) here. This introduces a bias in the data, North occurring artificially often. The other approach, usable only in classification tasks, is to introduce a new class that symbolizes no direction. This then creates the issue of a class representing, what would be in real-vector space, a zero-measure set. If then one were to classify both wind speed and direction of a station, they could arrive at a speed greater than zero but the "no direction" class. This represents a contradiction that is not easily remedied. Thus, the first option was taken. To account for the wind speed being naturally erratic at the hourly, or sub-hourly, scale a day averaging scheme was used to preprocess the data. Each measurement was considered to represent the station in time up until halfway between its time and the time of the next measurement; the endpoints of a day were considered to be valid for thirty minutes off of their associated end of the day. Our data then represents the bulk trend of a day, a timescale thought to be more physically significant with the added benefit in smoothing rapid transitions in data. This does introduce the possibility for measurements to represent far more time than they physically do, as in the case of a day containing two measurements nine hours apart, but this was accepted since it provides a philosophical consistency to the data. If instead, we only allowed representation windows to extend a fixed amount, we could then encounter cases where the middle of a day was devoid of measurement. Our data would then represent an "observed" day which, technically speaking, is true of all data; data only exists for time observed. Using our scheme of contiguous adjacent time-windows, although dependent on our assumption, does allow us to operate on "days" and not "observed days" which makes the interpretation of the data more physically relevant. Missing features within an entry were handled on a per-feature basis. That is, the day averaging of a feature operated on all data of that feature; a void being covered by the extension of adjacent measurements. This means that a day missing one feature at many different times did not result in that time's value for other features being discarded.

METHODS

Regression Over Space

One meaningful application of this data would be to attempt to predict a station's state in space for a given moment of time. This is done by using data from a specific region on one day to see if primarily topography, along with the other available feature data, can predict a station's wind speed or direction; thus this work will contain regressions for both wind speed and direction over space. This would be useful for applications in wind power; using incomplete knowledge of a region to find the point of highest wind speed so as to maximize power return. Since this analysis is over space, only data from the year 2015 was used; one would expect that years would likely resemble one another, ignoring the potential linear trends one might expect as years progress, thus examination of a single year is sufficient. Owing to the fact that time must be restricted to the scale of a day, 365 independent regressions are necessary, one per day. We then care about which model is best able to generalize across the entire year so metrics are taken as averaged across all of the days. In addition to the wind speed regression, SVR is also tried as a means of prediction since it is better suited to capturing complexity in a dataset via a Gaussian kernel.

Regression Over Time

Data for the regression over time came from a single weather station located in Turku, Finland. This site was chosen because it had data spanning back to 1901 and is still reporting weather observations today. Unfortunately, when selecting this station in particular, the availability of data was only checked in 1901 and 2015 with the assumption that reporting was consistent between these two years. The result of this naive decision left us with a gap in our data spanning from around 1906 to 1952.

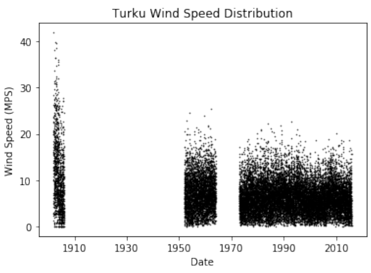

Figure 3. 1

Turku Speed Distribution Over Time

Additionally, as can be seen in Figure 3, there is a inconsistency in the wind speed measurements at the Turku station. We have wind measurements in the pre-1906 data that are significantly higher than those in the post-1952 data. This inconsistency was dealt with differently, depending on the regression technique used in the regressions over time.

Two separate methods were attempted for regressing over time. The first segmented the entire dataset from 1901-2016 by the day of year (1-365), and performed a regression for each day of the year over all years, for a total of 365 regressions. (A slight approximation was made for the leap years as we considered only the first 365 days in any year.) These regressions rely on the implicit assumption that a single day of the year will be correlated with that same day throughout time. Under this methodology we performed separate regressions aimed at predicting both wind speed and direction. The wind speed prediction utilized wind gust, temperature, wind direction and sea level pressure as a feature set, and the wind direction prediction utilized wind gust, temperature, wind speed, and sea level pressure as a feature set. For all regressions under this methodology, we tested both polynomial feature spaces and regularization to see if we could capture existing non-linearity or reduce any overfitting.

The second method for regressing over time did not segment the data by day of year. Under this methodology the month of the year was included in the feature set to verify if it could be used to predict wind speed or direction. Its a reasonable assumption that certain months of the year may on average have higher or lower wind speeds or similar wind directions. Because of the data inconsistency in the pre-1906 observations, post-1952 data was used for this method. This methodology again performed separate regressions aimed at predicting both wind speed and direction. The wind speed prediction utilized wind gust, temperature, wind direction, sea level pressure, and month of year as a feature set, while the wind direction prediction utilized wind gust, temperature, wind speed, sea level pressure and month of year as a feature set. All regressions under this methodology tested both polynomial features spaces and regularization to determine if we could capture existing non-linearity or reduce any overfitting.

Classification Over Space

In addition to forecasting wind speed and direction using regression techniques, we explored how viable a classification method would be for making daily predictions across a region of space. For this study, we looked at a combination of ensemble techniques to create a decision schema for predicting wind speed and direction across multiple stations in Finland. Specifically, we created multiple random forest classifiers to observe whether a prediction could be made with an acceptable amount of certainty for a region on a given day. Determining weather locally makes sense when considering day-to-day activity in that region; ignoring trends that we expect to see over longer periods of time, and under the assumption that data across multiple years will be closely related, we hoped to formulate wind forecasts for individual stations for a moment in time. Hence, for this investigation, we used only one year's worth of wind observations to make predictions. Therefore, each day became a data point and each weather station became sample on that day.

Data from 2015 was selected for this task as it proved to be very data rich; having approximately 57 stations consistently reporting hourly measurements. An unfortunate consequence of this dataset were frequent data imputation irregularities across time. Despite consistency in reported daily observations, as the ISD became more global, methods for recording different features altered or halted, and standard practice varied across stations. For the regression task, missing data posed a smaller threat to the success of the models because the sheer size of the set made up for 'missingness'. But for our classification task, the number of samples were based upon the number of stations that reported measurements for the region on a given day. Therefore, missing data only added to the sparsity of what was available. In addition, a randomly selected subset of those stations was isolated from the rest of the data so as to ensure a consistent testing set for logistic regression tasks, and to test the accuracy of the classifier. Thus, the sparsity of the set was increased.

To build the classifier required a non-bias discretization of the data, with the intention that discrete labels would improve predictive performance. *Lustgarten et al.,*[2008][4] demonstrate in *Improving Classification Performance with Discretization on Biomedical Datasets*, that "discretization can help improve significantly the classification performance of" Support Vector Machines and Random Forests, "as well as algorithms like Niäve Bayes, that are sensitive to the dimensionality of data." By construction, the wind speed measurements were continuous values in units of m/s and the preprocessed direction values were continuous radian angle measurements, denoting the amount by which an anemometer varied in a clockwise direction from true North. Following the clockwise construction of the dataset, directions were assigned a label of North (0), Northeast (1), East (2), Southeast (3), South (4), Southwest (5), West (6), and Northwest (7). Each of the eight values corresponds to one of eight equally spaced intervals in which a reported observation could exist. As mentioned in the *Data Preprocessing* section, calm winds made the data unreliable for an accurate prediction of the wind's direction, and the addition of a ninth direction class would yield a nonsensical result. Similarly, wind speeds were assigned an integer value representation of that speed's class. The Beaufort wind speed scale has evolved into a standard measure of a gale. Integer values between 0 and 12 denote how extreme wind conditions are. Wind speeds in m/s were classified according to the following table:

Speed Class:	Label Description:	Interval:
0	Calm	<0.3 m/s
1	Light Air	0.5 - 1.5
2	Light Breeze	2 - 3
3	Gentle Breeze	3.5 - 5
4	Moderate Breeze	5.5 - 8
5	Fresh Breeze	8.5 - 10.5
6	Strong Breeze	11 - 13
7	Moderate Gale	14 - 16.5
8	Fresh Gale	17 - 20
9	Strong Gale	20.5 - 23.5
10	Whole Gale	24 - 27.5
11	Storm	28 - 31.5
12	Hurricane	> 32 m/s

Table 1. Beaufort Wind Speed Scale

Classification of each station over 365 days meant that Random Forest classifiers for each day were constructed thusly: for each station, each of the direction labels were allowed 2 classes; 1 or 0. For a given observation, if a direction was given a value of 1, then for that day the mean wind direction was whichever direction (0 - 8) the 1 was assigned to. A zero indicated "not [*direction*]." For example, a 0 in the North (0) column would indicate the direction was "not North." Hence, a one-vs-all (OvA) was implemented on each class. This was necessary to remedy biases introduced by the interval construction for wind direction. Each of the eight Random Forests were composed of 50 Decision Trees, and the same eight models were given to each station for a particular day. As each forest classified a specific direction, each decision tree constituted a "vote" for a direction. An artificial probability distribution of the classifications at each station could then be created by applying a Softmax function to the ratio of votes assigned a value of 1, to total votes. The expectation was that the largest ratio would indicate the direction of the wind at an input station on a given day.

RESULTS

Regressions Over Space

Since Linear Regression with a basic (non-polynomial) feature set without regression is a deterministic process, the cost was used as a metric for the purposes of investigating which feature sets should be tested further. A solely topographical regression, Figure 4, revealed that while there are days on which average error is low, the bulk trend is that the regression is far too inaccurate. From there, combinations of topography, SLP, and temperature were tried without polynomial features.

Of that set of tests, SLP with topography had the best average score with regards to wind speed. Figure 5 shows the average errors for that test and while numerically there was improvement, qualitatively the result looks no different. Since the error associated with direction was far greater in proportion, the regression of direction was abandoned.

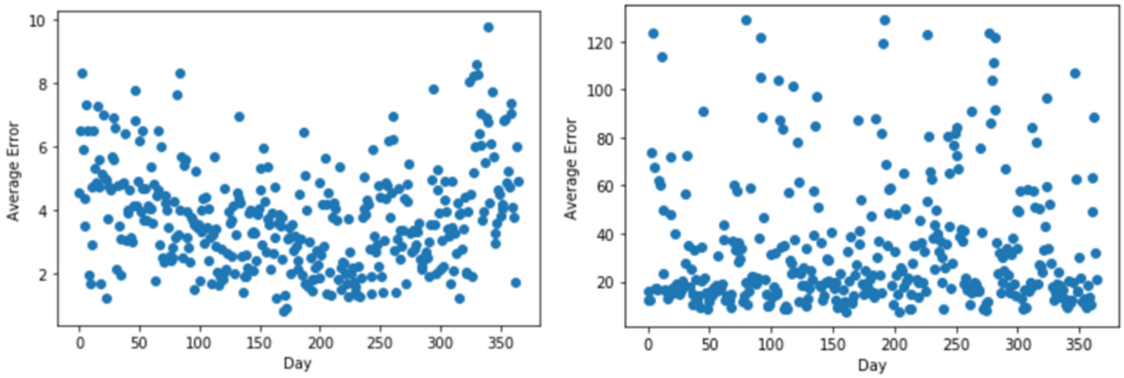

Figure 4. Regression results for (Left) wind speed and (Right) direction with only topographical features

Using this information, a regularized regression of wind speed was performed for second order polynomial features of SLP and topography. Using the validation set to assess the best regularization constant, a simple search found that the error decreased with increasing the regularization. This process plateaued in the limit of increasing regularization which means that the average of wind speed was best among the tested models at generalizing. Figure 6 shows a heavily regularized result. As shown, there appear to be some testing stations for whom the regression often is rather close. It's a bit of an optical trick since if one focuses on finding days where the error was higher for those stations, there are many examples. From this we see that none of the regressions tested give results above guessing the average in regards to being accurate over the entire year; on some days the bulk behavior was rather consistent thus a regression may improve upon average but these days were infrequent as seen looking at the earlier results.

Since SVR with Gaussian kernels is better at capturing complex data, an attempt with SVR was

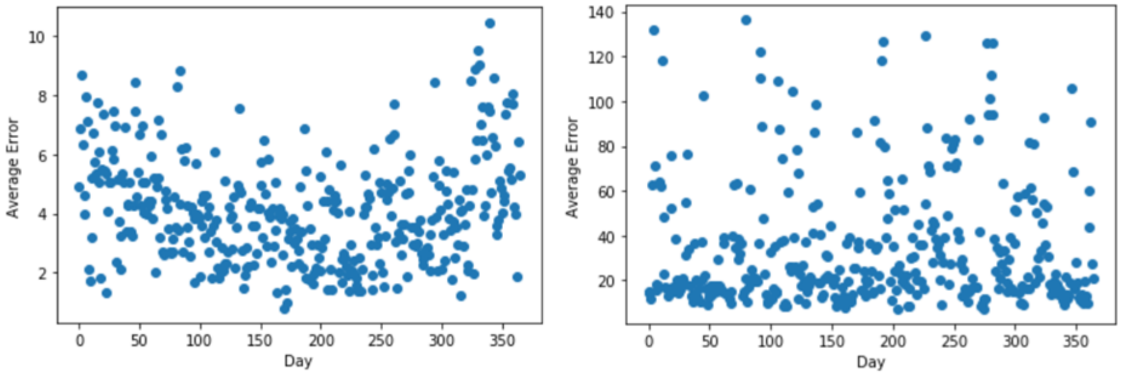

Figure 5. Regression results for (Left) wind speed and (Right) direction with sea-level pressure and topographical features

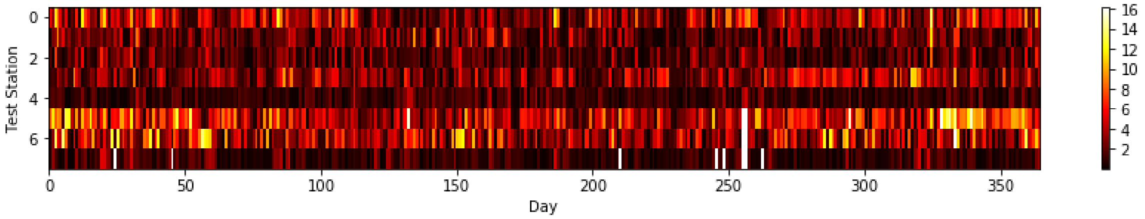

Figure 6. Regularized wind speed regression result

made. The error metric was the R^2 value over the region in question (validation or testing when appropriate). A grid search was conducted for parameters using the validation set which yielded a γ of 1^{-05} and a C of 3.16 but the R^2 value associated with these was -0.12 which means that even these locally optimal values don't yield good results. Applying these values to the testing set, we have Figure 8 which shows the R^2 value for the testing stations for each day of the year. As seen, the upper-bound of performance was the same as what would be achieved by guessing the average value.

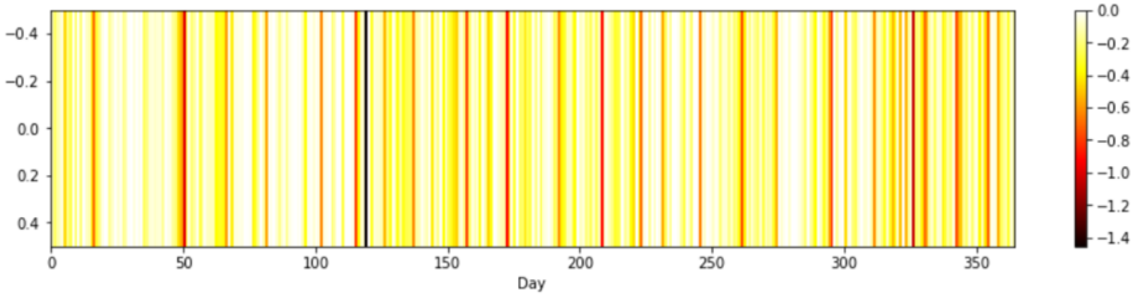

Figure 7. R^2 **SVR wind speed result**

Regression Over Time: Day of Year Over Time Method

For the regression by day of year over time method we began by utilizing the entire feature set to predict wind speed and direction and then tried some individual features as well. Initial regression to predict speed on the entire feature set with no polynomial feature expansion or regularization yielded an average 5-fold cross validation (CV) RMSE across days of the year equal to 6.694. Polynomial feature expansion did not seem to help much, though regularization with alpha = 80 did, yielding a 46% reduction in CV RMSE. However, utilizing the mean of the target variable yielded a 59% reduction in CV RMSE, implying that our model is no better than the mean. We found that increasing the regularization constant alpha always yielded lower CV RMSE. This implies that our coefficients were of no help to the model and ideally the model would converge to a horizontal line corresponding to the y-intercept. This horizontal line is equivalent to the mean of our target variable across years. This method always produced test $R^2 \sim 0$, implying that our mode performs no better than the mean.

Initial regression to predict direction on the entire feature set with no polynomial feature expansion or regularization yielded and average 5-fold CV RMSE across days of the year equal to 202.925. Polynomial feature expansion did not help, however regularization with alpha = 100 yielded a 58% reduction in RMSE. Again we found that further increases to the regularization constant alpha always resulted in lower CV RMSE, which implies that our model coefficients would ideally be driven to zero and the horizontal line corresponding to the y-intercept and the mean of our target variable would be the best predictor. This method always produced test $R^2 \sim 0$ implying that the model performs no better than the mean.

To look at the relative importance of the features used to predict speed under this methodology, we normalized our feature matrix X so that we could compare coefficient values. At alpha = 80, gust was by far the most important feature based on the absolute value of its coefficient value. Because of its relative importance we performed a regression to predict speed using gust as the sole feature. This regression actually performed slightly better than the mean, however the test R^2 statistic was still close to zero, implying that the model was not much better than the mean.

To look at the relative importance of the features use to predict direction under this methodology, we again normalized the feature matrix X so that we could compare coefficient values. At alpha = 100, not all coefficients had been driven to zero so we performed individual regressions for every feature. For all of these regressions, we found that increasing alpha always yielded lower CV RMSE, implying again that the mean was a better predictor than any model using the individual features.

Regression Over Time: Total Data Over Time Method

For the regression method that did not partition the data by day we again began by utilizing the entire feature set to predict the wind speed and direction and then looked at individual feature contribution as well. Initial regression to predict speed on the entire feature set with no polynomial feature expansion or regularization yielded a 5-fold CV RMSE of 2.55. Utilizing polynomial feature expansion, we gained a 3% improvement over the initial CV RMSE. Further application of a regularization constant with alpha = 800, yielded a 3.7% improvement over the initial CV RMSE. However, when we use the mean of the target variable we are also able to achieve 3.7% improvement over the initial CV RMSE, implying that our model performs no better than the mean. Interestingly, the test R^2 statistic for this model was equal to 0.44, which seems to imply that our model does explain some of the variation in speed. Analysis of the individual features yielded test R^2 values close to zero for temperature, wind direction, sea level pressure, and month but found a R^2 value of 0.42 for gust. This may explain the test R^2 statistic equal to 0.44 when we look at all of the features. Again, we find that gust is the only significant feature to predict wind speed, which clearly makes sense since the two are intrinsically related.

Initial regression to predict direction under this methodology yielded an 5-fold CV RMSE of 82.45. We were unable to significantly improve this result by using polynomial feature expansion or regularization. Again we found that the mean of the target yielded similar results implying that our model is no better than the mean. The test $R^2 = 0.03$ for this model, implying that the model does not do a good job explaining the variation in wind direction

Classification Over Space

The above table shows the result of applying the Softmax function to each classifier after fitting the models to the data. The first eight stations at day 1 are revealed. Below that is the overall distribution of wind direction across the region for the first day. Our expectation was that the classifiers would reveal the same distribution in the predicted outcomes. The findings in Table 2 likely reveal that for a given station, if the field denoting a direction is a column of all zeros, then the Random

Latitude	Longitude	North	Northeast	East	Southeast	South	Southwest	West	Northwest
61.200	28.467	0.1355	0.1355	0.1355	0.1355	0.1355	0.0540	0.1328	0.1355
60.383	22.110	0.1368	0.1368	0.1368	0.1368	0.1263	0.0627	0.1263	0.1368
61.943	28.945	0.1377	0.1377	0.1377	0.1377	0.1377	0.1127	0.0606	0.1377
60.717	28.733	0.1377	0.1377	0.1377	0.1377	0.1377	0.0582	0.1150	0.1377
61.000	28.567	0.1380	0.1380	0.1380	0.1380	0.1380	0.0608	0.1107	0.1380
61.033	21.783	0.1363	0.1363	0.1363	0.1363	0.1363	0.0511	0.1309	0.1363
60.250	20.750	0.1371	0.1371	0.1371	0.1371	0.1344	0.0604	0.1192	0.1371
61.050	28.217	0.1379	0.1379	0.1379	0.1379	0.1379	0.0572	0.1152	0.1379
*	*	*	*	*	*	*	*	*	*

Table 2. Day 1 Predictions for Wind Direction (first eight)

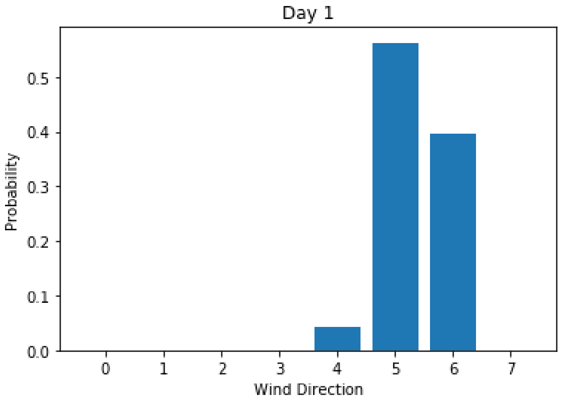

Figure 8. Distribution of Wind Direction over Region

Forest classifier will experience fewer instances of that class. Since the model is predicting based on the mode of the two classes for each direction, then a column with no 1's will have a larger bearing of the overall prediction. Thus, we have little certainty that classification of wind speed and direction performed in this way can adequately predict wind speed and direction for a region of space.

It is possible that a denser area of stations would reveal a higher degree of variance among the stations. The higher variability would change the resolution of the distribution for a region of space, and thereby provide a better depiction of the wind's activity. In addition, we could draw a larger spatial area of interest for our study, but the test would then become less meaningful because we lose significant resolution in our predictions. It is useless to know the overall direction of the wind for an entire country, when regional approximations tell us where to place turbines and other such instruments. The results of classification may simply reveal that the dataset or our methodology are poor forecasters.

CONCLUSION & FUTURE WORK

In summary, the attempted regressions in both space and time showed that predicting the average out-performed all of the models tested. The attempt at classification methods too showed that the machine learning techniques here employed were not able to achieve that for which we hoped. Unfortunately, a result of failure is unable to prove anything definitive. All that can be said from these results is that the methods here attempted on the problem phrased in the way it was here do not achieve good results. This could be an issue with data; we could be lacking in sufficient data or the appropriate features for the task. There is also the possibility that the problem as was formulated here has issues, such as our choice of methods simply being inappropriate for solving the problem and a less rigid structure, such as Artificial Neural Networks, may be better suited for the task. Lastly, it is entirely possible that predicting wind is not possible with machine learning and that the best one could hope for is predicting the average; while this would be disheartening, the negative results here only allow us to speculate and not disqualify even the most unlikely of possibilities.

Any continuation of this work must consider the necessity of spatially restricting the region upon the basis of sensible spatial locality since it invariably shrinks the usable data for machine learning tasks. While it makes rational sense that one would not ask what the weather was in Turkmenistan to inform them of conditions in Burlington Vermont, there must be some middle ground that determines that allowable extent of a spatial region. Attempting these methods on a region with a greater density of measurement stations would help alleviate data issues. Also, as mentioned above, using a different methods such as Artificial Neural Networks, due to its greater variation in available problem phrasings, may improve performance on these tasks.

CONTRIBUTIONS

Previous research that we looked at attempted to predict wind speed at target stations in a region by using wind speed data from various adjacent stations.[3] Our spatial regression method attempts a similar goal, however using raw meteorological data from surrounding stations that does not include wind speed. The finding that these meteorological features did not work well to predict wind speed is significant in that it may imply that non-speed features should not be used to predict target wind speeds. Before making this conclusion however, we would need to test many more regions with varying topography and sizes.

Given the volatile nature of wind, it is a challenge to predict the amount of power supplied to the grid at any one time. For this reason, many studies have focused on short term wind prediction to help utilities gain a better understanding of future power demand. These studies utilize both forecast models and surrounding wind speed measurements to make their predictions. We take a different tack in that we use only raw meteorological measurements that span back a significant amount of time (much farther than any of the studies we looked at), in an attempt to build a general model to predict wind speed that would be valid for any moment in time. Again, the finding that these raw meteorological measurements did not work well to predict wind speed is significant in that it may imply that these features are not amenable to wind speed prediction.

ROLE FOR EACH MEMBER

- David Hinckley: Regression/SVR over space
- David Landay: Classification Methods
- Andy: Regression over time

REFERENCES

[1] Sancho Salcedo-Sanz, Emilio G. Ortiz-Garcia, Angel M. Perez-Bellido, Antonio Portilla-Figueras, Luis Prieto. *Short term wind speed predictions based on evolutionary support vector regression algorithms*. Expert Systems with Applications, **38**,issue 4, 4052-4057, (2011).

[2] M.A. Mohandes, T.O. Halawani, S. Rehman, Ahmed A. Hussain. *Support vector machines for wind speed prediction* Renewable Energy, **29**, issue 6, 939-947 (2004).

[3] Mehmet Bilgili, Besir Sahin, Abjulkadir Yasar. *Application of artificial neural networks for the wind speed prediction of target station using reference stations data* Renewable Energy, **32**, issue 14, 2350-2360 (2007).

[4] Lustgarten JL, Gopalakrishnan V, Grover H, Visweswaran S. *Improving Classification Performance with Discretization on Biomedical Datasets* AMIA Annual Symposium Proceedings, **2008**;445-449 (2008).

[5] Chung-Hong Lee, Chien-Cheng Chou, Xiang-Hong Chung, Pei-Wen Zeng. *Applying Climate Big-Data to Analysis of the Correlation between Regional Wind Speed and Wind Energy Generation*. 3rd International Conference of Green Technology and Sustainable Development (2016)