Applying Weibull Distribution and Discriminant Function Techniques to Predict Damaged Cup Anemometers in the 2011 PHM Competition

Joshua Cassity[1], Christopher Aven[1], Danny Parker[1]

[1]*Miltec Research and Technology, A Ducommun Company, Oxford, MS, 38655, United States*

dparker@one.ducommun.com
jcassity@one.ducommun.com
caven@one.ducommun.com

ABSTRACT

Cup anemometers are frequently employed in the wind power industry for wind resource assessment at prospective wind farm sites. In this paper, we demonstrate a method for identifying faulty three cup anemometers. This method is applicable to cases where data is available from two or more anemometers at equal height and cases where data is available from anemometers at different heights. It is based on examining the Weibull parameters of the distribution generated from the difference between the anemometer's reported measurements and utilizing a discriminant function technique to separate out the data corresponding to bad cup anemometers. For anemometers at different heights, only data from the same height pair combinations are compared. In addition, various preprocessing techniques are discussed to improve performance of the algorithm. These include removing data that corresponds to poor wind directions for comparing the anemometers and removing data that corresponds to frozen anemometers. These methods are employed on the data from the PHM 2011 Data Competition with results presented.

1. INTRODUCTION

The issue of identifying faulty anemometers used during wind resource assessment at prospective wind farm sites has increased in recent years as wind energy grows in importance due to declining fossil fuel availability. In wind resource assessment, the need for effective wind estimation is critical. If anemometer readings differ from reality by a small amount, the cost in terms of return on investment can be large.

When data from anemometers at equal height above the ground (and therefore equal wind speeds in principle) are available, previous studies (Ye, Veeramachaneni, Yan., and Osadciw, 2009) have shown that the differences in wind speed between sensors at equal height can be characterized as a Weibull distribution. To identify broken anemometers the Weibull parameters of difference between wind speeds were estimated and then compared against thresholds for the shape and scale parameters. These thresholds where heuristically determined based on experience with previously good data. The researchers go on further to propose using the area under the cumulative distribution function (cdf) as a feature for discrimination. Though the method has demonstrated good results, there still remains the issue of analytically choosing an appropriate threshold. In this paper we utilize a discriminant analysis to generate the minimum-error-rate thresholds for distinguishing the bad sensor data sets from the good sensor data sets. The features we chose to discriminate on are the Weibull parameters characterizing the difference in the wind speeds between two anemometers. In cases where there were not anemometers paired at the same height a physics based model of the wind speed versus height was assumed to feed into the discriminant functions.

Discriminant analysis is a powerful set of tools that tries to analytically determine classification boundaries based on the statistical behavior of the features used to characterize the data. These boundaries can be further adjusted based on the probability of each of the classes occurring. Before the discriminant functions can be generated a certain amount of preprocessing must be done to condition the data. This is done to remove certain environmental and terrain effects that may skew the thresholds.

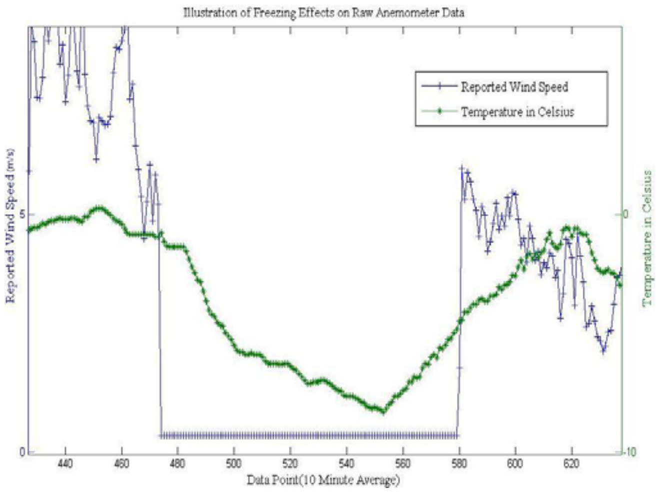

Figure 1-Effects of Freezing Conditions on Raw Anemometer Data

The data from the PHM 2011 Data Competition was organized as follows. It was divided into two groups: 'paired' anemometer data containing two sensors at equal height and 'shear' anemometer data containing either three or four sensors at different heights above the ground. The wind speed measurements for each sensor were averaged over ten minutes and provided in the data, along with maximum, minimum, and standard deviation within each averaged ten minute segment. In addition, wind direction and temperature data were provided alongside the anemometer readings.

The goal of the PHM 2011 Data Competition was to determine, from provided data, whether given cup anemometers were damaged and reporting erroneous readings. Per competition rules, anemometers that become frozen due to weather effects do not count as damaged. These had to be identified in order to prevent false diagnoses. There are 420 test data files for the 'paired' case and 255 test data files for the 'shear' case. In the 'paired' case, a point is awarded when a data file is properly diagnosed, that is, when both sensors are correctly marked as damaged or undamaged. In the 'shear' case, a point was awarded if the competitor correctly determined that a sensor (if any) is damaged in the data file. In addition to the test data files, there were 12 training data files for both the 'paired' and 7 training data files for the 'shear' case. For these files, the anemometers guaranteed to be good, that is, not broken. So it was not necessary in the 'shear' case to identify the specific damaged anemometer. In the 2011 PHM Competition, submissions were graded and a leader board was provided to show the rankings of each team relative to one another. However, the actual scores for each submission were obscured and the labeled data was not released, making it impossible to accurately compare one algorithm's performance to another.

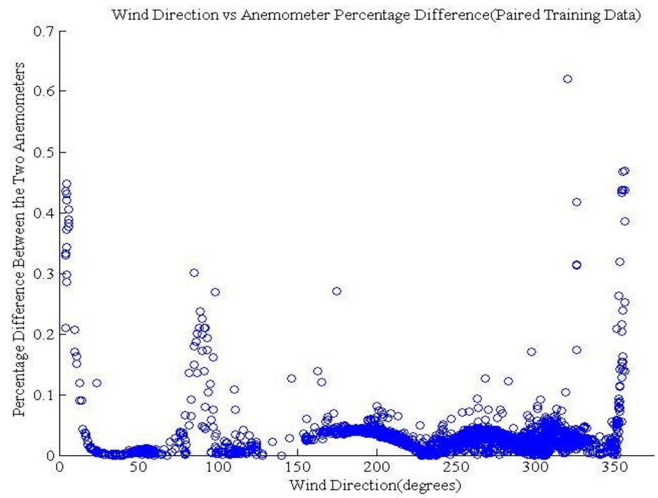

Figure 2-Wind Direction versus Reported Difference between Sensors at Equal Heights

The rest of the paper is organized as follows. In 'Methodology' the data preprocessing, discriminant function technique, 'initial guess' estimation for paired data corresponding to anemometers at equal height, and 'initial guess' estimation for shear data corresponding to anemometers at different heights are presented, and in 'Conclusion' we mention some concluding remarks and point towards areas of future research. Throughout the paper, we will refer to data that corresponds to undamaged sensors as 'good data' and data from damaged sensors as 'bad data'.

2. METHODOLOGY

2.1. Data Preprocessing

There were several problems with the raw data used that had to be addressed before methods could be applied to determine faulty anemometers. At times, the anemometers would freeze and stop moving, even if they were not broken, and this would skew the algorithm towards overpredicting bad sensors. The method used to accommodate this was to search the data for measurements that were both below freezing and stuck on the lowest possible wind speed reported by the anemometer. This data was then discarded under the assumption that the sensor was frozen and reporting incorrectly.

For the 'shear' data set, if even one of the anemometers was seemingly frozen, all the data for that unit of time was discarded. Figure 1 gives an example of frozen data within the data set. One can clearly see where the reported wind speed drops to near zero which corresponds to temperatures below freezing.

A second issue that turned up was that, due to the anemometer placement (whether at 90 degrees to each other or 180 degrees), the observed mean difference between wind speed for anemometers at equal heights (such as found in the 'paired' data) would vary according to wind direction. This caused the differences to be more spread out because of the statistics in a few particular directions. To deal with this problem, each 'paired' data file was divided into bins 30 degrees in size. Then the standard deviation of the percentage differences between the two anemometers at each ten minute average was calculated for every bin and the two bins with the highest wind speed difference between the two anemometers were discarded. This threw out wind directions that corresponded to the greatest difference in wind speeds between anemometers at equal height, thus causing the remaining data to have a much lower standard deviation. In Figure 2, we show the variation between reported differences in wind speed versus wind direction for a training data set. In this figure data from 0 and 90 degrees would be discarded. If there were more than two directions that skewed the statistics they could also be discarded.

2.2. Discriminant Function Technique

The Discriminant analysis is a method of pattern classification that is known to achieve the minimum-error-rate classification for a given feature set (Duda, R. O, Hart, P. E., Stork, D. G., 2001). For each classification, a separate discriminant function is derived and evaluated for each feature vector. Then a particular set of data is determined to be of the class with the highest discriminant function value. This in effect yields the minimum error rate classification in the assigning of a class. For this application, there are two classes: the good sensors and the bad sensors and the features used to describe the class are the shape and scale parameters of the Weibull distribution of the difference in wind speeds. The formula for computing a discriminant function g($\mathbf{x}$) for arbitrary covariance matrix is given by,

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{W}_i \boldsymbol{x} + \boldsymbol{w}_i^t \boldsymbol{x} + w_{i0} \quad (1)$$

Where,

$$\boldsymbol{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \quad (2)$$

$$\boldsymbol{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \quad (3)$$

And,

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2} ln|\boldsymbol{\Sigma}_i| + ln\, P(\omega_i) \quad (4)$$

Here $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ refer to the covariance matrix and mean vectors of the features, respectively. The mean vector can be understood as the center of a class and the covariance matrix describes how scattered the points in the class are distributed about the center. $P(\omega_i)$ refers to the a priori probability that a given test file belongs to the class which the discriminant function corresponds to. In equation 1, $\boldsymbol{x}$ refers to the feature vector for a particular set of data under consideration, that is, the Weibull parameters generated from the percentage difference between the two anemometers. This number could typically be obtained from the failure rate of the sensors. Since that was not known it was estimated by examining the percentage of the files that the initial estimation classified as bad. To apply the Discriminant Function Technique, it is necessary to have labeled or known data from both classes. Since there was no labeled bad data provided, a method had to be devised to get an initial guess at the 'bad sensor' class. Exactly how the initial classifications were determined for both the paired and shear cases is described in the next section. If more detailed information on discriminant function analysis is desired (Wolverton, C., Wagner, T., 1969) is an excellent resource.

2.3. Paired Data

Previous research (Ye et al, 2009) has demonstrated that anemometers at similar heights exhibit a Weibull distribution in the time domain differences between their mean reported wind speeds. In their paper, the shape and scale parameters of the Weibull distribution from a week's worth of data is estimated using a maximum likelihood estimation and results over a number of weeks are plotted on a graph with the scale parameter on the x-axis and shape parameter on the y-axis. Then, a visual investigation is performed and a 'cloud' of good performing sensors is identified by drawing an oval around the area of highest density, leaving the points outside the oval to be flagged as bad. This analysis was applied to obtain the initial bad set of files corresponding to good sensors. In our case, since the data in the PHM 2011 Data Competition was provided in sets of five days, we used five days as the interval of time for which to calculate scale and shape parameters. From this set of data, we employed two methods to label files as either good sensor or bad sensors for the discriminant analysis. Secondly, a hypothesis test was used to determine if a set of data came from a Weibull distribution from the good sensor class. The Kolmogorov–Smirnov (K-S) test (Eadie, W. T., Drijard, D., James, F. E, Roos, M., & Sadoulet, B., 1971) was implemented as the hypothesis test with the null hypothesis being that the data under consideration was from the same distribution as a good pair of sensors.

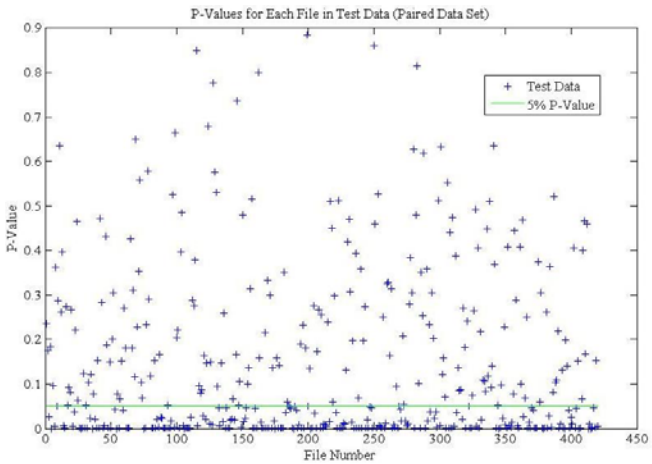

Figure 3-Confidence Level for each Test File in rejecting the null hypothesis

To accomplish this, the percentage difference between the anemometers was calculated for every 10 minute average in each of the training files (which were known to be good data). Each training file formed a distribution of percentage differences. Then, for each of the test files, the same process was undertaken and the K-S test was performed for the test file against each one of the 'good' distributions. If the null hypothesis, that the test file and training file were from the same distribution, was rejected for all 'good' distributions with 5% significance level, then the file was flagged as bad. On the other hand, if the test file's distribution could be matched with at least one training file at 5% significance level, then the test file was marked good. In order to be assured that the initial data in the bad class was truly from bad cases the intersection of the K-S set and the set derived from visual inspection of the data (that is, the files common to both sets) was used to obtain the set of bad sensors for applying the discriminant analysis.

Figure 3 shows the confidence level of rejecting the null hypothesis for each test file, given the set of training data files. As can be seen, this method is another form of thresholding where the threshold in this case is the significance level. A significance level below 5% indicates that there is a less than five percent probability that the null hypothesis is correct.

Figure 4 shows the plot of the Weibull parameters for the paired data set, both training and test data. Note the ‘cloud’ of data points in the lower left corner clustered together corresponding to the estimated 'good' data files. So in general ‘good’ data files have lower scale parameters than ‘bad’ data files. A single training data point seems disconnected from the rest of the training point, with a much higher than expected scale parameter. This point (circled in the figure) was discarded when choosing our threshold, as it corresponded to a data file where a large amount of points had been thrown out due to the wind direction preprocessing described earlier.

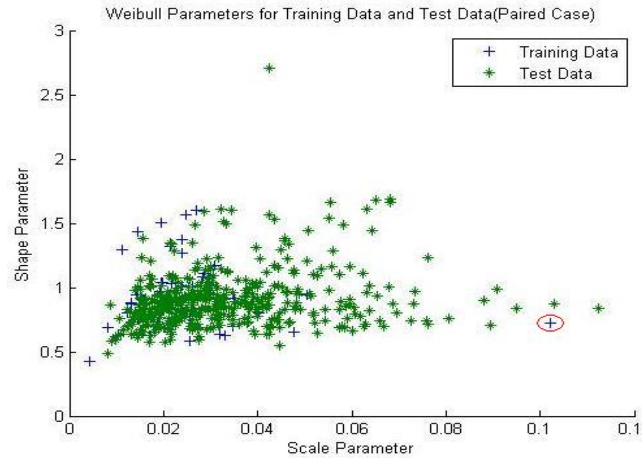

Figure 4-Scatter Plot of Weibull Parameters for all Training and Test Files in Paired Data

For the paired data, a competition requirement was to identify not only when a sensor was bad but which one of the pair was defective. In order to identify which specific anemometer has failed, the assumption was made that the anemometer with the lower mean reported wind speed will be the one that is bad. This was based on a consideration of the types of damage possible to cup anemometers. For example, a chipped or cracked cup will not hold wind as well as a normal one and as such should report a lower wind speed.

2.4. Shear Data

In the shear data, if there were sensors paired at each height then the previous methods for getting initial labeled data would be effective. However, that was not the case here. Data were given over various days with only one sensor at each height and to make matters more difficult, the heights were not consistent from file to file.

So to generate some initial labels for both good and bad classes, a physics-based model was employed. There are two different models that can be used to characterize the relationship between wind speed and height. The first, referred to as the wind profile power law (Oke, T, 1987). It is of the form,

$$\frac{u}{u_r} = \left(\frac{z}{z_r}\right)^{\alpha} \qquad (5)$$

Where u and z is the mean wind speed and height, respectively under consideration, u_r and z_r refer to the wind speed and height at a given reference point (usually 10 meters), and α is an empirically derived constant whose value depends on the stability of the atmosphere. For conditions of neutral stability, α is approximately 0.143. This equation assumes the relationship follows a simple power relation (Touma, J, S, 1977).

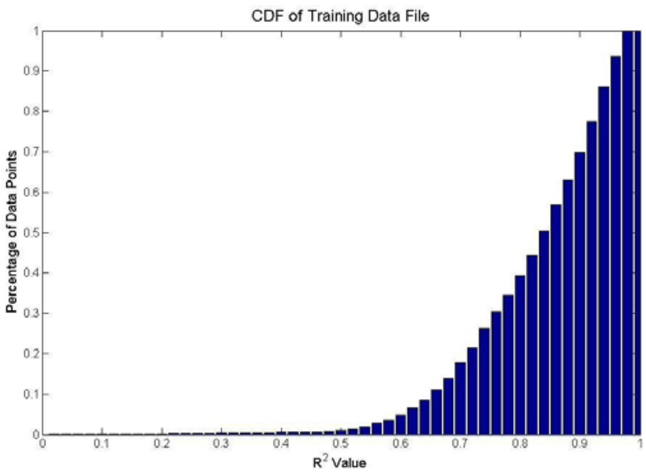

Figure 5-CDF of R^2 for a training file in the Shear data set

This equation does not take into account certain terrain features such as the roughness of the surface or the level of atmospheric stability, which can greatly affect the reported wind speeds. It only requires the mean wind speed at a 'reference' point, usually 10 meters. The second model, referred to as the 'logarithmic wind profile law' (Oke, 1987), is shown in equation 6,

$$u_z = \frac{u_*}{\kappa}\left[ln\left(\frac{z-d}{z_0}\right) + \psi(z, z_0, L)\right] \tag{6}$$

Equation 6 is valid from the surface up to around 1000m. It takes stability and surface roughness into account, but requires a number of known parameters, such as the zero plane displacement, friction velocity, and the Monin-Obukhov stability parameter, none of which were available to us.

Since the physical parameters were not known, a general *log*-linear relationship of the wind speed versus height from the ground was assumed. With the assumption that the data should fit a logarithmic curve, the exponential of the data was taken. The results of which should produce wind speeds that are a linear function of height. Once this conversion has been made, a simple linear regression analysis was performed to determine residuals and goodness of fit. The result of the linear fit analysis, R^2, gives a quantifiable value that can be used to classify on a sample by sample basis. However, the problem still remained on what an appropriate R^2 threshold should be to label a file as bad. To attempt at a systematic way to arrive at a threshold the R^2 values were then used to make histograms for estimating the probability density function (PDF), with careful consideration given to keeping bin edges equal. The PDF's were then integrated to form the cumulative density function (CDF). The CDFs represent the percentage of files equal to or below a particular R^2 value.

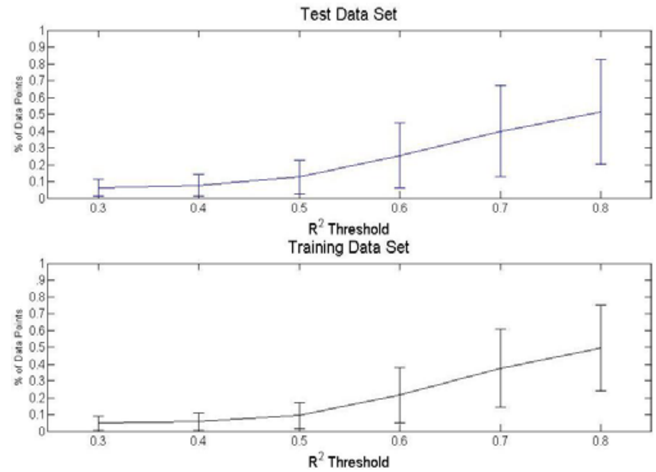

Figure 6-Mean and Standard Deviation for each R^2 value analyzed in Shear data

Figure 5 shows a CDF plot for a typical training data file. Notice that there are very few files that don't have at least a R^2 value above 0.7. This is in line with expectations that a *log*-linear fit does a good job modeling the wind profile as a function of height.

To arrive at an exact threshold value the CDF's were then analyzed at six values of R^2, ranging from 0.3 to 0.8, to determine what percentage of data points per file were less than or equal to the R^2 values of interest. The idea being that the training data would have less percentage of data points below a particular R^2 value when compared to test data that has potential bad sensors in it. At each threshold the mean and standard deviation of the number of files labeled as bad were then calculated for both training and test data. Figure 6 shows the plots of mean and standard deviation for each R^2 value analyzed.

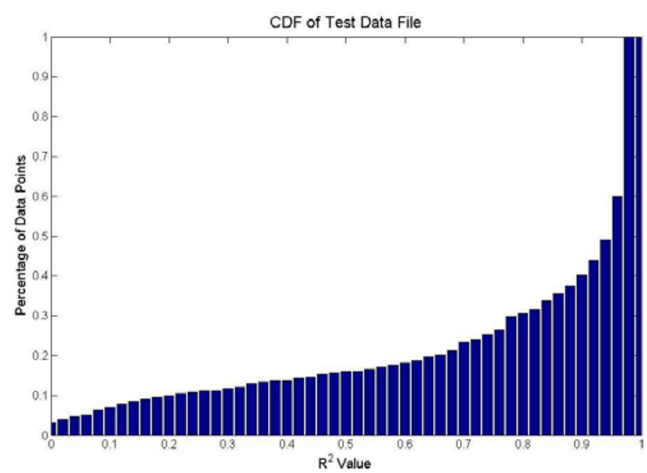

Figure 7-CDF of R^2 for a test file in the Shear data set

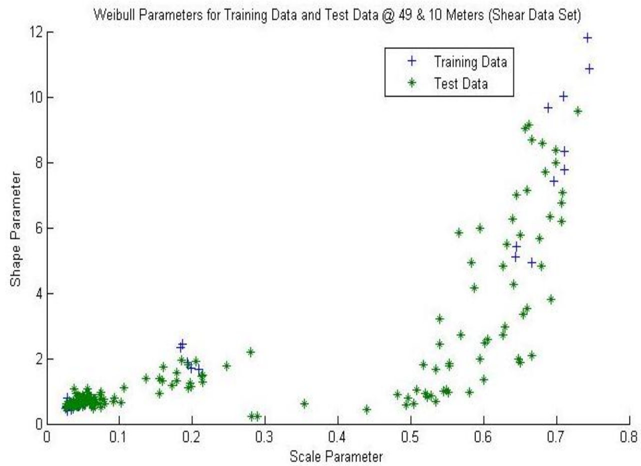

Figure 8-Scatter Plot of Weibull Parameters for all Training and Test Files in Shear Data (49 & 10 Meter Case)

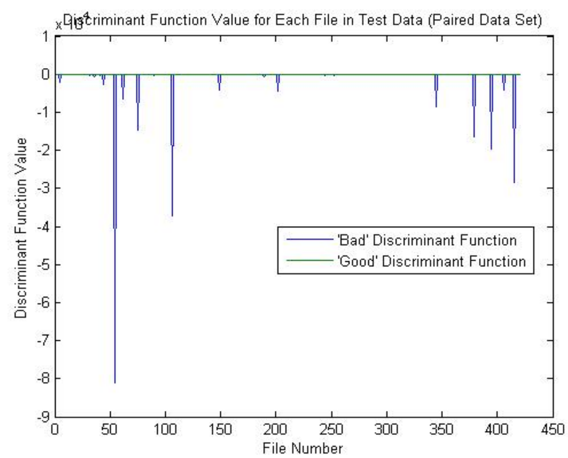

Figure 9-Discriminant value comparison for each test file in Paired Data Set

The R^2 value with the largest discrepancy between training and test data was then selected as the threshold value for labeling a file as bad. The assumption was made that the larger difference between the two data sets at a given R^2 value would indicate that the optimal threshold for discriminating between the two classes was achieved.

Since this threshold is on a per sample basis and the desire of the competition was to label the data per day, it had to be worked out how many files flagged during a day would cause the data to be labeled as bad. To come up with this threshold the per sample R^2 threshold was analyzed to determine the max percentage of training data points that were included in all bins less than or equal to the chosen threshold. Since it was known that all training data was considered good then it was assumed that whatever percentage of training files labeled as bad represented some acceptable percentage per day. Any days that had a higher percentage would then be labeled as bad. In practice a value slightly higher than this max value was then chosen as the percentage threshold for damage detection. This slight increase in percentage threshold value was selected to ensure that no training data was flagged as bad. The CDFs generated from every test file were then evaluated at the R^2 per sample threshold. Figure 7 shows the CDF of a typical test data file. The percentage of data points contained in all bins less than or equal to the per sample threshold was then compared to the percentage threshold obtained from the training data. If the percentage of test data points contained in those bins were greater than or equal than that threshold, the file was marked as bad.

Using this line of reasoning to label some files as the bad sensors class, we are able to extend the previously discussed discriminant analysis to the shear data.

However there is still some difficulty in using this approach for sensors at different heights. The problem is that the percentage difference between anemometers will increase or decrease with different height differences. To help alleviate this issue, all the training data was divided into height-pair combinations (such as 49 & 10 meters, 35 & 10 meters, etc), Weibull parameters were generated for the percentage difference between the respective sensors, and the discriminant function method was employed on these parameters using the files labeled earlier as the 'bad set'. This allowed us to compare each sensor to every other sensor in a data set. If a sensor goes bad, ostensibly this will show up in the Weibull parameters generated by the comparison of that sensor with all other sensor in the data set.

2.5. Discriminant Analysis Summary

In this section, the discriminant analysis employed in this paper will be briefly summarized. The analysis relies on creating discriminant functions for each desired classification (‘good’ and ‘bad’). The equations to create this function are given in equations (1)-(4). The necessary parameters are the feature vector, in this case, the scale and shape parameters derived from the Weibull distribution of the percentage difference between two sensors under test. The resulting feature vector is of dimension 2 x 1. By examining all the feature vectors for both groups of labeled data yields two mean vector of size 2 x 1 and two covariance matrices of size 2 x 2, one for each class. In order to obtain the data necessary for each discriminant function, the ‘training data’ provided by the competition was used as good data and a variety of methods outlined in the previous two sections were used to obtain an estimate of some initial good data. Once the feature vectors, mean vectors, and covariance matrices were obtained for both classes, the parameters of a Weibull distribution describing the difference between mean wind speeds of sensors at different heights were estimated. The resulting shape and scale parameters were used as the vector ***x*** in both discriminant functions, good sensors and bad sensors. The file was labeled as either good or bad based on which discriminant function produces the highest value. This process was repeated for every file under consideration.

After the discriminant functions have been created, they are evaluated at each file and the file is classified into the group whose discriminant function has the higher reported value. In our analysis, using the 'initial guess' presented in the previous two sections, we flagged 25% of the files as bad from the 'paired' data set and 58% of the files from the 'shear' data set. Figure 9 illustrates the outputs of the discriminant functions for each test file in the paired data. The function with the higher value will correspond to the minimum error rate classification of incorrect classification.

3. CONCLUSION

Building on the previous work of others in identifying that a Weibull distribution can statistically describe the differences between paired anemometers over short distances, we have proposed a conceptually simple method using discriminant functions for analytically determining classification thresholds. There are several complicating parameters like not having paired data at all heights and consistent heights, that are most likely artifacts of the competition and not indicative of real world monitoring. In addition, a real world application would generally also provide environmental information such as stability & surface roughness, along with the Monin-Obukhov stability parameter, which would enable the more accurate 'logarithmic wind profile law to be used. The performance of the data was improved by preprocessing to remove obviously faulty data and there was a rough attempt at estimating the probability of a sensor being bad. There are several ways in which this method could be improved in the future. One improvement could be had if the statistics between all sensors on a tower were modeled and used as features. This would allow for a more complete and robust description of an installation which in turn would allow for more powerful classification techniques to be applied.

Though the discriminant function yields the minimum-error-rate classification, it is highly sensitive to variations in the intial guess. We noticed that varying the files in the initial guess can dramatically alter the classifications. Therefore, finding the best method to obtain the initial guess of the bad files is critical.

Also, a better method for accommodating the variation in wind speed differences between anemometers at the same height with wind direction would offer some improvement. A simple method of doing so would be to find a function that characterizes the plot of wind speed differences versus wind direction as presented earlier and subtract the effect from the data. To accomplish this, it would be necessary to know the precise orientation of the anemometers ahead of time as this is not always possible to deduce from the plot of the data (sometimes wind may be from a small number of directions for the duration of a test and plot such as Figure 2 cannot be easily made). Another possible improvement would be if additional data such as surface roughness, and atmospheric stability information is available, then the log wind profile equation can be used in which should greatly improve the predictive ability for the shear data set.

REFERENCES

Duda, R. O, Hart, P. E., Stork, D. G., (2001). *Pattern Classification*. New York, NY: John Wiley & Sons, Inc

Eadie, W. T., Drijard, D., James, F. E, Roos, M., & Sadoulet, B. (1971). *Statistical Methods in Experimental Physics*. Amsterdam, North-Holland: Elsevier Science Ltd

Oke, T, R, (1987). *Boundary Layer Climates*, London, UK: Routledge

Touma, J, S, (1977). Dependence of the wind profile power law on stability for various locations, *J. Air Pollution Control Association*, Vol. 22, pp. 863-866.

Wolverton, C., Wagner, T. (1969), Asymptotically optimal discriminant functions for pattern classifiers, *IEEE Transactions on Information Theory*,

Ye, X., Veeramachaneni, K., Yan, Y., & Osadciw, L. A. (2009), Unsupervised Learning and Fusion for Failure Detection in Wind Turbines, *Proceedings of 12th International Conference on Information Fusion*. July 6-9,Seattle,WA