Comparison of novel SCADA Data Cleaning Technique for Wind Turbine Electric Pitch System

C McKinnon [1], K Tartt [1], J Carroll [1], A McDonald [2], C Plumley [3] and D Ferguson [4]

University of Strathclyde, 99 George Street, Glasgow, G1 1RD

E-mail: conor.mckinnon@strath.ac.uk

Abstract. Wind turbines typically do not operate in the ideal operating conditions, leading to abnormal behaviour that is reflected in their power curves. This abnormal behaviour can affect the performance of condition monitoring processes, as it may mask faulty behaviour. By cleaning other abnormal data, such as curtailment, models can learn the normal behaviour of the turbines.

This paper presents a novel cleaning technique that utilises a combination of data binning and the Mahalanobis distance. This removes between 5 to 6% of the data, without great loss of normal data. When compared against other data cleaning techniques, the one presented in this paper produces a more ideal power curve. This technique could improve the performance of data-based condition monitoring techniques.

1. Introduction

In the decade from 2009 to 2019, there has been an eleven-fold increase in offshore wind power generation in Europe, illustrated in figure 1, this has taken offshore wind generation from 2 GW of generation to 22 GW. There has also been over 100 GW of increased generation onshore at the same time[1]. This has been a substantial change to the energy mix of the EU, and there are even greater increases set to come in the next decade.

There is a massive drive to install more wind capacity in the EU, and this will lead to even larger farms of turbines in operation requiring greater oversight. Operators will also be overseeing a greater number of turbines than ever before. With this greater capacity, operations and maintenance (O&M) will require better techniques and strategies to improve efficiency and reduce the costs. According to Wood MacKenzie[2], O&M costs were projected to be $ 15 Billion in 2019, and 57% of this would be on unplanned repairs and corrective maintenance.

Wind farms offshore have a failure rate per year on average of 10 per turbine. Of this, 20% of failures are major maintenance actions. The pitch system has one of the highest failure rates, and one of the lowest repair times[3], however offshore a high failure rate can lead to exacerbated downtimes even if the repair time is low. This is due to the lower accessibility offshore, which can be accounted for by weather conditions and vessel availability. According to [4] the pitch system, which was considered part of the hub, had an average downtime of 0.64 days per year, and another paper[5] has stated that the system fails between 0.1 and 0.3 times per year.

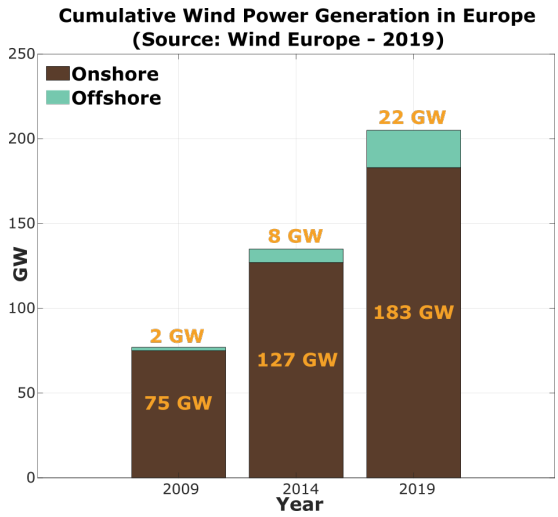

Figure 1. Illustration of increase in Offshore wind in Europe, from 2009 to 2019. Data taken from Wind Europe[1].

Turbine downtime can be reduced by utilising condition-based maintenance, over conventional methods such as corrective or scheduled maintenance. This approach can reduce the frequency of unplanned downtimes, and the associated costs. This requires both monitoring and analysing the component condition, and then scheduling maintenance based on the condition[6] - allowing to plan around weather and vessel availability. One method of analysis that is being considered is normal behaviour modelling, which tries to build a model of the normal data and predict future performance based on that - then if there is an error this can be analysed for potential warning signs.

The performance of condition monitoring techniques can be improved by using data pre-processing before the analysis. One such method is data-cleaning, where data outliers are removed from the data-set to improve the performance of normal behaviour modelling. In this paper a novel data cleaning technique is presented based on the use of the Mahalanobis distance. The data was cleaned through binning of wind turbine power curves, then applying Mahalanobis distance to each bin. This method was also compared against several other techniques for validation.

This paper aims to provide a novel user friendly data cleaning technique to remove anomalous data without much loss of information. The objectives of this paper is to test the presented technique on multiple turbines, and against other similar techniques. This testing is done on various metrics, to help assess the performance of the techniques examined. To the best of authors' knowledge, the use of Mahalanobis distance with data binning has not been used previously for cleaning of wind turbine power curves. The use of this in combination with an above rated filter has also previously not been considered.

The structure of the paper is as follows. Section 2 reviews previous literature in condition monitoring for wind turbines. Section 3 outlines the method used in this paper along with a description of the data in the case studies considered. Finally, section 4 presents and discusses the results from the method outlined in section 3.

2. Literature Review

Several papers have presented novel techniques for wind turbine data cleaning in recent years. One example of a novel data cleaning technique was produced by Xiaojun Shen et al.[7]. The authors used change point detection (CPD) on the variance of the binned power values taken

from SCADA data. This was the first step of data removal, and then a second step using 1.5 times the inter-quartile range (IQR) to remove outliers above the wind power curve, as CPD could not remove these effectively. The data deletion rate was used as a comparison metric, and proved to perform better than Local Outlier Factor (LOF) and IQR then CPD, instead of CPD then IQR.

A paper from A. Li Yuan et al.[8] used least squares to model a fitted power curve from SCADA data. This was fit to the average wind speed - power values from data bins. Upper and lower control limits, consisting of 3 standard deviations from the best fit line, were used to remove abnormal data.

Image based techniques have been used previously by Y. Su et al,[9]. The power curve was converted to a binary image with any pixel containing data converted to white. The data was considered anomalous based on their median distance to the bounday of a bright pixel group.

Another paper[10] has looked at the use of K-means clustering to initialise cluster centres before then clustering again with manifold spectral clustering.

Another image based method, from H. Long et al.[11], converted scatter data to a binary image. The principal part of the power curve is extracted and the Hu moments calculated. The extracted reference curve is considered normal, and anything outside this boundary is abnormal.

A paper by J. Wu[12] modelled a power curve by finding wind speed values and calculate the corresponding power values. Centre values of these values were used to plot the power curve, and upper and lower envelope lines obtained. This removed much of the outliers, then the data was binned and the Inter Quartile range was applied to the data to remove the rest of the outliers.

A similar technique from Shen, Fu and Zhou[7], first applied Change Point Detection to binned and ordered power values to first find the lower change point. This removed much of the outliers in the power curve. However, outliers still occurred above and below the curve, so the Inter Quartile Range was used to remove the rest.

Another image based method, from Z. Wang et al.[13], converted the curve to a binary image. First non-continuous vertical pixels were removed, then this was repeated for the horizontal direction. This removed much of the outliers in the curve.

A paper by Z. Lin et al.[14] compared the use of elliptic envelope and isolation forest to cleaned power curves from the Levenmouth 7 MW demonstrator turbine. This was done to pre-process data before use in a predictive deep learning neural network. Isolation forest was found to be better than Elliptic Envelope for this purpose.

A paper from Y. Bao et al.[15] utilised k-nearest neighbours in data bins to remove outliers before approximating the power cuvre with a least squares B-spline curve.

Another paper that utilises the inter quartile range (IQR) is from T. Yuan[16], which first clusters data using DBSCAN into core, border, and noise clusters. The noise is first removed, then the IQR is used to set upper and lower data limits to remove the remainder of the outliers.

3. Methodology

3.1. Data

The data provided for this paper was taken from the Supervisory Control and Data Acquisition (SCADA) system of four multi-megawatt wind turbines situated in South America. The SCADA system records data for multiple wind turbine components and down-samples it to 1 recording every 10 minutes. Each data record was approximately 2 years in length, and each was split evenly into training and testing datasets.

Two of the turbines provided did not fail within the recording period, these are classed as healthy turbines in this paper. The other two turbines did fail within the recording period, and these are classed as unhealthy. The unhealthy turbines failed from the same fault in the pitch system, which was located in the bearing. The healthy turbines were used to assess the technique's ability to retain important information when removing data. The unhealthy turbines were used to assess how well the overall method worked with those turbines that failed during

the time period examined.

3.2. Data Cleaning

Several techniques were compared in this paper, and each consisted of multiple stages. The first stage in the model was to either remove, any data with either a negative power value, or a power value over 2500 kW (for simplicity this is referenced to as the negative power filter). These outliers are typically due to sensor error or turbine downtime. This stage was accomplished through simple filtering of the data in Python, and the effect can be seen in figure 2

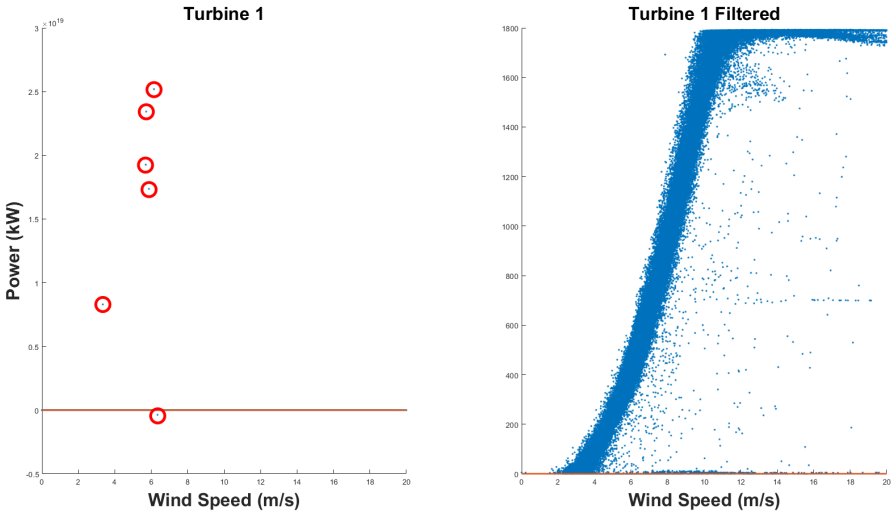

Figure 2. Comparison of Turbine 1 power curve, with and without filter. The red circles highlight where excessive readings occurred, and the orange line is positioned at 0 kW.

Second, data bins of roughly 1 m/s width were constructed. The mean power and wind speed values of these bins were found as cluster centres. The cluster centres found during binning were then used to find the Mahalanobis distance of each data-point in the bin. This is shown in equation 1:

$$D_{Mahalanobis}(x,y) = \sqrt{(x-y)C^{-1}(x-y)} \tag{1}$$

Mahalanobis distance appears to be more suitable for this data as it takes into account the covariance of the data. This is superior to the more conventional Euclidean distance, shown in equation 2, which assumes the data is distributed in a circle, whereas data in the power curve bins are distributed in truncated ellipses. The Euclidean distance was also compared in this paper, with the previous filtering and binning steps being the same. To remove data, a threshold of 2 standard deviations of the distance was used for both the Mahalanobis and Euclidean distances. This, in theory, should remove only around 5% of the data.

$$D_{Euclidean}(x,y) = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{2}$$

Similarly, Local Outlier Factor (LOF) was compared using the same filtering and binning steps. This technique, described in detail in [17] first finds the k nearest neighbours to each point in the data, then the "reachability distance" of this point to others in the dataset. The average of these distances are found for the k nearest neighbours, which is called the local

reachability distance. The LOF is then the average ratio between the local reachability distance of the point and its nearest neighbours.

Wind turbine blades are typically pitched to feather to control power output in wind turbines, this is typically done above rated wind speeds however during regimes of curtailment this can be done below rated. This can lead to straight horizontal lines under the power curve where the turbine is operating at a fixed power over a range of wind speeds. As another method for comparison, any datapoint corresponding to below rated pitch activity was removed. As this is one of the simpler techniques used currently, it was added as a baseline test to compare the slightly more complex methods against a relatively simple one. Here no binning was performed, however the negative power filter was still applied.

Finally, a comparison was made between the techniques described, and their counterparts that were only applied to data below rated. For the above rated data, a general filter of any power value lower than 1600 kW was removed. It was found that above rated the data was sparse, and it may not be suitable to use Euclidean or Mahalanobis distance as the standard deviation threshold assumes the data in a Gaussian Distribution. The power values above rated were scattered sparsely from 0 to over 2000 kW, and with most being clustered around rated power. The distribution of this is closer to a Poisson distribution, and therefore a filter to remove the tail of the data was considered. This change in distribution is brought on by the rated power acting as an upper bound on the data.

These techniques were compared against a technique from a paper by Shen, Fu, and Zhou[7] presented in Section 2. This involved the use of Change Point Detection (CPD) and Inter-Quartile Range (IQR). Binning was applied to the data, then data was initially removed using Change Point Detection, then Inter-Quartile Range was applied to the data to remove further outliers. This technique was compared against to provide context of this paper within the literature.

4. Results

4.1. Effectiveness of Technique

This section examines the effectiveness of the presented technique. The results for four turbines are compared, two of which did not fail in the time recorded and two did fail. For comparison, each turbine dataset was cleaned with and without the above rated filter presented in Section 3. Figure 3 presents the power curves for each of the turbines, here turbines 1 and 2 are considered healthy, and turbines 3 and 4 are unhealthy. The raw datapoints are plotted in blue, whilst the cleaned datapoints are in orange. The graphs on the right of the image also applied the above rated filter. The negative power filter was applied before the technique presented in this section.

Table 1 presents the percentage of data removed from the original datasets for each turbine by each cleaning technique, and the percentage of data removed from the filtered data. All models presented were applied on filtered data, except for the CPD-IQR method which was applied only on the raw data. The negative power filter appears to remove between 10 and 19% of data from the original, along with the outliers presented in figure 2. The models with the above rated filters applied are noted with an "AR" in table 1.

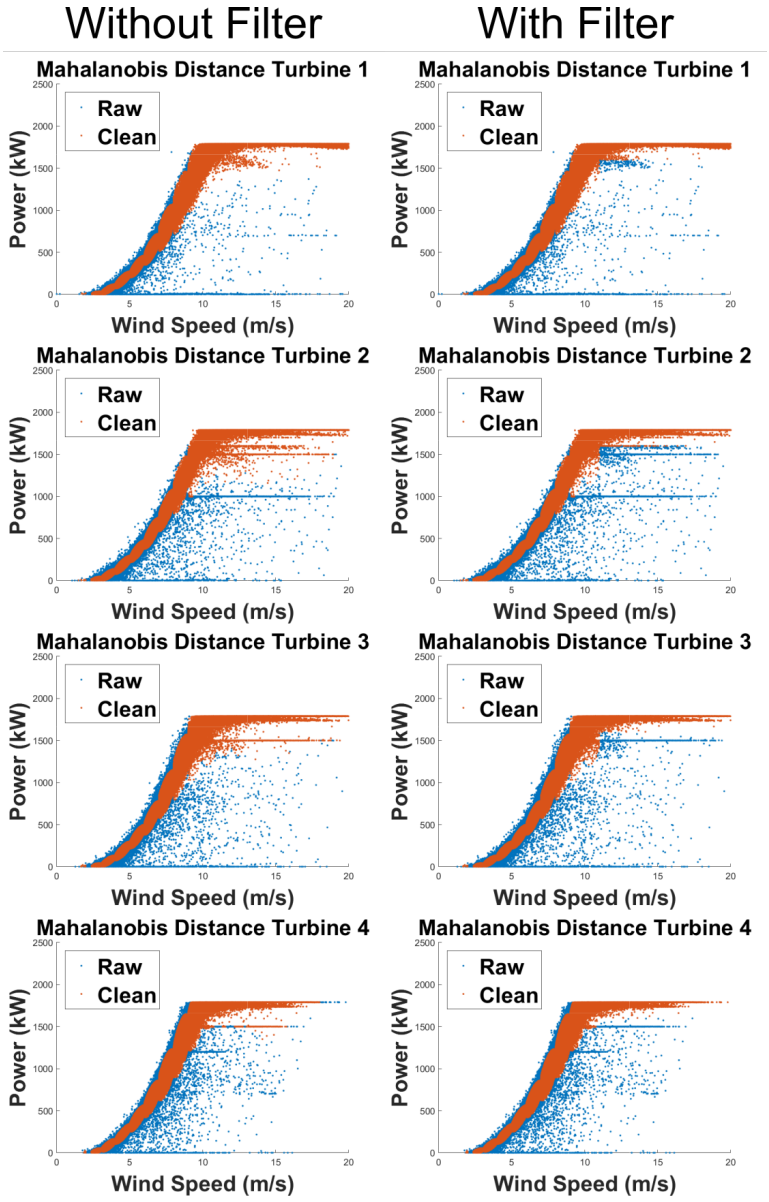

Figure 3. These graphs compare the effect of the cleaning technique with and without the above rated filter.

The effect of the above rated filter can be seen in figure 3, where more anomalies below the graph are removed. In particular for turbines 2, 3, and 4 the horizontal lines just below rated power, that typically indicate power curtailment, are filtered out. These were missed by the original cleaning model, and can affect normal behaviour models. By comparing the effect of the Mahalanobis model, with and without the above rated filter, it can be seen from table 1 that less than 1% of the data is removed by this filter. Wind turbines above rated power are controlled with the pitch system to remain at rated power at any wind speed, so any power values that vary from above rated can be considered different from normal wind turbine behaviour. Whilst wind power curtailment is planned behaviour by turbine operators, it is abnormal to the

ideal power curve which aims to produce rated power at all times above rated wind speed, and below cut-out. Therefore, it is important to remove this data as it can affect the performance of power prediction, or fault detection, as this planned behaviour will affect the model and produce errors during normal turbine operation at above rated wind speeds. While most of the above rated outliers are removed by the Mahalanobis cleaning method, some curtailment data remains, therefore the use of the above rated filter here is appropriate in removing the extra data that can affect normal behaviour modelling. It is possible that by using the median wind and power values to determine bin centres that this could remove the need for the above rated filter. The mean of the power values will be pulled down by sparse low power values, whereas a median would remain within the more dense cluster near rated.

The model is shown to have similar effects on each turbine, removing much of the outlier data below the curve, and the outliers found above the curve, such as those in figure 2. This is evident for the models with and without the above rated filter. Of the four turbines, only turbine 1 appears to have very little curtailment above rated, compared to the others. For all turbines, it appears the Mahalanobis method removes between 5 and 6% of the negative power filter data, and the same method with the above rated filter removes between 5 and 7.5%. As the threshold used is based upon the standard deviation this is to be expected.

By looking at figure 3, the amount of normal information being retained by each model can be compared for each turbine. Each turbine appears to have most of the normal data retained above rated, with some normal data being removed from below rated. There is also a slight step effect in the below rated regime due to the binning of the data and applying the Mahalanobis distance method. This could potentially be solved by slightly overlapping the bins and only removing data where outliers are detected in both bins. Below rated there seems to be data removed from the main cluster of the power curve, it may be that the threshold is too low for abnormal data below rated, and therefore a threshold of 3 standard deviations could maybe be applied in the future below rated. This would need to be tuned to retain as much of the normal data as possible.

A way to perhaps improve on this technique is to introduce a level of supervision into the technique. This could be done by including data labels, such as curtailment, turbine faults, and normal data, even as a method of evaluating the technique. This context would help to identify data for removal, however only some of this data could be removed when applied online as fault data would not be available in real time. This could potentially allow for a first step of removing any data flagged as curtailment, then further data cleaning could be done, any data left that is abnormal would be sparse and random in nature which is well suited to this technique.

Overall, this technique appears to remove much of the abnormal data, without sacrificing too much normal data to do so. The above rated filter works well at removing curtailment data that is missed by the unfiltered technique. The ideal power curve is preserved by the Mahalanobis distance method, however there are some areas that could be improved by future work, such as the loss of some normal data below rated.

4.2. Comparison of Cleaning Techniques

This section presents the results of the different cleaning techniques outlined in section 3. Figures 4 and 5 show the cleaned power curves for turbine 1 only, for each of the different techniques, with and without an above rated power filter respectively. The six techniques compared were the negative power filter, the Mahalaobis Distance, the Euclidean Distance, Local Outlier Factor, Below Rated Pitching, and the Change Point Detection[7].

Figure 4 presents the cleaned power curves for the models compared with the above rated filter. The negative power looks unchanged, however it has removed the negative power values. The Mahalanobis and Euclidean distance methods look very similar, however what is most noticeable is how squared the stepping is below rated for the Euclidean distance method. This is due to the Euclidean distance itself, as it assumes a circular distribution whereas the Mahalanobis

Table 1. Percentage of Data removed by each data cleaning technique. Results bolded are from method presented in this paper.

Model	Datapoint % Difference from Original				Datapoint % Difference from Filtered			
	Turb. 1	**Turb. 2**	**Turb. 3**	**Turb. 4**	**Turb. 1**	**Turb. 2**	**Turb. 3**	**Turb. 4**
1) Original	0.00	0.00	0.00	0.00	N/A	N/A	N/A	N/A
2) Negative Power Filter	-17.96	-10.21	-12.19	-18.39	0.00	0.00	0.00	0.00
3) Maha	**-22.14**	**-16.10**	**-17.19**	**-22.57**	**-5.10**	**-6.55**	**-5.69**	**-5.12**
4) Maha_AR	-22.24	-16.95	-17.84	-22.83	-5.22	-7.50	-6.44	-5.44
5) Euc	-20.76	-15.47	-15.95	-21.71	-3.42	-5.86	-4.28	-4.06
6) Euc_AR	-20.86	-16.35	-16.60	-22.00	-3.54	-6.83	-5.03	-4.42
7) LOF	-18.01	-10.27	-12.22	-18.43	-0.06	-0.06	-0.03	-0.05
8) LOF_AR	-18.26	-12.45	-13.04	-19.07	-0.37	-2.49	-0.97	-0.83
9) Pitch	-22.47	-17.40	-19.47	-27.42	-5.50	-8.01	-8.29	-11.06
10) CPD & IQR	-56.14	-29.56	-34.39	-27.29	-46.54	-21.55	-25.28	-10.91

distance takes into account the shape of the distribution. Therefore, the Mahalanobis distance takes into account the angle of the distribution below rated, which is why this step effect is less noticeable. Above rated they both appear to behave the same. This step effect is caused by the Euclidean distance assuming each bin cluster is distributed as a circle, and the data is bounded on both sides by the bin boundaries.

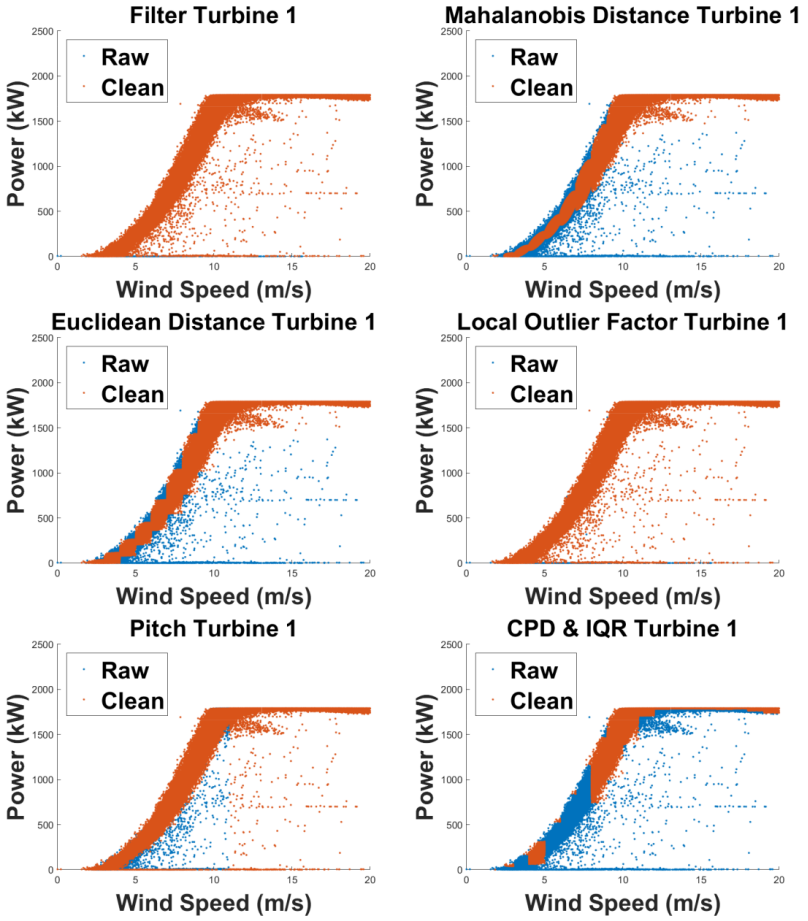

Figure 4. Each graph compares the clean and raw dataset for each cleaning technique without the above rated power filter.

The local outlier factor appears to make no difference to the data, and it is easy to mistake it for the filtered data. The below rated pitch method only seems to remove any data under the curve in the below rated regime, except for zero power. Above rated all of the outliers are included in the cleaned data, so this is not so suitable. Finally, the CPD & IQR method seems to perform very poorly. Much of the "normal" data is removed, along with other outliers. It is unknown what caused this as the method was reproduced as described in [7].

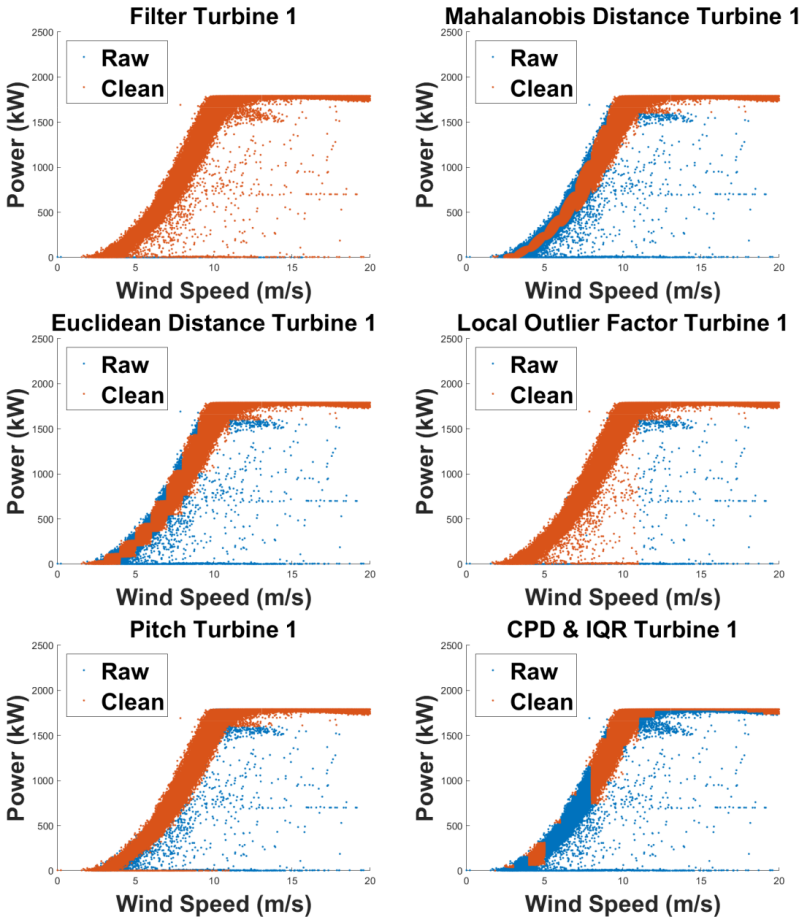

Figure 5. Each graph compares the clean and raw dataset for each cleaning technique with the above rated power filter.

Figure 5 presents the cleaned power curves using the above rated filter for all models other than the negative power filter method. For the Mahalanobis and Euclidean distance methods, the effect is much the same as discussed in the previous section. Any outliers beloww 1600 kW are removed, including any curtailment behaviour there might be. Local Outlier Factor and Pitch have above rated outliers removed, and improve their cleaning abilities over not using the above rated filter. There appears to be no discernible difference for the CPD & IQR method.

Overall, when the above rated filter is applied, and when it is not, the Mahalanobis method appears to retain more of the "normal" data than other techniques while removing as much abnormal data as possible. The technique itself is rather simple and user friendly, as it does not require anything beyond data binning and distance calculation. This is similar to the Euclidean distance, however the below rated curve is changed too much. Local outlier factor and the below rated pitch method are similarly user friendly, however their performance is unsatisfactory.

The two most successful techniques are the Euclidean and Mahalanobis distance methods, these both retain much of the normal data and remove all sparse abnormal data. Due to the introduction of the stepping effect below rated by the Euclidean distance, it is preferable to use the Mahalanobis distance. This technique also appears to be quite robust as the model performs similarly across all turbines examined. Further examination with models of a different turbine type would be useful to investigate whether this technique is still appropriate, and

further comparison with the CPD & IQR method may be required to assess how robust the model is.

5. Conclusion

This paper presented a novel data cleaning technique for pre-processing SCADA data in wind turbine normal behaviour modelling for condition monitoring. The method presented first used a filter to remove negative power values, then a combination of binning and the Mahalanobis distance to remove outliers from the power curve. A filter for the above rated regime of the turbine was also considered.

This technique was compared for 4 turbines, and against a number of other cleaning techniques. It was found that the technique removed only a small percentage of the data, between 5 and 6%, and retained much of the normal data in the power curves. The Mahalanobis technique, combined with the above rated filter, has been shown to be the most effective of the techniques presented. This paper shows that the Mahalanobis technique is not only user-friendly and conceptually simple, it also proved to be the best model examined. The model was also shown to be quite robust across all turbines, however some data was removed from the normal data below rated and potentially this could be solved by changing the threshold below rated.

Future work should be aimed at utilising this cleaning technique within the framework of a condition monitoring tool. This technique, combined with appropriate feature and model selection, could potentially improve condition based maintenance method.

Acknowledgements

This research was funded by EPSRC grant number EP/L016680/1.

[1] Wind Europe 2019 Wind Energy in Europe in 2019 - Trends and statistics URL https://windeurope.org/data-and-analysis/product/wind-energy-in-europe-in-2019-trends-and-statistics/
[2] 2019 Unplanned wind turbine repairs to cost industry $8 billion+ in 2019 URL https://www.woodmac.com/press-releases/unplanned-wind-turbine-repairs-to-cost-industry-8 − *billion* − *in* − 2019/
[3] Carroll J, McDonald A and McMillan D 2016 *WIND ENERGY* **19** 1107–1119 ISSN 11283602
[4] Faulstich S, Hahn B and Tavner P 2011 *Wind Energy* **14** 327–337
[5] Pinar Pérez J M, García Márquez F P, Tobias A and Papaelias M 2013 *Renewable and Sustainable Energy Reviews* **23** 463–472 ISSN 13640321 URL http://dx.doi.org/10.1016/j.rser.2013.03.018
[6] Nielsen J J and Sørensen J D 2011 *Reliability Engineering and System Safety* **96** 218–229 ISSN 09518320 URL http://dx.doi.org/10.1016/j.ress.2010.07.007
[7] Shen X, Fu X and Zhou C 2019 *IEEE Transactions on Sustainable Energy* **10** 46–54 ISSN 19493029
[8] Yuan L, Qiujuan H, Yi Y and Zuoxia X 2018 Abnormal State Analysis of Wind Turbines Based on the Power Curve *2018 International Conference on Power System Technology (POWERCON)* (Guangzhou: IEEE) pp 4135–4142 ISBN 9781538664612
[9] Su Y, Chen F, Liang G, Wu X and Gan Y 2019 *IEEE International Conference on Robotics and Biomimetics, ROBIO 2019* 1198–1203
[10] Hu Y, Xi Y, Pan C, Li G and Chen B 2020 *Renewable Energy* **146** 2095–2111 ISSN 18790682 URL https://doi.org/10.1016/j.renene.2019.08.043
[11] Long H, Sang L, Wu Z and Gu W 2020 *IEEE Transactions on Sustainable Energy* **11** 938–946 ISSN 19493037
[12] Wu J H, Shao Z G and Yang S H 2020 *Journal of Physics: Conference Series* **1639** ISSN 17426596
[13] Wang Z, Wang L and Huang C 2021 *IEEE Transactions on Instrumentation and Measurement* **70** ISSN 15579662
[14] Lin Z, Liu X and Collu M 2020 *Electrical Power and Energy Systems* **118** 105835 ISSN 0142-0615 URL https://doi.org/10.1016/j.ijepes.2020.105835
[15] Bao Y, Pan D, Wang X, Liao L and Yang Q 2017 Least-square B-spline Approximation Based Wind Turbine Power Curve Modeling *2017 Chinese Automation Congress* pp 6711–6716
[16] Yuan T, Sun Z and Ma S 2019 *Energies* **12** ISSN 19961073
[17] Xu X, Lei Y and Zhou X 2018 A LOF-based method for abnormal segment detection in machinery condition monitoring *2018 Prognostics and System Health Management Conference (PHM-Chongqing)* (IEEE) pp 125–128