Submitted to the Annals of Applied Statistics

COVARIATE MATCHING METHODS FOR TESTING AND QUANTIFYING WIND TURBINE UPGRADES

By Yei Eun Shin[†], Yu Ding[†] and Jianhua Z. Huang[†],

Texas A&M University [†]

In the wind industry, engineers perform retrofitting upgrades on in-service wind turbines for the purpose of improving power production capabilities. Considering how costly an upgrade can be, people often wonder about upgrade effect: whether it indeed improves turbine performances, and if so, how much. One cannot simply compare power outputs for the purpose of assessing a turbine's improvement, as wind power generation is affected by an array of environmental covariates, including wind speed, wind direction, temperature, pressure as well as other atmosphere dynamics. For a fair comparison to discern the upgrade effect, it is critical to have these environmental effects controlled for while comparing power output differences. Most existing approaches rely on establishing a power curve model and let the model account for the environmental effects. In this paper, we propose a different approach, which is to devise a covariate matching method to ensure the environmental covariates to have comparable distribution profiles before and after an action of upgrade. Once the covariates are matched, paired t-tests can be applied to the power outputs for testing the significance of the upgrade effect. The relative increase in power production can also be quantified. The proposed approach is simple to use and relies on fewer assumptions than the power curve modeling approach.

1. Introduction. Wind power is one of the fastest growing renewable energy resources [DOE (2015)]. As large wind farms are built, cost considerations are essential for effective wind farm management [Byon et al. (2013)]. One of the costly management actions for in-service turbine fleet is to perform retrofitting upgrades, so that outdated or malfunctioning wind turbines can restore or even improve their power generation capability [Khalfallah and Koliub (2007)]. It is, therefore, not a surprise that operators want to know whether the benefits from an upgrade outweigh the expenses of doing it, including material and labor cost. This inquiry motivates researchers to scrutinize turbine performances before and after an upgrade. It becomes

[*]The authors would also like to acknowledge the generous support from their sponsors. Ding was partially supported by NSF grant CMMI-1300236 and Qatar National Research Fund NPRP 7-953-2-357. Huang was supported by NSF grant DMS-1208952.

Keywords and phrases: Causal inference, Mahalanobis distance, Matching methods, Nearest neighbor matching, Observational study, Wind power curve

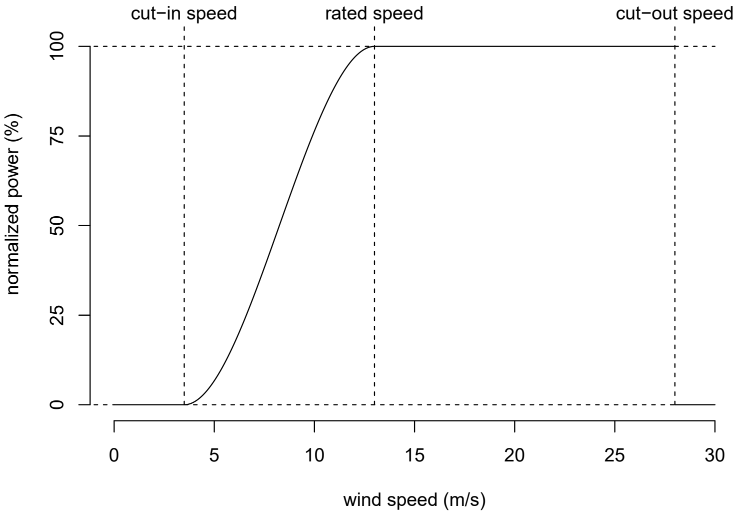


FIG 1. *Wind power curve. Wind turbine produces higher power as wind speed increases. A turbine starts power production at the cut-in speed, reaches its full operation at the rated speed, and stops producing power at and beyond the cut-out speed. Power outputs are normalized by the rated power.*

the research question we aim to answer in this paper, and if an upgrade does indeed improve turbine performances, we also want to quantify the improvement.

When it comes to comparing turbine performances between the periods before and after an upgrade, it is unreasonable to merely compare power outputs of the two periods because wind power generation is affected by an array of environmental covariates, such as wind speed, wind direction, temperature, air pressure and other atmosphere dynamics. Each of the environmental covariates observed before an upgrade may probabilistically distribute differently from the period after an upgrade. These incomparable input conditions cause different wind power outputs and could mislead the conclusion: for example, if too many windy days are there after an upgrade, high power generation might happen due to not only the upgrade effect but more so due to the high wind speed. For a fair comparison, therefore, these environmental effects need to be controlled for while comparing power outputs.

To handle the problem explained above, the dominating approach is to establish a model estimating wind power outputs conditioned on the observations of environmental covariates, so that the model can be used to compare the estimated power outputs between the two periods by setting the same input conditions. Such a model, if taking wind speed as a single input, is known as a power curve, explaining the functional relationship between wind power output and wind speed input [Ackermann and Söder (2005)]; Figure 1 presents an example.

To estimate a power curve using actual wind speed and power observations, the International Electrotechnical Commission [IEC (2005)] recommended the use of a binning method, which discretizes wind speed into intervals of, say, 0.5 meters per second (m/s) width and then uses the wind power data and wind speed records, averaged in respective intervals, to fit a smooth curve. Other curve fitting methods are also developed for estimating a power curve based on wind speed [Yan, Osadciw, Benson, and White (2009); Kusiak, Zheng, and Song (2009); Uluyol, Parthasarathy, Foslien, and Kim (2011); Osadciw, Yan, Ye, Benson, and White (2010); Albers (2012)], but they may be different from the binning method in specifics.

A common drawback of the IEC like approaches is that they regard wind speed too heavily as a factor driving the power production. While it is true that wind speed is the most significant effect in wind power generation, other environmental effects cannot be ignored. In an effort to include other environmental factors into an extended power curve model, the effect of wind direction was incorporated, in addition to wind speed [Nielsen, Nielsen, and Madsen (2002); Sanchez (2006); Pinson, Nielsen, Madsen, and Nielsen (2008); Jeon and Taylor (2012); Wan, Ela, and Orwig (2010)]. Most recently, Lee, Ding, Genton, and Xie (2015a) and Lee, Ding, Xie, and Genton (2015b) developed one of the first truly multivariate-dependency wind power models that allows all aforementioned environmental covariates to be included. Understandably, such a model, if fitted separately before and after an upgrade, could be used to compare a turbine's performance by setting input conditions at the same values.

In this paper we advocate a different approach. Its basic idea is as follows. Suppose that one can select a large enough subset of wind turbine data before and after an upgrade, such that they have comparable distribution profiles of the environmental covariates. Then one can simply compare the wind power outputs of the two periods within that selected subset. The appeal of such a direct comparison approach is its simplicity. Unlike the model-based approaches (to fit a power curve is to estimate a model), it relies on fewer assumptions. Additionally, the direct comparison approach is quick to be carried out in practice, and its working mechanism is easy to be understood by engineers. The last point is important because a method is less likely to have real impact in practice until it is understood and thus accepted by practitioners.

Covariate matching methods are rooted in the statistical literature. In stabilizing the non-experimental discrepancy between non-treated and treated subjects of observational data, Rubin (1973) adjusted covariate distributions by selecting non-treated subjects that have a similar covariate condition as that of treated ones. Through the process of matching, non-treated and treated groups become only randomly different on all background covariates, as if these covariates were designed by experimenters. As a result, the outcomes of the matched non-treated and treated groups, which keep the originally observed values, are comparable under the matched covariate conditions. For more discussion on covariate matching methods, please refer to Stuart (2010).

In this paper, we propose a covariate matching method tailored towards wind application, in which records from a turbine before and after an upgrade correspond to non-treated and treated subjects, respectively. We follow the four key steps for a matching method, introduced in Stuart (2010), of which the first three steps represent the *design* of a matching method, whereas the fourth step represents the *analysis* of the matched outcomes:

1. Define the measure of closeness;
2. Implement a matching method;
3. Diagnose the quality of the resulting matched samples;
4. Analyze the outcome and estimate the treatment effect.

Specifically in our approach, we use the Mahalanobis distance [Mahalanobis (1936)] in Step 1 to determine whether an individual is a good match to another. In Step 2, we adopt an idea of the $k:1$ nearest neighbor matching method [Rubin (1973)]. In Step 3, we rely primarily on density plots as our diagnostic tool. As the last step, we analyze the matched outcomes through paired t-tests and compute the improvement an upgrade makes.

We want to note that in the field of wind power analysis, there exist *analog* techniques, which have a similar idea to the matching methods, in that they search for and utilize a set of observations that have the most similar weather condition to the specific time point. Since these analog approaches typically aim at forecasting, they then estimate the probability distribution of the future state of atmosphere [Delle Monache, Eckel, Rife, Nagarajan, and Searight (2013)]. However, the covariate matching methods discussed above, including the proposed one, differ from the analog forecasting approaches, in that the covariate matching methods aim at investigating a treatment effect, or specifically, an upgrade effect in our context. They also do the investigation without any estimation procedure unlike the other approaches. Another difference is that the analog methods follow a timeline to find the most similar weather path to the time of interest, whereas the covariate matching methods break the time order of non-treated records to construct the counterpart of treated ones.

The remainder of this paper is organized as follows. In Section 2, we describe the data structure. In Section 3, we propose a matching method for handling wind turbine data. Section 4 presents an outcome analysis, including the quantification of the upgrade effect. Section 5 performs a sensitivity analysis to verify our approach's capability in estimating the upgrade effect and to compare it with a power curve modeling approach. We make a few further remarks concerning the proposed matching method in Section 6. Finally, we summarize the paper in Section 7.

2. Data structure.

In this study, we use data obtained from the authors of Lee et al. (2015b). For this reason, we study the same two upgrade cases as in Lee et al. (2015b). We would like to explain briefly the setting under which the data are obtained.

This study involves two pairs of turbines, which are distant apart enough, so that one pair of turbines does not affect the other pair. Within a pair, one turbine is called a test turbine on which an upgrade is applied, while the other one is called a control turbine of which no change is made. We deem the two turbines in a pair are identical for practical considerations, as they are of the same type from the same manufacturer and started their service at the same time. Both turbines in each pair are also associated with a meteorological mast, which houses sensors to measure several environmental conditions. Figure 2, similar to Figure 5 in Lee et al. (2015b), illustrates the layout of the two turbine pairs and their associated mast.

As in Lee et al. (2015b), we consider two types of upgrade: one is known as a vortex generator installation [Øye (1995)] and the other one is a pitch angle adjustment [Wang, Tang, and Liu (2012)]; both actions are believed to make the upgraded turbine to produce more wind power under the same environmental conditions. The vortex generator installation is physically carried out on a test turbine in a pair and we call this pair the *experimental* pair, whereas the pitch angle adjustment is not physically carried out but simulated on a test turbine; we call the turbine pair with the simulated upgrade the *mimicry* pair.

The following data modification is done to the test turbine data in the mimicry pair. The actual wind turbine data, including both power production data and environmental measurements, are taken from the actual tur-

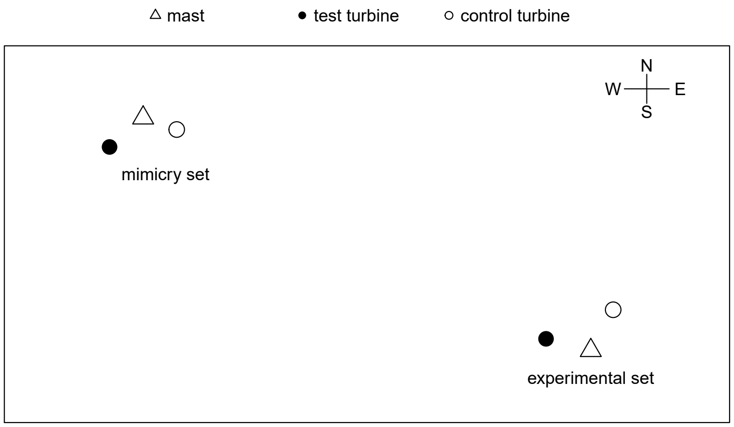


Fig 2. *Wind farm layout. This layout shows the relative locations of turbines and masts on a wind farm. Wind power production is measured at each turbine, and environmental conditions are measured by sensors at the nearby meteorological mast. An experimental pair includes an actually-upgraded test turbine (a vortex generator installation) and its control turbine, whereas a mimicry pair includes an artificially-upgraded test turbine (a pitch angle adjustment) and its control turbine.*

bine pair operation. Then, the power production from the designated test turbine on the range of wind speed over 9 m/s is increased by 5%, namely multiplied by a factor of 1.05; see Figure 3 for an illustration. This simulation of an pitch angle adjustment is motivated by Wang et al. (2012). Including the simulated data set in our study helps us get a sense of how well a proposed method can detect a power production change due to an upgrade and how accurately it can quantify the change.

We denote the power output of a turbine by P (in kilowatts), so that P^{ctrl} and P^{test} are associated with a control turbine and a test turbine, respectively. In this study, power output values are normalized by the rated power, to protect the identities of the turbine manufacturer and the wind farm operator.

Environmental conditions directly measured at a meteorological mast are: wind speed, V, wind direction, D, ambient temperature, T, and air pressure, Q. Using these measurements, the values of additional environmental covariates can be computed, including air density, A, wind shear, W, and turbulence intensity, I, using the following formulas:

- air density, $A = \frac{Q}{R \cdot T}$ (kg/m^3), where $R = 287$ (Joule/(kg·K)) is a gas constant;
- wind shear, $W = \frac{\ln(V_2/V_1)}{\ln(g_2/g_1)}$, which represents a vertical variation of wind,

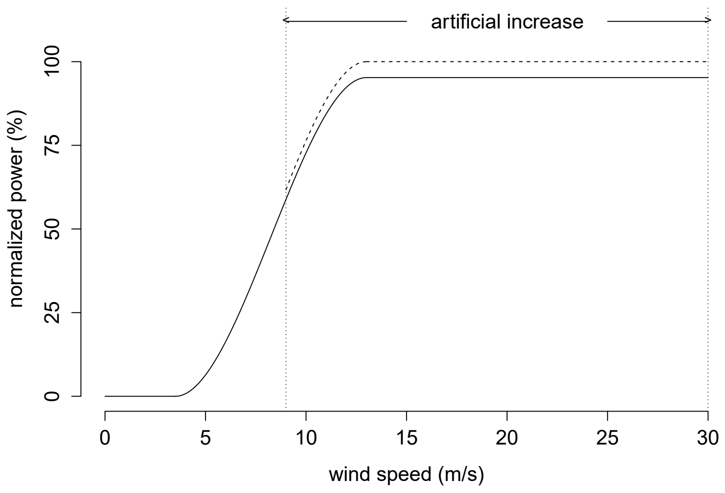


FIG 3. *The modification in the mimicry test turbine data as if a pitch angle adjustment were applied. The power on the range of wind speed over* 9 *m/s is increased by* 5%.

where V_1 and V_2 are wind speeds measured at heights $g_1 = 80$ m and $g_2 = 50$ m, respectively;

- turbulence intensity, $I = \frac{\hat{\sigma}}{V}$, where $\hat{\sigma}$ is the standard deviation of wind speed in a 10-minute duration.

The air density A represents the combined effect of temperature and pressure; once the air density is included to explain wind power outputs, temperature and pressure are no longer needed. The wind shear W and turbulence intensity I measure certain aspects of atmospheric dynamics that wind speed itself does not fully represent.

As such, each data set has five explanatory covariates, (V, D, A, W, I), and two power outcomes, $(P^{\text{ctrl}}, P^{\text{test}})$. Note that wind turbine data are arranged into 10-min blocks, so that the values of (V, D, A, W) are the averages of the 10-min intervals and I is the ratio of the standard deviation of wind speed in the 10-min blocks over the average wind speed of the same block. This 10-min block data arrangement is commonly used in the wind industry.

For the experimental pair, we have 14 months worth of data in the non-treated period (i.e., before the upgrade) and 5 weeks worth of data in the treated period (i.e., after the upgrade), whereas for the mimicry pair, we have 8 months worth of data in the non-treated period and 7 weeks in the treated period. Note that it is preferable to have a much larger set in the non-treated period than the treated. That is because a sufficiently large candidate pool to match can avoid too many of repeatedly selected individuals, and therefore the matched subset of the non-treated period reflects reality such as varying weather conditions.

3. Matching methods.

Our investigation starts off with exploring the discrepancy of the covariate distributions. Figure 4 demonstrates for each covariate the difference in empirically fitted density functions between the non-treated and treated periods. The last subplot in both the upper and lower panel is the density function of the power output of the respective control turbine. For the control turbine, as it is not modified, the distribution of its power output is supposed to be comparable, should the environmental conditions be maintained the same. But the data show otherwise, suggesting the existence of environmental influence, which confounds the upgrade effect in power outputs.

Let us introduce a few notations and terminologies. The environmental covariate vector is denoted by $\mathbf{X}$. In this study, $\mathbf{X} := (V, D, A, W, I)^T$, but it can include more variables, should their measurements be available. The data pair $(\mathbf{X}, P)$ forms a data record, containing the value of the environmental covariates and its corresponding power outputs. The data records collected before the upgrade form the non-treated data group, whereas those collected after the upgrade form the treated group. Let S_{bef} and S_{aft} be the index set of the data records in the non-treated and treated group, respectively. Let Y_S denote the values of a covariate Y for data indices in S. For example, $V_{S_{\text{bef}}}$ is the vector of all wind speed values that are observed before the upgrade.

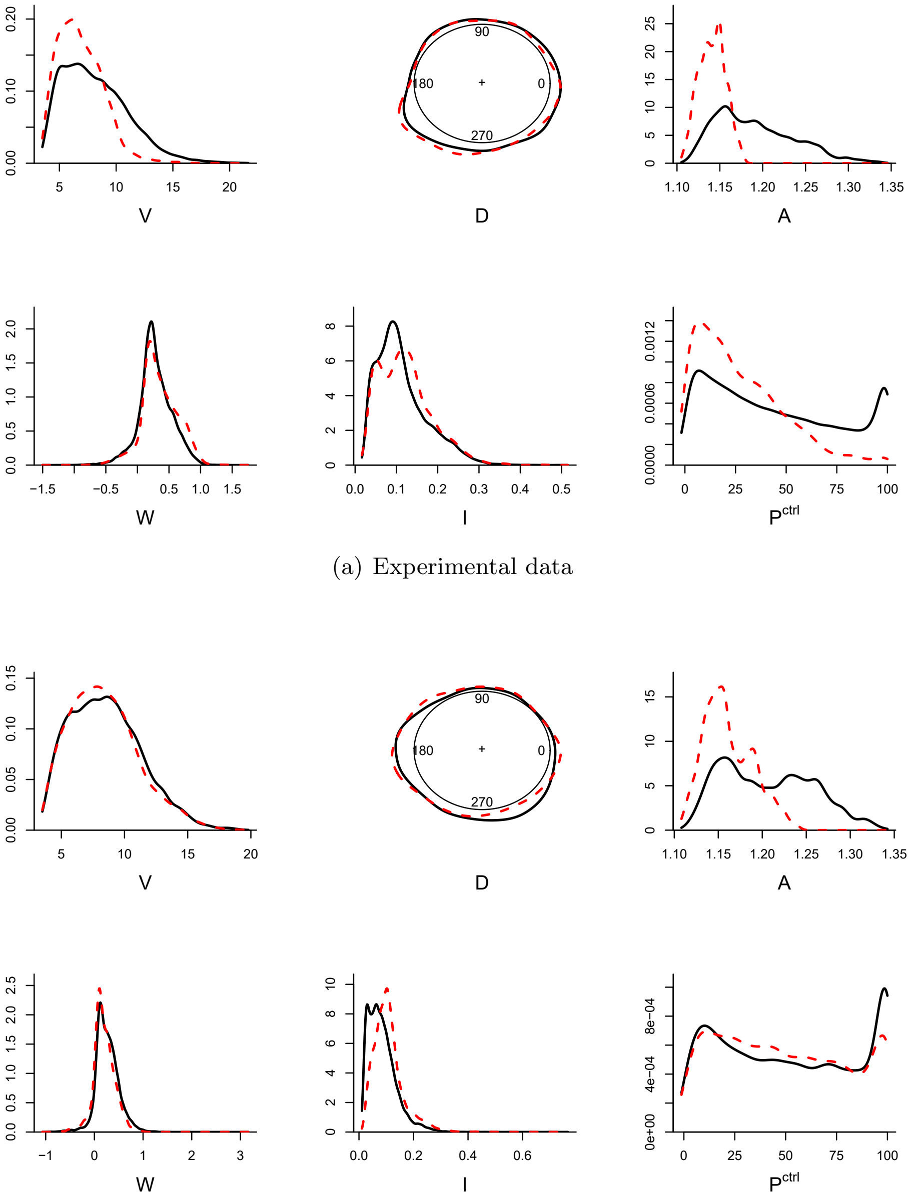
This section presents a matching method to create comparable distribution profiles of covariates. Before going through the four-step procedure of developing a matching method, as mentioned in Section 1, we first describe the preprocessing steps in Sections 3.1 and 3.2. Then, Sections 3.3, 3.4, and 3.5 describe Step 1, 2 and 3, respectively. Step 4 is discussed in Section 4.

3.1. *Hierarchical Subgrouping.*

The first action of preprocessing is to narrow down the set from which we will perform the data records matching subsequently. The reason for this preprocessing is to alleviate a computational demand arising from too many pairwise combinations when comparing two large size data sets.

This objective is fulfilled via a procedure we label as *hierarchical subgrouping*. The idea goes as follows.

1. Locate a data record in the treated group, S_{aft}, and label it by the



(a) Experimental data

(b) Mimicry data

Fig 4. *Overlapped density functions of unmatched covariates and power output of control turbine; solid line = before upgrade (non-treated), dashed line = after upgrade (treated).*

index j.

2. Select one of the covariates, for instance, wind speed, V, and designate it as the variable on which we measure similarity between two data records.
3. Go through the data records in the non-treated group, S_{bef}, by selecting the subset of data records such that the difference, in terms of the designated covariate, between the data record j in S_{aft} and any one of the records in S_{bef} is smaller than a pre-specified threshold. When V is in fact the one designated in Step 2, the resulting subset is then labeled by placing V as a subscript to S, namely S_V.
4. Next, designate another covariate and use it to prune S_V in the same way as one prunes S_{bef} into S_V in Step 3. This produces a smaller subset nested within S_V. Then continue with another covariate until all covariates are used.

The order of the covariates in the above hierarchical subgrouping procedure is based on the importance of them in affecting wind power outputs; according to Lee et al. (2015a), it is V, D, A, W, and I, from the most important to the least important. We will discuss more about the matching order of covariates in Section 6.1. Also note that wind direction D is a circular variable and an absolute difference between two angular degrees is between 0 and π; we then adopt a circular variable formula from Jammalamadaka and Sengupta (2001) to calculate the difference between two D values.

The above process can also be written in set representation. For a data record j in S_{aft}, we define subsets of data records in S_{bef}, hierarchically chosen, as

$$\begin{aligned}
S_V &:= \{i \in S_{\text{bef}} : |V_i - V_j| < \alpha_V \sigma(V_{S_{\text{bef}}})\};\\
S_D &:= \{i \in S_V : \pi - |\pi - |D_i - D_j|| < \alpha_D \sigma(D_{S_V})\};\\
S_A &:= \{i \in S_D : |A_i - A_j| < \alpha_A \sigma(A_{S_D})\};\\
S_W &:= \{i \in S_A : |W_i - W_j| < \alpha_W \sigma(W_{S_A})\};\\
S_I &:= \{i \in S_W : |I_i - I_j| < \alpha_I \sigma(I_{S_W})\},
\end{aligned}$$

where $\sigma(Y)$ is the standard deviation of Y and α_Y is a thresholding coefficient. We discuss how to determine these α's in Section 3.5. This hierarchical subgrouping establishes the subsets nested as such: $S_I \subset S_W \subset S_A \subset S_D \subset S_V \subset S_{\text{bef}}$. Consequently, the data records in the last hierarchical set S_I have the closest environmental conditions as compared with the data record j in S_{aft}.

This hierarchical subgrouping procedure shares certain similarity with the coarsened exact matching (CEM) approach [Iacus, King, and Porro (2012)], in that it performs the data records matching on broader ranges of covariates and builds factor-sized strata. Unlike CEM, however, the strata from our procedure have a hierarchical and nested structure that CEM does not have.

3.2. *Unmeasured Factors.* There could be other environmental conditions, in addition to V, D, A, W and I, which may affect wind power production while not measured. For instance, humidity is one variable that was shown to have an appreciable impact on wind power production for offshore wind turbines [Lee et al. (2015a)] but for the wind farm data we worked with, humidity was not measured.

The possible existence of unmeasured environmental factors presents the risk of causing a distortion in comparison, even when the aforementioned measured environmental factors are matched between the treated and non-treated groups. In order to alleviate this risk, we make use of the power output of the control turbine in each turbine pair, P^{ctrl}. What we propose is to further narrow down from the most nested subset produced in Section 3.1, S_I, by taking the following action – we select records from S_I whose P^{ctrl} values are comparable to the P^{ctrl} value of a data record j in S_{aft}. Specifically, this amounts to continuing the hierarchical subgrouping action in Section 3.1, producing a S_P, a subset of S_I, based on P^{ctrl}, such that

$$S_P := \{i \in S_I : |P_i^{\text{ctrl}} - P_j^{\text{ctrl}}| < \alpha_P \sigma(P_{S_I}^{\text{ctrl}})\}.$$

We perform this procedure for all data records in the treated group so that each record j in S_{aft} has its matched set $S_{P,j}$. In the case that $S_{P,j}$ is an empty set, we then discard the respective index j from S_{aft}. Because of this, S_{aft} may shrink after the subgrouping steps.

What we do in this subsection is essentially to use the control turbine to calibrate the conditions affecting the test turbine. A similar idea was tried by Albers (2012), but his approach is different from ours. Albers used a power curve based approach, in which the author fitted a *relative* power curve between the control and test turbines and hoped using that can calibrate the conditions for the test turbine. The rationale behind Albers's relative power curve is not as transparent as our subgrouping procedure and that approach is still model-based rather than direct comparison; in fact, it involved several modeling steps in its analysis.

3.3. *Mahalanobis Distance.* Denote $S_{P,j}$ as a set of candidate matches of data records in the non-treated group to a data record j in the treated group. Our next goal is to choose a data record in $S_{P,j}$ that is the closest to a data record j. For this purpose, we need to define a dissimilarity measure to quantify the closeness between two data records.

We decide to use the Mahalanobis distance [Mahalanobis (1936)] as our dissimilarity measure, which is popularly used in the context of multivariate analysis. It re-weighs the Euclidean distance between two covariate vectors with the reciprocal of a variance-covariance matrix. Before presenting the definition of the Mahalanobis distance between two wind turbine data records, we first introduce a transformed covariate vector, denoted by $\mathbf{X}^*$, such that

$$\mathbf{X}^* := (V \cos D, V \sin D, A, W, I)^T.$$

Using $\mathbf{X}^*$ makes it easier to deal with the circular wind direction variable D. The Mahalanobis distance ($\mathtt{MD}_{ij}$) between a data record j in S_{aft} and a data record i in $S_{P,j}$ is defined as

$$\mathtt{MD}_{ij} := \sqrt{(\mathbf{X}_i^* - \mathbf{X}_j^*)^T \Sigma^{-1} (\mathbf{X}_i^* - \mathbf{X}_j^*)},$$

where $\Sigma = \text{Cov}(\mathbf{X}^*_{S_{\text{bef}}})$. Obviously, the larger an $\mathtt{MD}$ value, the more dissimilar two data records.

Alternatively, the propensity score can be used as a dissimilarity measure [Rosenbaum and Rubin (1983)]. The propensity score has an advantage for a large number of covariates, whereas the Mahalanobis distance works quite well when there are fewer than eight continuous covariates [Zhao (2004)]. Moreover, since the Mahalanobis distance can reflect the interaction among covariates, which indeed exists in our data as described in Section 6.1, we choose the Mahalanobis distance rather than the propensity score.

3.4. *One-to-one matching.* As the simplest form of the $k : 1$ nearest neighbor matching, introduced by Rubin (1973), we perform the $1 : 1$ matching; it selects, for each treated record j, the non-treated record with the smallest distance from j. As the size of the matching candidates for each treated subject is reduced while undertaking the subgrouping step, there is no need to search in the entire non-treated group but simply within the resulting subgroup.

In a set representation, given $S_{P,j}$ and $\mathtt{MD}_{ij}$ from Section 3.2 and 3.3, respectively, we select the data record i_j in $S_{P,j}$ that has the smallest Mahalanobis distance as the best match to data record j in S_{aft}. That is, the data record i_j is found such that

$$i_j = \arg\min_{i \in S_{P,j}} \mathtt{MD}_{ij},$$

for each j in S_{aft}. In case that two or more are tied for the smallest value, we choose one of them randomly. After this step, each data record j in the treated group has one non-treated counterpart i_j, with the exception of those already discarded during the subgrouping step. We define the index set of the matched data records from the non-treated group as

$$S^*_{\text{bef}} := \{i_j \in S_{\text{bef}} \,|\, j \in S_{\text{aft}}\}.$$

As such, the data records in S_{aft} are now individually paired to those in S^*_{bef}.

It should be noted that we allow replacement in our matching procedure. In other words, i_j is not eliminated from the candidate set S_P, even though it has matched to j once. When the next data record $j+1$ is selected from S_{aft}, the same non-treated data i is thus possible to be matched again. We believe that allowing replacement helps achieve a fair matching because the data records in S_{aft} have no presumed order to be paired in advance. We will provide further discussions related to the matching with replacement in Section 6.2.

3.5. *Diagnostic.* After performing the matching procedure, it is crucial to diagnose how much the discrepancy of the covariate distributions has been removed, as compared to the original (unmatched) data set. Only after the diagnostics signifies a sufficient improvement, an outcome analysis is then ready to perform in the next step.

We measure the discrepancy of distributions in two ways, numerically and graphically. For the numerical diagnostics, the standardized difference of means (`SDM`) is used as a measure of dissimilarity of a covariate between the treated and non-treated groups [Rosenbaum and Rubin (1985)];

$$\texttt{SDM} := \frac{\overline{Y}_{S_{\text{aft}}} - \overline{Y}_{S_{\text{bef}}}}{\sigma(Y_{S_{\text{aft}}})},$$

where Y is one of the covariates, and $\overline{Y}_S$ denotes the average of Y in the set of S. The `SDM` decreases if the matching procedure indeed reduces the discrepancy between the two groups. As shown in Table 1, `SDM` decreases significantly for all covariates. A previous study [Rubin (2001)] found that `SDM` should be less than 0.25 to render the two distributions in question comparable. Otherwise, the differences between the distributions of covariates in the two groups are regarded as substantial.

For the graphical diagnostics, we overlap the empirical density function of each covariate as well as that of the control turbine power, associated with the treated group and the matched subset of the non-treated group. We can visually inspect the discrepancy between the two density functions and see

TABLE 1

Numerical diagnostics. See the decrease of `SDM` *after the matching. The matching procedure indeed reduces the discrepancy between the two periods*

	V	D	A	W	I	P^{ctrl}
Unmatched	0.6685	0.0803	3.2715	0.2312	0.1382	0.8132
Matched	0.0142	0.0026	0.0589	0.0721	0.0003	0.0083

(a) Experimental data

	V	D	A	W	I	P^{ctrl}
Unmatched	0.0605	0.1647	1.6060	0.2759	0.4141	0.0798
Matched	0.0077	0.0029	0.0263	0.0158	0.0111	0.0036

(b) Mimicry data

if they are similar enough. An example is shown in Figure 5, in which we observe the well-matched distributions of covariates after the matching process. The improvements in term of distribution similarity are clearer when compared to Figure 4, which demonstrates the dissimilarity in covariate distributions of the unmatched original set.

Either the numerical or the graphical diagnostics may fail to provide credible evidence to perform an outcome analysis; for example, `SDM` increases, rather than decreases, or some non-overlapped bumps are observed in the density plots. If this happens, we adjust the thresholding coefficients α's and repeat the procedures of Section 3.1 and 3.2 until a well-matched set is obtained. It should also be noted that, if the size of S_{aft} after the matching loses too many data records, and this can happen when too small α's are applied, we suggest to enlarge the size of S_{aft} prior to the matching process, so that we can secure a sufficient amount of representative weather conditions in the matched S_{aft}.

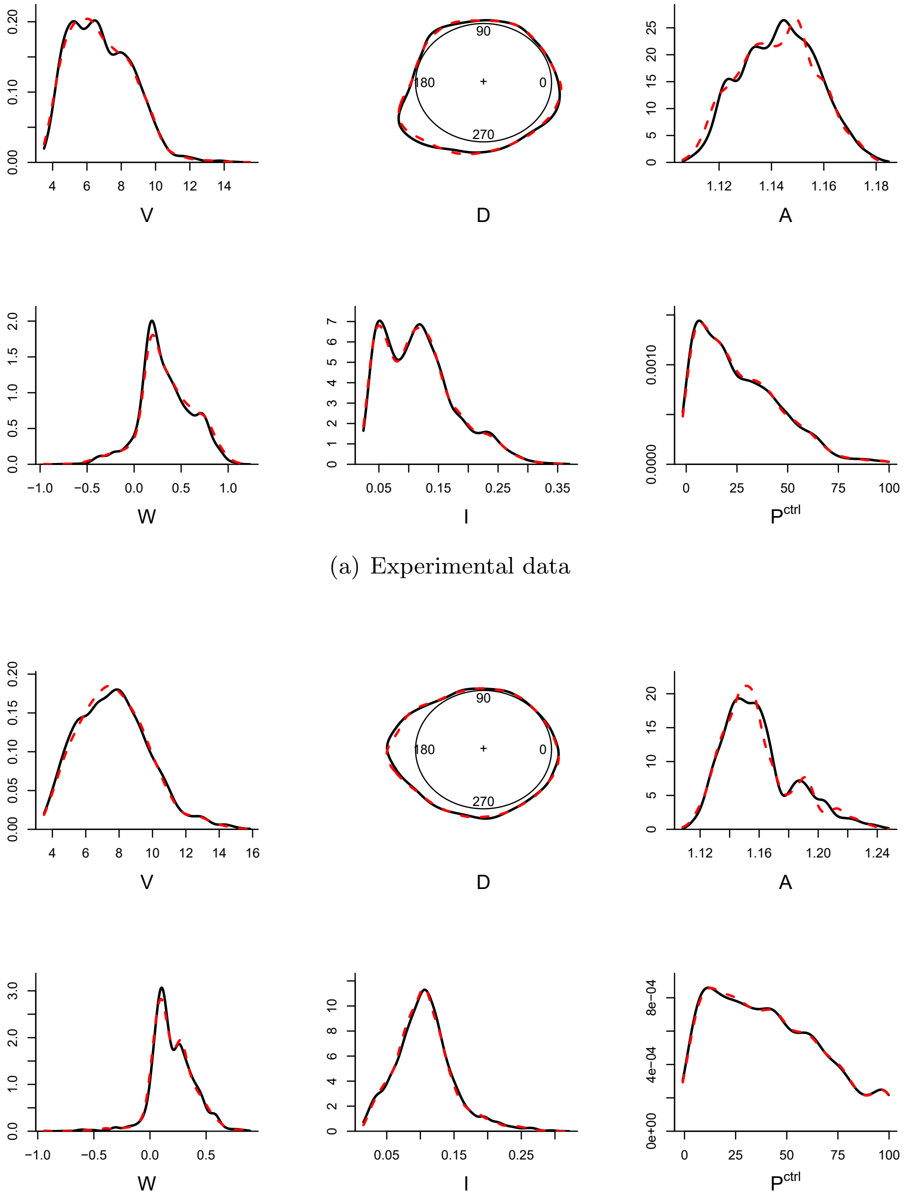
4. Outcome analysis.

This section describes the outcome analysis, Step 4 of a matching method as outlined in Section 1. It fulfills the research goal of testing the significance of the upgrade effect and quantifying its improvement in terms of extra power production under comparable environmental conditions.

4.1. *Paired t-tests.*

From the matching procedure, we have the paired data records of the two groups, (i_j, j) where $i_j \in S^*_{\text{bef}}$ and $j \in S_{\text{aft}}$. Using these paired indices, we can retrieve the paired test power outputs, $(P^{\text{test}}_{i_j}, P^{\text{test}}_j)$. The power output pair can be interpreted as repeated measurements under comparable environmental conditions, which makes the power outputs also comparable.

As such, we apply a t-test to analyze the difference of the two paired test



(a) Experimental data

(b) Mimicry data

FIG 5. *Overlapped density functions of matched covariates as well as that of power output of control turbine; solid line = before upgrade (non-treated), dashed line = after upgrade (treated). Compare this figure to Figure 4 and notice the improvement in agreement between the pairs of density plots.*

TABLE 2

Outcome analysis. The results of paired t-tests and upgrade quantification

t-stat	p-value	UPG
3.015	0.003	1.13%

(a) Experimental data

t-stat	p-value	UPG
7.447	< 0.0001	3.16%

(b) Mimicry data

outcomes, $D_j = P_j^{\text{test}} - P_{i_j}^{\text{test}}$. The assumption of independence is met; this will be reviewed in Section 6.2. It tests the null hypothesis that the expected mean of the difference is zero, that is $H_0 : E(\overline{D}) = 0$, where $\overline{D}$ is the sample mean of $\{D_j : j \in S_{\text{aft}}\}$. Accordingly, the test statistic t is

$$t := \frac{\overline{D}}{s/\sqrt{n}},$$

where s and n are the sample standard deviation and the sample size of $\{D_j : j \in S_{\text{aft}}\}$, respectively. If the test concludes a significant positive mean difference, the upgrade on the test turbine is then concluded as effective.

In Table 2, the first and second cells show the results from a paired t-test. In both datasets, the tests show a significant upgrade effect at the 0.05 level.

4.2. *Quantification.* Reporting a percentage value representing the relative increase in power production is a typical way to quantify an improvement of a turbine's performance after an upgrade. As such, we quantify the upgrade effect (UPG) in percentage terms by computing

$$\texttt{UPG} := \frac{\sum_{j \in S_{\text{aft}}} (P_j^{\text{test}} - P_{i_j}^{\text{test}})}{\sum_{j \in S_{\text{aft}}} P_{i_j}^{\text{test}}} \times 100,$$

where $i_j \in S_{\text{bef}}^*$ is the counterpart of $j \in S_{\text{aft}}$.

The quantification results are shown in the third cell of Table 2. Recall that we have increased the test turbine power in the mimicry pair by 5% for wind speed 9 m/s and above, which translates to a 3.11% increase for the whole wind spectrum. Our quantification shows an improvement of 3.16% overall, which appears to present a fair agreement with the simulated amount. If the quantification amount is to be trusted, the vortex generator installation enables a turbine to produce 1.13% more wind power than without the upgrade.

4.3. *Mean Comparison.* In Figure 6(a), we present the boxplot of P^{test} data for the both datasets under the unmatched conditions (i.e., the original data) and the matched conditions (i.e., the matched subset of the original data). We noticed that the unmatched data of the experimental set show a higher mean power before the upgrade than after. This mean power pattern is, however, reversed on the matched data, as expected. The interpretation of the mean power pattern of the unmatched data is obvious; the difference in the environmental covariates causes the wind turbine to produce more wind power in the period before the upgrade, so the upgrade effect is overwhelmed and not detectable. Even though the unmatched data seemingly shows an improvement in power production like the mimicry data in Figure 6(b), the imbalanced profile of weather conditions should be noticed, and so the matching is required to stabilize their discrepancy. This analysis demonstrates the benefit of executing this matching procedure before comparing the test power outputs and quantifying its net effect.

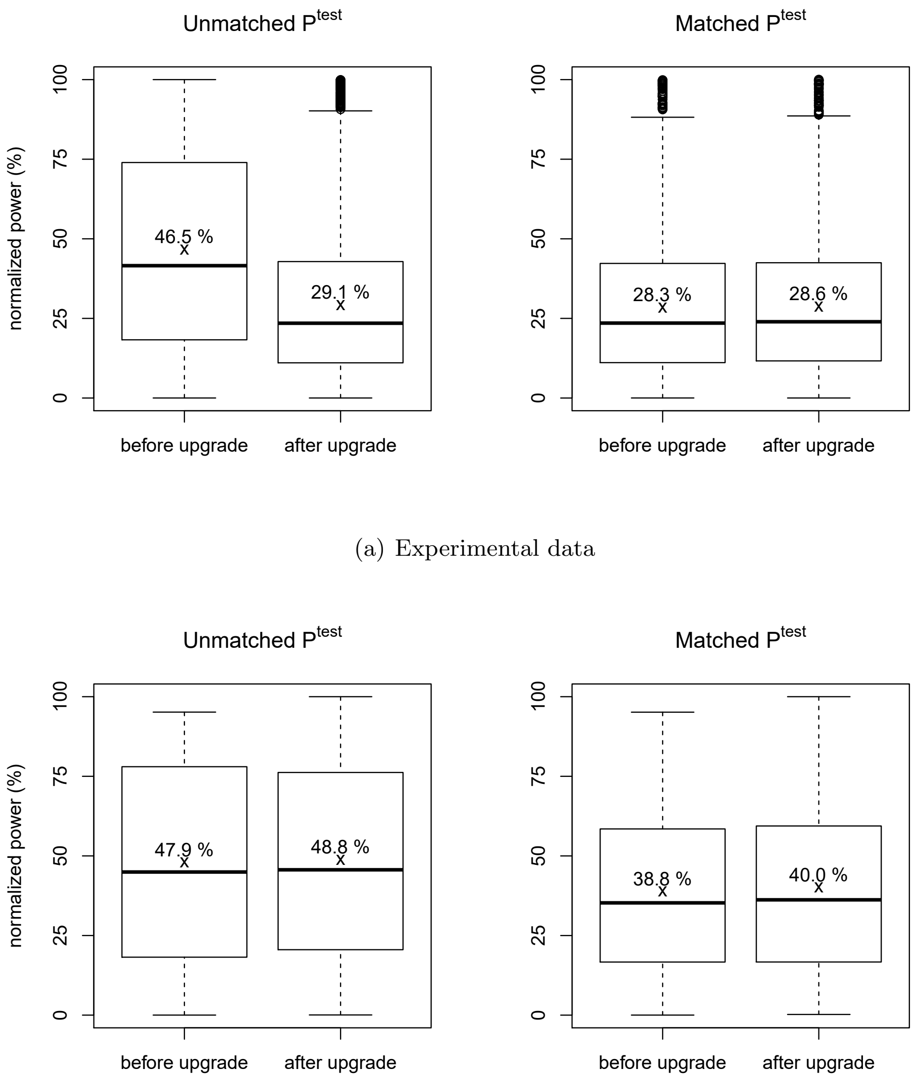
5. Sensitivity analysis. Recall that the mimicry pair is analyzed for the purpose of getting a sense of how well a proposed method can estimate a power production change, owing to a turbine upgrade. While only the 5% simulated improvement is used when illustrating the methodology in Section 3 and 4, this section re-performs the matching on various degrees of improvement. There are two reasons for this practice: (a) to see how sensitive the proposed method is in terms of estimating the power production change when the change magnitude varies (in Section 5.1), and (b) to compare the proposed matching method to the kernel plus method proposed by Lee et al. (2015b) (in Section 5.2).

5.1. *Sensitivity of estimating changes.* Considering how the mimicry pair is created, it is unreasonable to use the nominal power increase rate, denoted by r, to represent the power change magnitude over the entire spectrum of wind power. This is because the nominal power increase rate is applied only to the partial range of wind power corresponding to wind speed higher than 9 m/s. Therefore, when it comes to verifying the estimation quality in the mimicry case, we should compute the effective power increase rate, denoted by r', such as

$$r' := \frac{\sum_{j \in S_{\text{aft}}} P_j^{\text{test}} \{1 + r \cdot \mathcal{I}(V_j^{\text{test}} > 9)\} - \sum_{j \in S_{\text{aft}}} P_j^{\text{test}}}{\sum_{j \in S_{\text{aft}}} P_j^{\text{test}}},$$

where $\mathcal{I}$ is an indicator function.

As shown in Table 3, as r changes from 2% to 9%, r' changes from 1.25% to 5.6%. This range of the power improvements is considered practical for the detection purpose. If an improvement is smaller than 1%, it is going to be considerably hard for detection, and given the amount of noises in



(a) Experimental data

(b) Mimicry data

Fig 6. *Boxplots of the normalized test power values;* x *points, referred to by the label in percentage above it, are the mean of the respective normalized P^{test}. The upgrade effect is revealed in the matched test powers while confounded in the unmatched test power.*

Table 3

r = nominal power improvement rate; r' = effective power improvement rate; `UPG` *and* `DIFF*` *estimate r' through the matching method and the kernel plus method, respectively.*

r	2%	3%	4%	5%	6%	7%	8%	9%
r'	1.25%	1.87%	2.49%	3.11%	3.74%	4.36%	4.98%	5.60%
`UPG`	1.74%	2.21%	2.68%	3.16%	3.63%	4.11%	4.58%	5.05%
`UPG`/r'	1.4	1.2	1.1	1.0	1.0	0.9	0.9	0.9
`DIFF*`	1.97%	2.56%	3.15%	3.73%	4.30%	4.86%	5.42%	5.97%
`DIFF*`/r'	1.6	1.4	1.3	1.2	1.1	1.1	1.1	1.1

wind and power measurements, no known method can do an adequate job. On the other hand, when an improvement is greater than 6%, it becomes a bit unrealistic due to technology limitations, and if indeed so, the detection becomes easier – it is possible that even the standard IEC binning method can detect this level of change. That is why we choose this specific range to test the sensitivity of our method.

The middle two rows in Table 3 compare `UPG` to r'. We notice that `UPG` considerably overestimates r' when r' is small (smaller than 2%); the overestimation is as much as 40% for the smallest change at 1.25%. But the estimation quality of `UPG` gets stabilized as r' increases. In fact, for the last six cases, the differences between `UPG` and r' are within 10%. This result reflects the reality that the smaller degree of turbine upgrade is indeed difficult to estimate and demonstrates the merit of the proposed matching method.

5.2. *Comparison between the matching method and the kernel plus method.* The best benchmark method for upgrade quantification is the kernel plus method presented in Lee et al. (2015b). In this section, we compare the covariate matching method with the kernel plus method.

The metric quantifying a turbine's improvement used by Lee et al. (2015b) is labeled as `DIFF`, which indicates a percentage value measuring the power production difference before and after the turbine upgrade. Although `DIFF` has a similar concept to `UPG` in this paper, there is a subtle difference that needs to be addressed. In Lee et al. (2015b), `DIFF` values are computed for the test and control turbine separately, which are denoted by $\texttt{DIFF}_{\text{test}}$ and $\texttt{DIFF}_{\text{ctrl}}$, respectively. However, `UPG` uses the control turbine's record as a baseline reference during the matching process, so deals solely with and represents the net effect. For that reason, the metric from the kernel plus method, to be fairly compared with `UPG`, should be $\texttt{DIFF*} := \texttt{DIFF}_{\text{test}}$ - $\texttt{DIFF}_{\text{ctrl}}$, which also adjusts the test turbine outcomes using the control turbine as a baseline.

This adjusted metric `DIFF*` is then estimated for each r and compared to

r' in the last two rows of Table 3. As we notice here, the kernel plus method also considerably overestimates the small r' values and does better as r' gets bigger. The degree of overestimation of $\texttt{DIFF}^*$ is severer than that of $\texttt{UPG}$; while $\texttt{DIFF}^*/r'$s have 10% or more values over all of r values, $\texttt{UPG}/r'$s are mostly within 10% and even make almost correct estimations at $r = 5\%$ and 6%. Therefore, the covariate matching method outperforms the kernel plus method for the practical range of improvement rate, from $r = 2\%$ to 9%.

If applied to the experimental turbine pair, our analysis in Section 4.2 shows $\texttt{UPG} = 1.13\%$. On the other hand, $\texttt{DIFF}^*$ from the kernel plus method is 1.48%. This result is anticipated, in that the kernel plus method tends to overestimate a little more, and both methods are in fact less accurate when estimating a small improvement such as 1% or less.

Please note that $\texttt{DIFF}^*$ values reported here are different from those reported in Lee et al. (2015b). This discrepancy is due to the different use of data; while Lee et al. (2015b) use 2-week-after-upgrade worth of data in their analysis, we use in this study 7-week-after-upgrade worth of data for the mimicry turbine pair and 5-week-after-upgrade worth of data for the experimental pair, as our covariate matching requires a longer duration to ensure a sufficient amount of data.

6. Remarks.

This section presents further discussion of a few issues arising in our research undertaking. Section 6.1 reviews in more details about the priority order and the interaction effect of the environmental covariates as well as how the right order can benefit the analyses. Section 6.2 discusses the issue of replacement while matching data records and affirms that the independence assumption of a t-test is approximately satisfied.

6.1. *The priority order and interaction of covariates.*

The priority order of the environmental covariates used in the hierarchical subgrouping procedure in Section 3.1 is as the following: wind speed, wind direction, air density, wind shear and turbulence intensity. The importance of wind speed V is obvious and it is universally agreed to be the most important factor affecting wind power production. Wind direction D also matters a great deal even though wind turbines have a yaw control mechanism that is supposedly to track wind direction and point the turbine towards the direction from which the wind blows. Nonetheless, a score of studies showed that this tracking is not perfect, and consequently, including wind direction as one covariate can significantly reduce the prediction error of wind power [Lee et al. (2015a); Jeon and Taylor (2012); Wan et al. (2010)].

The effects of the next tier of factors, namely air density A, wind shear W and turbulence intensity I, come more in the form of interacting with

TABLE 4

Numerical diagnostics when matching with a reversed priority order, P^{ctrl}, I, W, A, D, V; notice less decreased `SDMs` *of D, A, W and P^{ctrl} than those of Table 1 (b), which implies that a poorly defined order may lead to an unsatisfactory quality of matching.*

	V	D	A	W	I	P^{ctrl}
Unmatched	0.0605	0.1647	1.6060	0.2759	0.4141	0.0798
Matched	0.0022	0.0036	0.0377	0.0208	0.0055	0.0085

the two main effects, wind speed and wind direction. Lee et al. (2015a) illustrated, in Figure 4 of their paper, the existence of interaction effects between these second-tier factors and the wind speed/direction.

We believe the nested structure of our hierarchical subgrouping helps handle the priority of the main and interacting covariates. The variance-covariance matrix in the Mahalanobis distance (Section 3.3) also captures the interaction effects through the covariance terms and incorporates them in the calculation of the dissimilarity measure.

If a priority order is poorly defined, the quality of matching may not be as satisfactory as compared to a well-defined order. To show some numerical evidence of this argument, we conducted the matching on the mimicry set with a reversed order, $P^{\text{ctrl}}, I, W, A, D, V$; their numerical diagnostics are shown in Table 4. Comparing this result to Table 1 (b), the `SDMs` of D, A, W and P^{ctrl} with the reversed order are greater than those with the proper order. It should be noted that the thresholding degrees in Table 4 are the same as those in Table 1 for a fair comparison. However, as long as those `SDMs` are acceptable to perform an outcome analysis, the significance and quantification of turbine improvement does not change dramatically. The analysis using the reversed order leads to a `UPG` $= 3.33\%$ with p-value $< .0001$, which is similar to that with the well-defined order (`UPG` $= 3.16\%$, while true value $= 3.11\%$).

Still, although an outcome analysis appears to show a certain degree of robustness under acceptable `SDMs`, one might as well make use of the priority information, if known, since it helps find the acceptable matched set much more efficiently. If a priority order of covariates is unknown, it is recommended to perform some statistical analysis using, for example, random forests [Breiman (2001)], which can measure the importance of covariates, before applying the matching method.

6.2. *Matching with replacement and assumption of independence.* Recall from Section 3.4 that we allow replacement when carrying out the matching procedure. Because of this, a data record in the non-treated dataset S_{bef} could possibly be paired with two or more data records in the treated dataset S_{aft}.

A potential problem of allowing replacement is that the replication of the same data records may cause a violation of independence of outcome variables. In order to settle this issue, information about frequency weights, such as the relative number of replications, may need to be taken into account [Stuart (2010)].

In our application, however, replacement does not seem to cause too much of a problem, for the following reasons: (a) such replication happens rather rarely by starting with the much larger set of non-treated period than the treated; (b) we in fact analyze the differences between the treated period (not replicated, so independent) and the non-treated period (possibly replicated, so dependent), and taking differences further reduces the dependence caused by replication.

7. Summary. We are interested in statistical inference about the upgrade effect on wind turbine performance. It is a challenging issue because the upgrade effect on wind power production could be biased and confounded by unmanageable environmental conditions. Some of these conditions are measured on a wind farm, while others are unknown or not measured. We propose a covariate matching method, allowing for a fair and direct comparison of power outcomes without establishing power curve models.

Compared to the current studies on wind power analysis, our matching method entertains several advantages: (a) it does not compare the estimated power outputs from the fitted power curve models, but compares the observed power outputs directly; (b) by using the control turbine power output as a benchmark, our method takes into account both measured and unmeasured environmental conditions; (c) when future technology innovations allow additional environmental covariates to be measured, their inclusion in our matching method is straightforward and it does not complicate the subsequent analysis steps. By testing on both experimental data and simulated data, the proposed matching method appears to be sensitive to detecting small to moderate changes resulting from an upgrade on a wind turbine.

Acknowledgement. We thank the editor, associate editor, and reviewers for their detailed, valuable comments which substantially helped us to improve the quality of the paper.

References.

Thomas Ackermann and Lennart Söder. Wind power in power systems: an introduction. *Wind Power in Power Systems*, pages 25–51, 2005.

A. Albers. Relative and integral wind turbine power performance evaluation. In *Proceedings of the 2012 European Wind Energy Conference & Exhibition*, pages 22–25, 2012.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Eunshin Byon, Lewis Ntaimo, Chanan Singh, and Yu Ding. Wind energy facility reliability and maintenance. In P. M. Pardalos, S. Rebennack, M. V. F. Pereira, N. A. Iliadis, and V. Pappu, editors, *Handbook of Wind Power Systems: Optimization, Modeling, Simulation and Economic Aspects*, pages 639 – 672. Springer-Verlag, Berlin, 2013.

Luca Delle Monache, F Anthony Eckel, Daran L Rife, Badrinath Nagarajan, and Keith Searight. Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10):3498–3516, 2013.

DOE. Windexchange: US installed wind capacity 2015. Technical report, U.S. Department of Energy's Energy Efficiency & Renewable Energy Website, 2015. Available at http://apps2.eere.energy.gov/wind/windexchange/wind_installed_capacity.asp.

Stefano M. Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

IEC. Wind turbines – part 12-1: Power performance measurements of electricity producing wind turbines; iec tc/sc 88. Technical report, International Electrotechnical Commission 61400-12-1:2005, 2005.

S. Rao Jammalamadaka and Ashis Sengupta. *Topics in Circular Statistics*, volume 5. World Scientific, 2001.

Jooyoung Jeon and James W. Taylor. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107(497): 66–79, 2012.

Mohammed G. Khalfallah and Aboelyazied M. Koliub. Suggestions for improving wind turbines power curves. *Desalination*, 209:221–229, 2007.

Andrew Kusiak, Haiyang Zheng, and Zhe Song. Wind farm power prediction: a data-mining approach. *Wind Energy*, 12(3):275–293, 2009.

Giwhyun Lee, Yu Ding, Marc G. Genton, and Le Xie. Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *Journal of the American Statistical Association*, 110(509):56–67, 2015a.

Giwhyun Lee, Yu Ding, Le Xie, and Marc G. Genton. A kernel plus method for quantifying wind turbine performance upgrades. *Wind Energy*, 18(7):1207–1219, 2015b.

Prasanta C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

Torben S. Nielsen, Henrik A. Nielsen, and Henrik Madsen. Prediction of wind power using time-varying coefficient functions. In *Proceedings of the XV IFAC World Congress on Automatic Control*, Barcelona, Spain, 2002.

Lisa Ann Osadciw, Yanjun Yan, Xiang Ye, Glen Benson, and Eric White. Wind turbine diagnostics based on power curve using particle swarm optimization. In Lingfeng Wang, Chanan Singh, and Andrew Kusiak, editors, *Wind Power Systems (Green Energy and Technology)*, pages 151–165. Springer-Verlag, Berlin, 2010.

Stig Øye. The effect of vortex generators on the performance of the ELKRAFT 1000 kW turbine. In *Aerodynamics of Wind Turbines: 9th IEA Symposium*, pages 9–14, Stockholm, Sweden, 1995. ISSN:0590-8809.

Pierre Pinson, Henrik A. Nielsen, Henrik Madsen, and Torben S. Nielsen. Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1):59–71, 2008.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Donald B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1): 159–183, 1973.

Donald B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188, 2001.

Ismael Sanchez. Short-term prediction of wind energy production. *International Journal of Forecasting*, 22(1):43–56, 2006.

Elizabeth A. Stuart. Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1):1, 2010.

Onder Uluyol, Girija Parthasarathy, Wendy Foslien, and Kyusung Kim. Power curve analytic for wind turbine performance monitoring and prognostics. In *Annual Conference of the Prognostics and Health Management Society*, volume 2, Publication Control Number 049, Montreal, Canada, 2011.

Yih-Huei Wan, Erik Ela, and Kirsten Orwig. Development of an equivalent wind plant power curve. Technical Report NREL/CP-550-48146, National Renewable Energy Laboratory, 2010. Available at http://www.nrel.gov/docs/fy10osti/48146.pdf.

Lin Wang, Xinzi Tang, and Xiongwei Liu. Blade design optimisation for fixed-pitch fixed-speed wind turbines. *ISRN Renewable Energy*, Article ID 682859, 2012.

Yanjun Yan, Lisa Ann Osadciw, Glen Benson, and Eric White. Inverse data transformation for change detection in wind turbine diagnostics. In *Proceedings of the 22nd IEEE Canadian Conference on Electrical and Computer Engineering*, pages 944–949, St. John's, Newfoundland, Canada, 2009.

Zhong Zhao. Using matching to estimate treatment effects: data requirements, matching metrics, and monte carlo evidence. *Review of Economics and Statistics*, 86(1):91–107, 2004.

Department of Statistics,
Texas A&M University,
College Station, TX 77843-3143
E-mail: syeeun@stat.tamu.edu
jianhua@stat.tamu.edu

Department of Industrial and
Systems Engineering,
Texas A&M University,
College Station, TX 77843-3131
E-mail: yuding@tamu.edu