Copula Graphical Models for Wind Resource Estimation

Kalyan Veeramachaneni
CSAIL, MIT
Cambridge, MA
kalyan@csail.mit.edu

Alfredo Cuesta-Infante
Universidad Rey Juan Carlos
Madrid, Spain
alfredo.cuesta@urjc.es

Una-May O'Reilly
CSAIL, MIT
Cambridge, MA
unamay@csail.mit.edu

Abstract

We develop *multivariate copulas* for modeling multiple joint distributions of wind speeds at a wind farm site and neighboring wind source. A *n-dimensional* Gaussian copula and multiple copula graphical models enhance the quality of the prediction site distribution. The models, in comparison to multiple regression, achieve higher accuracy *and* lower cost because they require less sensing data.

1 Introduction

This paper addresses wind resource assessment: the problem of determining if there will be enough wind in the ideal speed range that will endure at a potential wind farm or "*site*", over a 20+ year timespan. The end-to-end pipeline of a resource assessment service spans from automatic site-neighbor data extraction from public, online sources (ASOS database), through site-neighbor data synchronization in preparation for generative modeling, modeling, backcast (where historical data at neighboring sites is passed through a model to obtain predictions at the site) to estimation of the industry standard Weibull distribution from the derived predictions. Our focus in this paper is on the single most critical factor in assessment: achieving the most accurate *backcast* while incurring *minimal* financial expense. This implies integrating geographically proximal public wind data sources and building models that better represent the data (accuracy) while concurrently reducing the duration of anemometer sensing during the assessment period (expense).

The primary contribution of this paper is the use of *multivariate copulas* for modeling multiple joint distributions of wind speeds at the site and a publicly available neighboring wind source. We construct a *n-dimensional* Gaussian copula and multiple copula graphical models to enhance the quality of the prediction site distribution. This modeling step is embedded within a widespread methodology called Measure-Corelate-Predict (MCP) [Gross and Phelan, 2006; Bass *et al.*, 2000; Bailey *et al.*, 1997; Lackner *et al.*, 2008].

For demonstration we use speed and direction data from an actual site in the state of Massachusetts where we have assembled data from anemometer sensing carried over for a period of two years. We compare our models with multiple regression methods, where it achieves higher accuracy with less sensing data – sometimes with only 3 months. The industry standard method, multiple regression, achieves a reasonable accuracy with 8 months of data, an industry standard period. Thus we achieve better accuracy at a lower cost.

We proceed by describing MCP while introducing notation in Section 2. Section 3 describes the real wind resource estimation scenario and the dataset we utilized throughout this paper to demonstrate our methods. Section 4 describes the copula modeling. Section 5 is the demonstration. We intentionally reference related work throughout the paper, in context of discussion.

2 Measure-Corelate-Predict (MCP)

We consider wind resource estimation derived by a methodology known as Measure-Correlate-Predict or MCP. In terms of notation, the wind at a particular location is characterized by speed denoted by x and direction θ. Wind speed is measured by anemometers and wind direction is measured by wind vanes. The 360^o direction is split into multiple bins with a lower limit (θ_l) and upper limit (θ_u). We give an index value of $J = 1 \ldots j$ for the directional bin. We represent the wind speed measurement at the test site (where wind resource needs to be estimated) with y and the other sites (for whom the long term wind resource is available) as x and index these other sites with $M = 1 \ldots m$.

The three steps of MCP are:

MEASURE Short term sensing measurements on the site are collected. This is denoted by $Y = \{y_{t_k} \ldots y_{t_n}\}$. Measurements can be collected using anemometers on the site, a newly-constructed meteorological tower, or even remote sensing technologies such as sonar or lidar. Different measurement techniques incur different costs that dictate their feasibility for different projects. Measurements from nearby sites for the same period are gathered. These sites, called *historical sites*, have additional data for the past 10–20 years. These are denoted by $X = \{x^{1\ldots m}_{t_k \ldots t_n}\}$ where each $x^i_{t_k \ldots t_n}$ corresponds to data from one historical site and m denotes the total number of historical sites. Historical data that is not simultaneous in time to the site observations used in modeling will be used in the PREDICT step.

CORRELATE A single directional model is first built correlating the wind directions observed at the site with

simultaneous historical site wind directions. Next, for each directional interval, called a (directional) bin, of a 360° radius, a model is built correlating the wind speeds at the site with simultaneous speeds at the historical sites, i.e. $Y_{t_i} = f_{\theta_j}(x_{t_i}^{1\cdots m})$ where $k \leq i \leq n$. The data available from the site at this stage is expected to be sparse and noisy.

PREDICT To obtain an accurate estimation of long term wind conditions at the site, we first divide the data from the historic sites (which is not simultaneous in time to the site observations used in modeling) into subsets that correspond to a directional bin. Prediction of the long term site conditions follows two steps:

A : We use the model we developed for that direction f_{θ_j} and the data from the historic sites corresponding to this direction $x_{t_1\cdots t_k-1}^{1\cdots m}|\theta_j$ to predict what the wind speed $\mathbf{Y_p} = y_{t_1\cdots t_k-1}$ at the site would be. For a new observation $\mathbf{x}$ we have to predict y. For this we form the conditional first by

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}, y)}{\int_y P(\mathbf{x}, y)dy}. \quad (1)$$

Our predicted $\hat{y}$ maximizes this conditional probability

$$\hat{y} = \arg\max_{y \in Y} P(y|\mathbf{x}). \quad (2)$$

Note that the term in the denominator of eq.(1) remains constant, hence for the purposes of finding the optimum we can ignore its evaluation. We simply evaluate this conditional for the entire range of Y in discrete steps and pick the value of $y \in Y$ that maximizes the conditional.

B : With the predictions $\mathbf{Y_p}$, from **A** above, we estimate parameters for a Weibull distribution. This distribution is our answer to the wind resource assessment problem. We generate a distribution for each directional bin.

The goal is to generate a predicted long term wind speed distribution in each direction which will be as close as possible to the real (as yet unexperienced) distribution. The result from MCP, i.e. the statistical distribution in each bin, is then used to estimate the energy which can be expected from a wind turbine, given the power curve supplied by its manufacturer. This calculation can be extended over an entire farm if wake interactions among the turbines are taken into account. See [Wagner *et al.*, 2011] for more details. Note that distribution not only captures the mean, but also variance in this speed. This is critical for assessment of long term wind resource and the long term energy estimate.

A variety of methods are developed in [Rogers *et al.*, 2005] to evaluate the accuracy of the predicted wind speed distribution. One method measures the accuracy in terms of ratios between true and actual parameters of the Weibull distribution. That is, true shape versus estimated shape and true scale versus estimated scale. To completely capture any possible inaccuracy in the predicted distribution, we measure a symmetric Kullback-Leibler distance. It is important to note that this measure is different than the mean-squared error or mean-absolute error which measure the accuracy in terms of difference between each predicted value and the true observation. Methods that minimize these errors would not necessarily accurately express how close the approximation is to the true distribution. As a measure of predictive accuracy we compare the final estimated Weibull distribution to the ground truth distribution using Kullback-Leibler (KL) divergence. The lower this value, the more accurate the prediction:

$$D_{(Y||\hat{Y})} = KL(P_Y(y)||P_{\hat{Y}}(\hat{y})). \quad (3)$$

KL divergence derives the distance between two probability distributions:

$$D_{\mathrm{KL}}(P_Y(y)\|P_{\hat{Y}}(\hat{y})) = \sum_i P_Y(y=i) \ln \frac{P_Y(y=i)}{P_{\hat{Y}}(\hat{y}=i)} \quad (4)$$

For baseline comparison, we also developed a linear regression model which is used quite extensively in wind resource assessment [Bass *et al.*, 2000; Rogers *et al.*, 2005].

We now proceed to describe the our machine learning approaches for wind resource assessment.

3 A real world scenario and dataset

To evaluate and compare our different algorithms, we acquired wind data collected using anemometers from the rooftop of Museum of Science in Boston where a wind vane is also installed. These anemometers are inexpensive and consequently noisy. The museum is located amongst buildings, a river and is close to a harbor as shown in Figure 1. This provides us with a site that is topographically challenging. At this location we have approximately 2 years worth of data collected at a frequency of 1 sample/second with 10 minute averages stored in a separate database. To derive the wind resource assessment we train using data from the first year. This data is split into three datasets we call D_3, D_6 and D_8. The split D_3 has data for 3 months. The split D_6 has 3 additional months for a total of 6 and D_8 has yet 2 more months for a total of 8. We divide each dataset and the second year's dataset (to serve as our test data, i.e. ground truth) further into 12 directional bins of equal sizes starting at compass point North (0°).

We use airport wind data from the public ASOS (Automated Surface Observing System) database for sources of neighboring-site data. This data is regularly accessed by the wind industry for correlation purposes. The airports' locations are shown in Figure 1 (right).

4 Multivariate copulas

Previous modeling techniques assume a Gaussian distribution for wind speed and direction variables and a Gaussian joint distribution. It is arguable however that Gaussian distributions do not accurately represent the wind speed distributions. In fact, conventionally a univariate Weibull distribution [Burton *et al.*, 2001] is used to parametrically describe wind sensor measurements. A Weibull distribution is likely also chosen for its flexibility because it can express any one of multiple distributions, including Rayleigh or Gaussian.

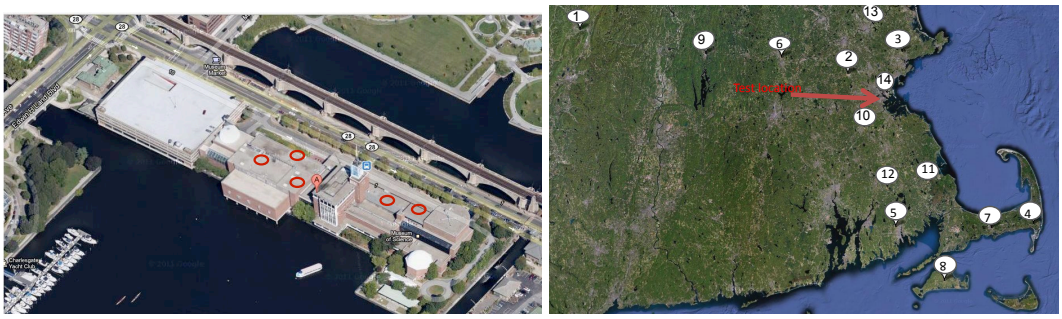


Figure 1: Left: Red circles show location of anenometers on rooftop of Museum of Science, Boston. Right: Neighboring-site data from fourteen airports (marked with circles) is used in the MCP correlation step.

To the best of our knowledge, however, joint density functions for non-Gaussian distributions have not been estimated for wind resource assessment. In this paper, to build a multivariate model from marginal distributions which are not all Gaussian, we exploit *copula* functions. A *copula* framework provides a means of inference after modeling a multivariate joint distribution from training data.

Because copula estimation is less well known, we now briefly review copula theory. We will then describe how we construct the individual parametric distributions which are components of a copula and then, how we couple them to form a multivariate density function. Finally, we present our approach to predict the value of y given $x_{1\ldots m}$.

A copula function $C(u_1, \ldots u_{m+1}; \theta)$ with parameter θ represents a joint distribution function for multiple *uniform* random variables $U_1 \ldots U_{m+1}$ such that

$$C(u_1, \ldots u_{m+1}; \theta) = F(U_1 \leq u_1, \ldots U_{m+1} \leq u_{m+1}). \quad (5)$$

Let $U_1 \ldots U_m$ represent the cumulative distribution functions (CDF) for variables $x_1, \ldots x_m$ and U_{m+1} represent the CDF for y. Hence the *copula* represents the joint distribution function of $C(F(x_1) \ldots F(x_m), F(y))$, where $U_i = F(x_i)$. According to Sklar's theorem any *copula* function taking marginal distributions $F(x_i)$ as its arguments, defines a valid joint distribution with marginals $F(x_i)$. Thus we are able to construct the joint distribution function for $x_1 \ldots x_m, y$:

$$F(x_1 \ldots x_m, y) = C(F(x_1) \ldots F(x_m), F(y); \theta) \quad (6)$$

The joint probability density function (PDF) is obtained by taking the $m + 1^{th}$ order derivative of eqn. (6)

$$\begin{aligned} f(x_1 \ldots x_m, y) = & \\ & \frac{\partial^{m+1}}{\partial x_1 \ldots \partial x_m \partial y} C(F(x_1) \ldots F(x_m), F(y); \theta) \\ & = \prod_{i=1}^{m} f(x_i) f(y) c(F(x_1) \ldots F(x_m), F(y)) \end{aligned} \quad (7)$$

where $c(.)$ is the *copula* density. Thus the joint density function is a weighted version of independent density functions, where the weight is derived via *copula* density.

4.1 Gaussian copula

First we consider a multivariate Gaussian copula to form a statistical model for our variables given by

$$C_G(\Sigma) = F_G(F^{-1}(u_1) \ldots F^{-1}(u_m), F^{-1}(u_y), \Sigma) \quad (8)$$

where F_G is the CDF of multivariate normal with zero mean vector and Σ as covariance and F^{-1} is the inverse of the standard normal.

Estimation of parameters: There are two sets of parameters to estimate. The first set of parameters for the multivariate Gaussian copula is Σ. The second set, denoted by $\Psi = \{\psi, \psi_y\}$ are the parameters for the marginals of $\mathbf{x}, y$. Given N *i.i.d* observations of the variables $\mathbf{x}, y$, the log-likelihood function is:

$$\begin{aligned} L(\mathbf{x}, y; \Sigma, \Psi) &= \sum_{l=1}^{N} \log f(\mathbf{x}_l, y_l | \Sigma, \Psi) \\ &= \sum_{l=1}^{N} log \left\{ \left(\prod_{i=1}^{m} f(x_{il}; \psi_i) f(y_l; \psi_y) \right) \right. \\ &\qquad \left. c(F(x_1) \ldots F(x_m), F(y); \Sigma) \right\} \end{aligned} \quad (9)$$

Parameters Ψ are estimated via[Iyengar, 2011]

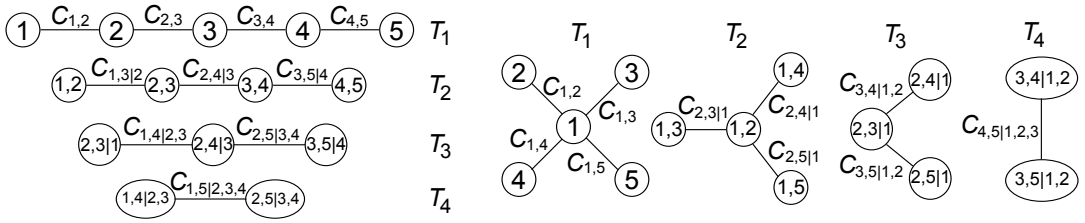
$$\begin{aligned} \hat{\Psi} = \arg\max_{\Psi \in \psi} \sum_{l=1}^{N} log \left\{ \left(\prod_{i=1}^{m} f(x_{il}; \psi_i) f(y_l; \psi_y) \right) \right. \\ \left. c(F(x_1) \ldots F(x_m), F(y); \Sigma) \right\} \end{aligned} \quad (10)$$

A variety of algorithms are available in literature to estimate the MLE in eq. (10). We refer users to [Iyengar, 2011] for a thorough discussion of estimation methods. For more details about the *copula* theory readers are referred to [Nelsen, 2006].

4.2 Vine models

Before giving details of how to construct a vine, we present three examples of how to derive the factorization of a multivariate probability distribution in terms of bivariate copulas. Consider first only two variables x_1 and x_2. The joint density, $f_{12}(x_1, x_2)$, can be factorized in two ways. First, using the chain rule:

$$= f_1(x_1) f_{2|1}(x_2|x_1)$$



(a) $f(x_1,\ldots,x_5) = (c_{1,2}c_{2,3}c_{3,4}c_{4,5})(c_{1,3|2}c_{2,4|3}c_{3,5|4})(c_{1,4|2,3}c_{2,5|3,4})(c_{1,5|2,3,4}) \prod_{i=1}^{5} f(x_i)$

(b) $f(x_1,\ldots,x_5) = (c_{1,2}c_{2,3}c_{3,4}c_{4,5})(c_{1,3|2}c_{2,4|3}c_{3,5|4})(c_{1,4|2,3}c_{2,5|3,4})(c_{4,5|1,2,3}) \prod_{i=1}^{5} f(x_i)$

Figure 2: Three examples of 5-dimensional copula constructions: (a) a D-vine; (b) a C-vine with the condition on T_1 that u_1 connects with every u_j for $j \neq 1$; The n-copula density is written below each graphical model.

and second using copulas due to Sklar's theorem (7):

$$= f_1(x_1) f_2(x_2) c_{12}(F_1(x_1), F_2(x_2)).$$

From these two, we can derive that

$$f_{2|1}(x_2|x_1) = f_2(x_2) \cdot c_{12}(F_1(x_1), F_2(x_2))$$

by canceling out the first term $f_1(x_1)$ in both. We can generalize this as

$$f_{p|q}(x_p|x_q) = f_p(x_p) c_{pq}(F_p(x_q), F_q(x_q)).$$

For the remaining part of this section variables will be omitted in probability densities and distributions since their subscripts give all the information. Thus, the last expression can be rewritten as

$$f_{p|q} = f_p c_{pq}. \tag{11}$$

Next, let us consider three variables x_1, x_2 and x_3. Like before, there are several factorizations of the joint density, f_{123}, due to the chain rule. For instance, it could be factorized as: $f_1 f_{23|1}$ or

$$= f_1 f_{2|1} f_{3|21}. \tag{12}$$

Expanding $f_{23|1}$ using Sklar's theorem we get

$$f_1 f_{23|1} = f_1 f_{2|1} f_{3|1} c_{23|1}; \tag{13}$$

where $c_{23|1}$ denotes the copula density $c(F_2(x_2|x_1), F_3(x_3|x_1))$.

According to (11) $f_{3|1} = f_3 c_{31}$, hence by replacing $f_{(3|1)}$ in (13) we get:

$$f_1 f_{23|1} = f_1 f_{2|1} f_3 c_{31} c_{23|1}. \tag{14}$$

Since (14) and (12) are both factorizations of the joint f_{123}, by equating (12) and right hand side of (14) and canceling terms on both sides we get:

$$f_{3|21} = f_3 c_{31} c_{23|1}. \tag{15}$$

It is convenient to generalize this last expression, (15), as:

$$f_{p|qr} = f_p c_{pr} c_{pq|r}. \tag{16}$$

Finally consider a last example with four variables and the two following factorizations:

$$f_{1234} = f_{12} f_{3|12} f_{4|123} = f_{12} f_{34|12}.$$

Sklar's theorem applied to $f_{34|12}$, using (16), and straightforward manipulation leads to

$$f_{4|123} = f_{34|12}/f_{3|12} = c_{34|12} f_{4|21} = f_4 c_{41} c_{42|1} c_{43|12}.$$

Again, the last expression can be generalised as:

$$f_{p|qrs} = f_p c_{pq} c_{pr|q} c_{ps|qr}. \tag{17}$$

Hence, using (11), (16) and (17), any factorization of the 4 variable joint density can be expressed in terms of bivariate copulas. For instance:

$$f_{1234} = f_1 f_{2|1} f_{3|21} f_{4|321} = (f_1 f_2 f_3 f_4)(c_{12} c_{23} c_{34})(c_{13|2} c_{24|3})(c_{14|23}). \tag{18}$$

A vine is a graphical representation of one factorization of the n-variate probability distribution in terms of $n(n-1)/2$ bivariate copulas by means of the chain rule. It consists of a sequence of levels and as many levels as variables. Each level consists of a tree (no isolated nodes and no loops) satisfying that if it has n nodes there must be $n-1$ edges. Each node in tree T_1 (level 1) is a variable and edges are couplings of variables constructed with bivariate copulas. Each node in tree T_2 (level 2) is a coupling in T_1, expressed by the copula of the variables; while edges are couplings between two vertices that must have one variable in common, becoming a conditioning variable in the bivariate copula. Thus, every level has one node less than the former. Once all the trees are drawn, the factorization is the product of all the nodes.

An example with 5 variables is given in Figure 2; showing two different vines together with their resulting joint pdf factorization.

For the sake of clarity we give the details of how to construct the panel of the left and how to derive its expression.

1. Construct T_1, the first tree, in two steps:
   (a) Draw in a row one node for each variable.
   (b) Draw the edges by linking two adjacent nodes with a copula.
2. Construct T_2 as follows:
   (a) Draw in a row one node for each edge in T_1. Write the variables of the edges in T_1 inside the nodes. For example in Figure 2 (left), the first edge in T_1 is $c_{1,2}$ which makes the first node of T_2 $1, 2$.

(b) Draw the edges by linking two adjacent nodes with a copula.
Write the bivariate copula such that the variables are those that only appear in one node or another, conditioning to the variables that are repeated in both; e.g. if nodes are $C(1,2)$ and $C(2,3)$, the conditioning variable is x_2 because only 2 appears in both.

3. Repeat until the last tree, which has only one node.
4. The factorization is the product of every node. □

Vines were initially presented in [Bedford and Cooke, 2001], [Cooke *et al.*, 2007]. A comprehensive compilation of vine methods can be found in [Kurowicka and Joe, 2011]. In a nutshell, each vertex in T_j with $j > 1$ is a coupling in T_{j-1} expressed by a bivariate copula with $j-2$ conditioning variables which are those in common in the vertices coupled at T_{j-1}. Edges in T_j are formed between 2 vertices of T_i that add another common variable. The factorization is just the product of all the edges in all the trees, times the product of all the marginals. Let $c_j(v_1, v_2|\mathbf{w})$ be the density of an edge in T_j, so that $v_1, v_2 \in \{u_1, \ldots, u_d\}$ and $v_1 \neq v_2$ are the variables of that edge, conditioned by the set $\mathbf{w} = \{w_1, \ldots, w_{j-1}\} \in \{u_1, \ldots, u_d\}$, with $w_k \neq v_1 \neq v_2$ for $1 \leq k \leq j-1$. Then, if $\mathcal{E}$ is the set of all edges in the vine, the factorization of the graphical model is:

$$f_{x_1,\ldots,x_n} = \left(\prod_{i=1}^{n} f_i(x_i)\right)\left(\prod_{j\in\mathcal{E}} c_j(v_1, v_2|\mathbf{w})\right). \qquad (19)$$

In practice, however, it is usually recommended to avoid constructing all the trees because a full vine will only represent the actual underlying joint pdf if the bivariate copulas used are rightly chosen and accurately estimated [Salinas-Gutiérrez *et al.*, 2010],[Haff *et al.*, 2010].

In this paper we use two vine constructions: *Drawable*-vines (D-vines) and *Canonical*-vines (C-vines), which are constructed according to different rules. All the models were fully obtained, in the sense that all the trees are developed, so that a study of how the depth influences the prediction can be done. The methodology in all of them is similar:

1. Transform all the variables by means of their marginals. In other words, compute
$$u_i = F_i(x_i), \quad \text{with} \quad i = 1, \ldots, 15$$
and compose the matrix $\mathbf{u} = u_1, \ldots, u_n$, where u_i are their columns.
2. For the construction of the first tree T_1, assign one node to each variable and then couple them by maximizing the measure of association considered. Different vines impose different constraints on this construction. When those are applied different trees are achieved at this level. Figure 2 shows that D-Vine and C-Vine lead to different T_1.
3. Select the copula that best fits to the pair of variables coupled by each edge in T_1. Details about how this step is carried out are given in next subsection.
4. Let $C_{ij}(u_i, u_j)$ be the copula for a given edge (u_i, u_j) in T_1. Then for every edge in T_1, compute either $v^1_{j|i} = \partial/\partial u_j C_{ij}(u_i, u_j)$ or similarly $v^1_{i|j}$, which are conditional *cdfs*. When finished with all the edges, construct the new matrix with $\mathbf{v^1}$ that has one less column $\mathbf{u}$.
5. Set $k = 2$.
6. Assign one node of T_k to each edge of T_{k-1}. The structure of T_{k-1} imposes a set of constraints on which edges of T_k are realizable. Hence the next step is to get a linked list of the *accesible nodes* for every node in T_k.
7. As in step 2, nodes of T_k are coupled maximizing the measure of association considered and satisfying the constraints impose by the kind of vine employed plus the set of constraints imposed by tree T_{k-1}.
8. Select the copula that best fit to each edge created in T_k.
9. Recompute matrix $\mathbf{v^k}$ as in step 4, but taking T_k and $\mathbf{v^{k-1}}$ instead of T_1 and $\mathbf{u}$.
10. Set $k = k + 1$ and repeat from 6 until all the trees are constructed.

The rest of this section is devoted to the selection of the copula and the particular constraints for each vine.

Copula function selection

In step 4 above, there are multiple options for choosing the copula between a pair of variables. In this paper we consider three parametric copulas [1] : Clayton, Frank and Gumbel because with these we cover a wide range of tail dependences.There is a number of ways to evaluate which one of the copulas fits better to a dataset. Two of the most popular methods are to compare the empirical density function of the copula with the theoretical one [Genest and Rivest, 1993], and to compare the upper or lower tail functions [Venter, 2001].

In this paper we employ a triple-check fitting based on the latter. For this, we first compute an empirical Copula (non-parametric). We then numerically compute upper and lower tails given this empirical Copula and calculate the area under the tails given by a_l and a_u. The upper and lower tail concentration functions are respectively defined as:

$$R(z) = [1 - 2z + C(z,z)]/(1-z)^2 \quad \text{and} \quad L(z) = C(z,z)/z^2 \qquad (20)$$

Hence, given out three candidate copulas (Clayton, Frank and Gumbel), the procedure proposed for selecting the one that best fit to a dataset of pairs $\{(u^j, v^j)\}_{j=1,2,\ldots}$, is as follows:

1. Estimate the most likely parameter θ of each copula candidate for the given dataset.
2. Construct $R(z|\theta)$. Calculate the area under the tail for each of the copula candidates.

[1] These are three popular archimedean families that can be found in many mathematical packages such as R or Matlab.

3. Compare the areas: a_u achieved using empirical copula against the ones achieved for the copula candidates. Score the outcome of the comparison from 3 downto 1, 3 being the best and 1 is the worst.
4. Proceed as in steps 2- 3 with the lower tail and function L.
5. Finally the sum of empirical upper and lower tail functions is compared against $R + L$. Scores of the three comparisons are summed and the candidate with the highest value is selected.

D-Vine Model.

In a D-Vine every node in tree T_1 has degree 2 except two nodes, with degree 1, which can be seen as the extremes. Figure 2a shows an example of D-Vine of five variables. The advantage of D-Vine is that, once T_1 is constructed, it uniquely determines the rest of the trees that compose the vine. Hence, learning the model is the task of finding the best assignment of variables to nodes in T_1. Throughout this paper, Kendall's τ is employed as measure of association to decide how to couple nodes. The procedure is then as follows.
Let $\mathbf{u}$ be the matrix of the marginal *cdf* values of the dataset $\mathbf{x}$;

1. Compute $M_\tau = [\tau_{ij}]$, with $i > j$, the matrix of Kendall's τ between every possible coupling of variables. Notice that M_τ only requires values in its upper triangular part because it is symmetric.
2. Find $\hat{\tau} = \max([\tau_{ij}])$. Let $(\texttt{left}, \texttt{right})$ be the coordinates of $\hat{\tau}$.
   Then initialize $T_1 = [\texttt{left}, \texttt{right}]$.
3. For k = 1 to m-1
   (a) Set $\texttt{left} = T_1(1)$ and $\texttt{right} = T_1(end)$
   (b) Set the $\texttt{left}^{th}$ and $\texttt{right}^{th}$ columns of M_τ to zero.
   (c) Find `leftNew`, the variable that couples with `left` with maximum Kendall's $\tau = \tau_L$. Similarly, find `rightNew`, the variable that couples with `right` with maximum Kendall's $\tau = \tau_R$.
   (d) If $\tau_L > \tau_R$, then $T_1 = [\texttt{leftNew}, T_1]$; else $T_1 = [T_1, \texttt{rightNew}]$

C-Vine Model

In a C-Vine, for every tree, one *anchor* node is connected with all the others. Figure 2b shows an example with five variables. If the anchor node of T_1 is the variable of the site of interest, then edges represent its dependence with respect to the rest of variables. The criterion for selecting the anchor node of T_k, for $k > 1$ is the following:

1. Set $i = 1$ and compute τ_{ij}, the Kendall's of the data associated to node i and node j in T_k, for every $j \neq i$.
2. Do $\tau_{[i]} = \sum_j \tau_{ij}$
3. Repeat for all i.
4. The anchor node = $\arg\max_i (\tau_{[i]})$.

5 Results and discussion

In this section, we present the results obtained using the described wind resource assessment techniques on data acquired from the roof top anemometers at the Boston Museum of Science. We also examine the improvement in performance of each of the algorithms as more data is made available in the form of 3, 6 and 8 months of training sets. Additionally we study how much benefit is obtained when more expressive copula models are employed. To this end, we constructed the following 5 multivariate copulas and multiple regression (LRR).

Multivariate copulas constructed	
n-Gaussian:	A multivariate Gaussian copula
4-tree C-Vine:	An incomplete C-vine, with depth four.
Full C-Vine:	The complete C-vine.
4-tree D-Vine:	An incomplete D-vine, with depth four.
Full D-Vine:	The complete D-vine.

Results are presented in Tables 1a-c for D_3, D_6, D_8 respectively. Each table shows the KL distance between the ground truth distribution and the distribution estimated based on the predictions provided by each technique for the year 2 dataset (the test dataset) and for every bin. Their right-most column is the sum of the KL distance per bin, which gives an overall performance measure for every model. Tables are sorted according to this value. In addition, the minimum value of each bin and the model that attains most of these minimums have been highlighted.

5.1 Comparison of algorithms

First we compare algorithms when the same amount of data is available to each one of them for modeling. It is clear from Tables 1a-c that model *LRR* is the worst one with a large margin between it and the second and third worst, which are *n-Gaussian* and *4-t D-Vine*. The remaining models attain very good results, with KL distances that range from 0.02 to 0.4. This result suggests that the model needs to incorporate a variety of dependence structures. Figure 4's heatmaps make it easier to extract qualitative conclusions across bins, models and training set size. Bands corresponding to C-vine models are, globally, darker than the rest for D_3, D_6 and D_8. The figure indicates that models are better for west oriented bins (7-12) than for east oriented ones (1-6).

For a combined comparison, we consider two metrics for every model: (i) the sum of KL distance for every bin (right-most column in Tables 1a-c), and (ii) the number of bins with minimum KL distance vs other models (highlighted cells). Sometimes the incomplete version of a vine performs better, or at least, equal to the full version. The most plausible explanation for this is that every tree added to a construction requires new copula estimations which always introduces another source of possible errors. In the C-Vine model the variable at the site of interest is the anchor node of the first tree. It influences every node in the whole C-vine which seems to explain the model's good performance compared with the incomplete D-Vine model. Because this variable is much less influential in D-Vine, one can expect to need more depth in order to attain similar results, as indeed happens.

As additional result, Figure 3 shows an example of truncated C-Vine; the one that corresponds to 8 month training

KL distance with 3 month training set

Model	1	2	3	4	5	6	7	8	9	10	11	12	Sum
Full C-Vine	0.125	0.103	0.033	0.112	0.112	0.090	0.076	0.059	0.034	0.041	0.042	0.063	0.891
4-t C-Vine	0.114	0.119	0.028	0.106	0.130	0.089	0.083	0.059	0.034	0.050	0.035	0.049	0.894
Full D-Vine	0.141	0.106	0.138	0.088	0.086	0.104	0.043	0.058	0.015	0.048	0.055	0.072	0.954
4-t D-Vine	0.162	0.139	0.109	0.101	0.105	0.110	0.051	0.061	0.018	0.071	0.037	0.080	1.043
n-Gaussian	0.093	0.118	0.089	0.197	0.128	0.134	0.108	0.027	0.018	0.057	0.041	0.051	1.059
LRR	0.201	0.268	1.698	3.123	1.291	0.723	0.615	0.427	0.111	0.024	0,128	0,0554	8.664

(a)

KL distance with 6 month training set

Model	1	2	3	4	5	6	7	8	9	10	11	12	Sum
Full D-Vine	0.086	0.051	0.119	0.120	0.116	0.100	0.049	0.036	0.063	0.096	0.062	0.085	0.984
4-t C-Vine	0.059	0.088	0.155	0.143	0.063	0.091	0.066	0.043	0.060	0.093	0.067	0.055	0.985
Full C-Vine	0.071	0.108	0.133	0.149	0.084	0.095	0.068	0.055	0.063	0.097	0.079	0.059	1.061
4-t D-Vine	0.106	0.063	0.147	0.162	0.173	0.098	0.059	0.048	0.062	0.109	0.054	0.087	1.169
n-Gaussian	0.121	0.149	0.141	0.133	0.120	0.096	0.078	0.043	0.047	0.099	0.088	0.092	1.208
LRR	0.236	0.396	0.361	0.963	0.783	0.592	0.528	0.389	0.064	0.036	0.948	0.057	5.355

(b)

KL distance with 8 month training set

Model	1	2	3	4	5	6	7	8	9	10	11	12	Sum
4-t C-Vine	0.064	0.070	0.169	0.071	0.052	0.050	0.066	0.083	0.108	0.090	0.080	0.053	0.955
Full C-Vine	0.067	0.073	0.120	0.087	0.071	0.041	0.058	0.078	0.127	0.091	0.095	0.051	0.959
Full D-Vine	0.103	0.116	0.168	0.158	0.112	0.085	0.077	0.058	0.094	0.110	0.101	0.106	1.288
n-Gaussian	0.136	0.143	0.135	0.110	0.129	0.118	0.099	0.065	0.085	0.122	0.095	0.105	1.342
4-t D-Vine	0.124	0.145	0.203	0.237	0.205	0.134	0.094	0.061	0.101	0.116	0.087	0.094	1.602
LRR	0.357	0.545	0.378	0.523	0.417	0.288	0.364	0.319	0.058	0.035	0.113	0.067	3.468

(c)

Table 1: KL distance for every model and every training set. Each column represents a directional bin but the right-most one, which is the sum of them. The lower the sum, the better the model overall performance.

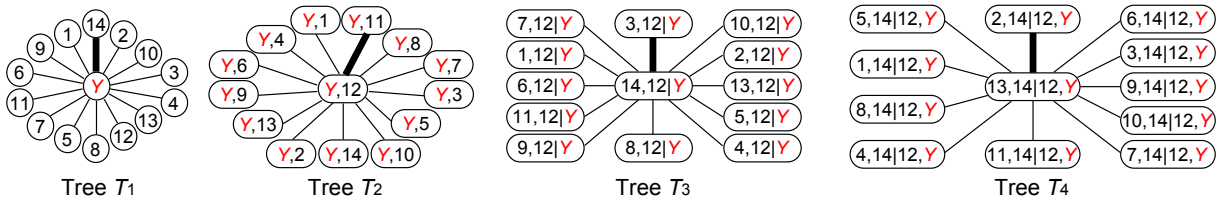


Figure 3: C-Vine truncated in the 4th level constructed out of the data from the 8 month training set and the 1st directional bin. Numbers represent the fourteen airports, and Y is the site of interest (Museum of Science, Boston). The bold edge in every tree represents the couple with highest Kendall's τ, and τ decreases clockwise.

set for the first directional bin; which is highlighted in Table 1c as its best model. The bond between nodes with highest Kendall's tau is represented with a bold edge and edges are deployed clockwise as τ decreases.

5.2 Increasing the data available for modeling

We next examine the robustness of each technique as progressively more data is made available to it for modeling. Figure 5a compares the sum of KL distances of each model for training sets D_3, D_6 and D_8. C-Vine does not significantly change as more/less data is incorporated whereas the other models get worse with more data. This may indicate overfitting which could be a disadvantage of such sensitive, tunable models.

When we examine the minimum KL distance attained with each train set for every bin (not shown),the highest difference in KL distance is less than 0.1. In other words, there is always at least one model out that performs similar to the best one in case of lost data or when more data is available.

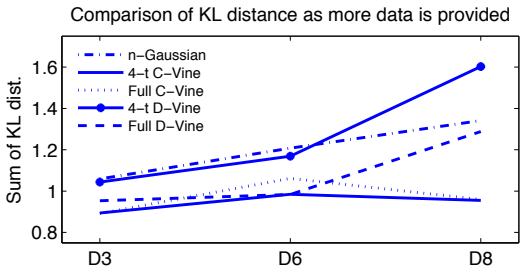


Figure 5: Comparison of performance increasing the data available for modeling. Sum of KL distances for all bins and for all models when D_3, D_6 and D_8 are employed.

6 Conclusions

In this paper we presented copula based approaches for Wind resource estimation. Copula based approach allow us form a joint distribution with Weibull marginals and allow us capture non-linear correlations between the variables. In addition,

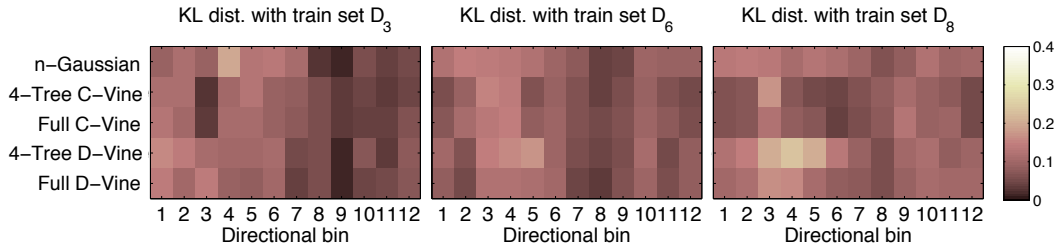


Figure 4: Comparison of different techniques when 3 months worth of data is modeled and integrated with longer term historical data from 14 airports. These results were derived using D_3, D_6, and D_8, and then compared with KL distance to the Weibull distribution estimate of the second year of measurements at the Boston Museum of Science.

we presented a methodology to construct a variety of copula models by factorizing the joint in different ways. With its ability to capture long tails and tail dependencies these models allowed us to estimate the wind resource at the new site with as little as 3 months of data. This is a significant achievement in the wind resource estimation domain where ability to estimate the wind resource accurately in less amount of time allows better planning. Such estimation from reduced amount of time/data is highly beneficial for offshore wind technology development where site-measurement campaigns are extremely expensive.

References

[Bailey *et al.*, 1997] BH Bailey, SL McDonald, DW Bernadett, MJ Markus, and KV Elsholz. Wind resource assessment handbook: Fundamentals for conducting a successful monitoring program. Technical report, National Renewable Energy Lab., Golden, CO (US); AWS Scientific, Inc., Albany, NY (US), 1997.

[Bass *et al.*, 2000] JH Bass, M. Rebbeck, L. Landberg, M. Cabré, and A. Hunter. An improved measure-correlate-predict algorithm for the prediction of the long term wind climate in regions of complex environment. 2000.

[Bedford and Cooke, 2001] T. Bedford and R. M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32:245–268, 2001.

[Burton *et al.*, 2001] T. Burton, D. Sharpe, N. Jenkins, and E. Bossanyi. *Wind energy: handbook*. Wiley Online Library, 2001.

[Cooke *et al.*, 2007] R.M. Cooke, O. Morales, and D. Kurowicka. Vines in overview. In *Invited Paper 3rd Brazilian conference on statistical Modelling in Insurance and Finance, Maresias, March 25-30*, 2007.

[Genest and Rivest, 1993] Christian Genest and Louis-Paul Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 88(423):1034–1043, September 1993.

[Gross and Phelan, 2006] Richard C. Gross and Paul Phelan. Feasibility study for wind turbine installations at museum of science, boston. Technical report, Boreal Renewable Energy Development, October 2006.

[Haff *et al.*, 2010] Ingrid Hobk Haff, Kjersti Aas, and Arnoldo Frigessi. On the simplified pair-copula construction - simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296 – 1310, 2010.

[Iyengar, 2011] S.G. Iyengar. Decision-making with heterogeneous sensors-a copula based approach. *PhD Dissertation*, 2011.

[Kurowicka and Joe, 2011] Dorota Kurowicka and Harry Joe. *Dependence Modeling: Vine Copula Handbook*. World Scientific Publishing Company, 2011.

[Lackner *et al.*, 2008] M.A. Lackner, A.L. Rogers, and J.F. Manwell. The round robin site assessment method: A new approach to wind energy site assessment. *Renewable Energy*, 33(9):2019–2026, 2008.

[Nelsen, 2006] R.B. Nelsen. *An introduction to copulas*. Springer Verlag, 2006.

[Rogers *et al.*, 2005] A.L. Rogers, J.W. Rogers, and J.F. Manwell. Comparison of the performance of four measure-correlate-predict algorithms. *Journal of wind engineering and industrial aerodynamics*, 93(3):243–264, 2005.

[Salinas-Gutiérrez *et al.*, 2010] R. Salinas-Gutiérrez, A. Hernández-Aguirre, and E.R. Villa-Diharce. D-vine eda: A new estimation of distribution algorithm based on regular vines. In *GECCO'10 Portland USA*, pages 359–365, 2010.

[Venter, 2001] Gary Venter. Tails of copulas. In *Proc. of ASTIN Colloquim of International Actuarial Association, Washington, USA.*, 2001.

[Wagner *et al.*, 2011] M. Wagner, K. Veeramachaneni, F. Neumann, and U.M. O'Reilly. Optimizing the layout of 1000 wind turbines. In *Scientific Proceedings of European Wind Energy Association Conference (EWEA 2011)*, 2011.

Appendix

For the sake of completeness we include below relevant expressions of the copulas used for constructing vines. All the following tables include:

- The range of the copula parameter θ.
- The bivariate copula cdf, $C(u, v)$.
- The conditional cdf of v given u, $F(v|u) = \frac{\partial}{\partial u} C(u, v)$.
- The bivariate copula pdf, $c(u, v) = \frac{\partial^2}{\partial u \partial v} C(u, v)$.
- The Kendall's tau given the copula parameter, τ.

In addition, for Frank copulas it is useful to define

$$g_z = e^{-\theta z} - 1$$

.

Clayton copula

$\theta \in (0, \infty)$

$C(u, v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta}$

$F(v|u) = u^{-\theta-1}\left(u^{-\theta} + v^{-\theta} + 1\right)^{-\frac{\theta+1}{\theta}}$

$c(u, v) = (\theta + 1)\,(uv)^{-\theta-1}\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{2\theta+1}{\theta}}$

$\tau = \theta/(\theta + 2)$

Frank copula

$\theta \in (-\infty, \infty)$

$C(u, v) = \frac{-\ln(1+g_u g_v/g_1)}{\theta}$

$F(v|u) = \frac{g_u g_v + g_v}{g_u g_v + g_1}$

$c(u, v) = \frac{-\theta g_1 (1+g_{u+v})}{(g_u g_v + g_1)^2}$

$\tau = 1 - \frac{4}{\theta} + \frac{4}{\theta^2} \int_0^\theta t/\left(e^t - 1\right) dt$

Gumbel copula

$\theta \in [0, \infty)$

$C(u, v) = \exp\left(-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{1/\theta}\right)$

$F(v|u) = C(u, v) \frac{\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{\frac{1}{\theta}-1}}{\theta(-\ln u)^{1-\theta}}$

$c(u, v) = \frac{C(u,v)}{uv} \frac{\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{\frac{2}{\theta}-2}}{(\ln u \ln v)^{1-\theta}}$

$\tau = (\theta - 1)/\theta$