

---

# Copula-Based Wind Resource Assessment

---

**Kalyan Veeramachaneni, Teasha Feldman-Fitzthum, Una-May O’Reilly**  
ALFA Group, CSAIL, MIT  
{kalyan, unamay}@csail.mit.edu; feldmant@mit.edu  
**Alfredo Cuesta-Infante**  
Felipe II College, Universidad Complutense de Madrid  
acuestai@ucm.es

## Abstract

We introduce the use of multivariate Gaussian Copula modeling to improve the accuracy of wind resource assessment. The technique also serves to lower assessment costs because it requires less sensing data than conventional methods.

## 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. [8] generally outlines the process and challenges involved in assessment. Herein we focus on the single critical factor in assessment: achieving the most accurate forecast while incurring minimal financial expense. This implies integrating geographically proximal public wind data sources for better accuracy (accuracy) while concurrently reducing the duration of anemometer sensing during the assessment period (expense).

Our first means of achieving improvement is an end-to-end “Automated Wind Resource Assessment as a Service” that could be deployed to the web or cloud. Figure 1 shows how this service spans from automatic site-neighbor data extraction from public, online sources (see 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 industry standard Weibull distribution formulation of the assessment. The service’s automation eliminates work currently being done manually on a per-assessment basis.

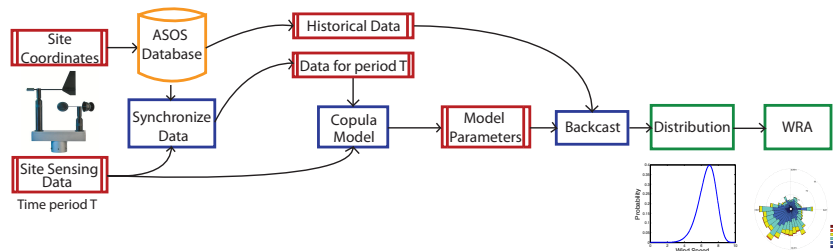


Figure 1: An Automated Wind Resource Assessment Service.

Our second means of improvement is the primary contribution of this paper. We use *multivariate Gaussian copulas* for modeling multiple joint distributions of wind speeds at the site and a publicly available neighboring wind source. This generative modeling step is embedded within a widespread methodology called Measure-Correlate-Predict (MCP) [1–3,5]. In contrast, state of art for modeling in the wind industry is linear regression. For demonstration we use speed and direction data from three actual wind farms in Indiana, Nebraska, and Maine where the available sensor data ranges from 3 to 6 months. We compare our probabilistic model with linear regression, where it achieves higher accuracy with less sensing data in all three cases. Thus we achieve better accuracy at a lower cost.

We proceed by describing MCP while introducing notation in Section 2. Section 3 describes the copula modeling. Section 4 is the demonstration.

## 2 Measure-Corelate-Predict (MCP )

In terms of notation, the wind at a particular location is characterized by speed denoted by  $x$  and direction  $\theta$ . The  $360^\circ$  direction is split into multiple bins with a lower limit ( $\theta_l$ ) and upper limit ( $\theta_u$ ). We give an index value of  $J = 1 \dots j$  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  $M = 1 \dots m$ . The three steps of MCP are:

**MEASURE:** Short term sensing measurements at the site plus ones at neighboring wind recording stations are collected and synchronized. Neighbor data for the past 10-20 years is reserved for back-cast in the PREDICT step. Sensing measurements are denoted by  $Y = \{y_{t_k} \dots y_{t_n}\}$ . Neighboring sites, also called *historical* are denoted by  $X = \{x_{t_k \dots t_n}^{1 \dots m}\}$  where each  $x_{t_k \dots t_n}^i$  corresponds to data from one historical site and  $m$  denotes the total number of historical sites.

**CORRELATE:** For each bin a directional model is built correlating the wind directions observed at the site with simultaneous neighboring site wind directions. Using likelihood parameter estimation we build a multivariate distribution with the probability density function  $f_{\mathbf{X},Y}(\mathbf{x}, y)$ , where  $\mathbf{x} = \{x_1 \dots x_m\}$  are the wind speeds at the historic sites and  $y$  is the wind speed at the site.

Next, for each directional bin, a model is trained, in our case using a multivariate Gaussian copula described in Section 3, correlating the wind speeds at the site with simultaneous speeds at the historical sites, i.e.  $Y_{t_i} = f_{\theta_j}(x_{t_i}^{1 \dots m})$  where  $k \leq i \leq n$ . Notationally, we refer to a model training point as  $l \in \{1 \dots L\}$  and a point for which we have to make prediction as  $k \in \{1 \dots K\}$ . We drop the notation for time after having time synchronized all the measurements across locations and the subscript for directional bin. Now when we refer to a model, it is the model for a particular bin  $j$ .  $f_Z(z)$  refers to a probability density function of the variable (or set of variables)  $z$ .  $F_Z(z)$  refers to cumulative distribution function for the variable  $z$  such that  $F_Z(z = \alpha) = \int_{-\infty}^{\alpha} f_Z(z)$  for a continuous density function.

Given the directional model, we predict the probability density of  $y$  that corresponds to a given test sample  $\mathbf{x}_k = \{x_{1_k} \dots x_{m_k}\}$  by estimating the conditional density  $f_Y(y|\mathbf{x}_k)$ . The conditional can be estimated by:

$$f_{Y|\mathbf{X}=\mathbf{x}_k}(y|\mathbf{x}_k) = \frac{f_{\mathbf{X},Y}(\mathbf{x}_k, y)}{\int_y f_{\mathbf{X},Y}(\mathbf{x}_k, y) dy}. \quad (1)$$

**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. We use the model we developed for that direction  $f_{\theta_j}$  and the data from the historic sites corresponding to this direction  $x_{t_1 \dots t_{k-1}}^{1 \dots m} | \theta_j$  to predict what the wind speed  $\mathbf{Y}_p = y_{t_1 \dots t_{k-1}}$  at the site would be. A point prediction of  $\hat{y}_k$  is made finding the value for  $y$  that maximizes the conditional.

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

Then, with the predictions  $\mathbf{Y}_p$ , we estimate parameters for a Weibull distribution expressing the mean and variance in speed. This is critical for assessment of long term wind resource and the long term energy estimate. The bins' distributions are the *assessment*. The assessment, 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 [9] for more details.

We are able to measure assessment accuracy in our demonstrations because we have sequestered future wind statistics from the site. We measure a symmetric Kullback-Leibler distance. This is intentionally different from mean-squared error or mean-absolute error because these errors would not necessarily accurately express how close the approximation is to the true distribution.

### 3 Copula Modeling

The crux of the methodology is the joint density function of the model. A simple and straightforward choice would be the multivariate Gaussian with Gaussian marginals. However conventionally the univariate densities  $f_{X_i}(x_i)$  are described with Weibull distributions. Copula theory neatly solves this problem, see [6] for more details. A copula function extracts the underlying joint behavior, which can be assumed to be multivariate Gaussian and allows individual behavior (parametric distributions) to be coupled with it as marginals. We first construct the individual parametric distributions, then we couple them to form a multivariate density function. Finally we predict the value of  $y$  given  $x_{1..m}$ . In detail:

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

$$C(u_1, \dots, u_{m+1}; \theta) = F(U_1 \leq u_1, \dots, U_{m+1} \leq u_{m+1}). \quad (3)$$

Let  $U_1 \dots U_m$  represent the cumulative distribution functions (CDF) for variables  $x_1, \dots, x_m$  and  $U_{m+1}$  represent the CDF for  $y$ . Hence the *copula* represents the joint distribution function of  $C(F(x_1) \dots F(x_m), F(y))$ , 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  $x_1 \dots x_m, y$  given by

$$F(x_1 \dots x_m, y) = C(F(x_1) \dots F(x_m), F(y); \theta) \quad (4)$$

The joint probability density function (PDF) is obtained by taking the  $m + 1^{th}$  order derivative of the eq. (4); leading to the Sklar's theorem formulation for densities:

$$f(x_1 \dots x_m, y) = \prod_{i=1}^m f(x_i) f(y) c(F(x_1) \dots F(x_m), F(y)). \quad (5)$$

where  $c(\cdot)$  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. In order to satisfy the assumption of an underlying multivariate gaussian dependence structure, we employ the Gaussian copula given by

$$C_G(\Sigma) = F_G(F^{-1}(u_1) \dots F^{-1}(u_m), F^{-1}(u_y), \Sigma) \quad (6)$$

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

There are two sets of parameters to estimate. The first set of parameters for the multivariate Gaussian copula is  $\Sigma$ . 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:  $L(\mathbf{x}, y; \Sigma, \Psi) = \sum_{l=1}^N \log f(\mathbf{x}_l, y_l | \Sigma, \Psi) = \sum_{l=1}^N \log \{(\prod_{i=1}^m f(x_{li}; \psi_i) f(y_l; \psi_y)) c(F(x_1) \dots F(x_m), F(y); \Sigma)\}$  Parameters  $\Psi$  are estimated, per [4], via:

$$\hat{\Psi} = \arg \max_{\Psi \in \psi} \left\{ \sum_{l=1}^N \log \left\{ \left( \prod_{i=1}^m f(x_{li}; \psi_i) f(y_l; \psi_y) \right) c(F(x_1) \dots F(x_m), F(y); \Sigma) \right\} \right\} \quad (7)$$

A variety of algorithms are available in literature to estimate the MLE in eq. (7), see [4] for a thorough discussion. To obtain predictions from a copula, for a new observation  $\mathbf{x}$  we form the conditional first by

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}, y)}{\int_y P(\mathbf{x}, y) dy}. \quad (8)$$

Our predicted  $\hat{y}$  maximizes this conditional probability  $\hat{y} = \arg \max_{y \in Y} P(y|\mathbf{x})$ . Note that the term in the denominator of eq.( 8) 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.

### 4 Demonstrations and Results

We demonstrate our approach with datasets from 3 actual wind farms we call “Indiana”, “Nebraska” and “Maine”. The topography of the sites varied from simple flat terrain in Indiana to more rough and complex terrain in Maine to rough, complex terrain with high seasonal variability in Nebraska.

We received sensing data collected at each the site from AWS Truepower and automatically, using ASOS, retrieved wind data from neighboring airports, see Figure 2. Data is collected at a frequency of 1 sample/second with 10 minute averages. For *Indiana* we had three months of site sensing, i.e. training data, for *Nebraska* and *Maine* we had six months. We received additional, later observed, sensing data (constituting the prediction period) from AWS Truepower which we used as the "ground truth". 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. For baseline comparison, we also developed a linear regression model which is used quite extensively in wind resource assessment [2, 7].

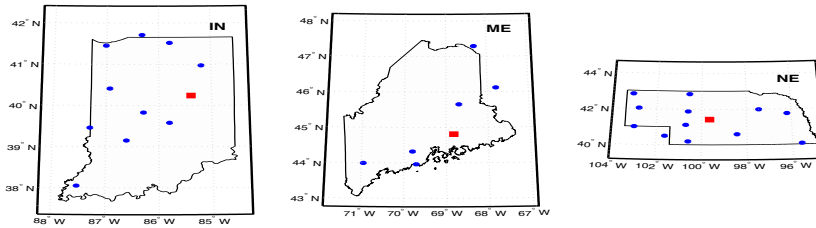


Figure 2: Airports employed in Indiana(IN), Maine(ME) and Nebraska(NE).

Figure 4 compares all methods' KL divergence with ground truth when the models are trained with 3 or 6 months of sensing data. With the exception of one directional bin, for one farm (NE, bin 8), the copula modeling generates consistently more accurate assessments. Linear regression frequently struggles to come close to ground truth with 3 months data. It improves with 6 months of data but still is always inferior to a copula. Figure 4 indicates that a copula's accuracy improves with more sensing data (3 to 6 months). At 3 months it is better than a linear model trained with 6 months data.

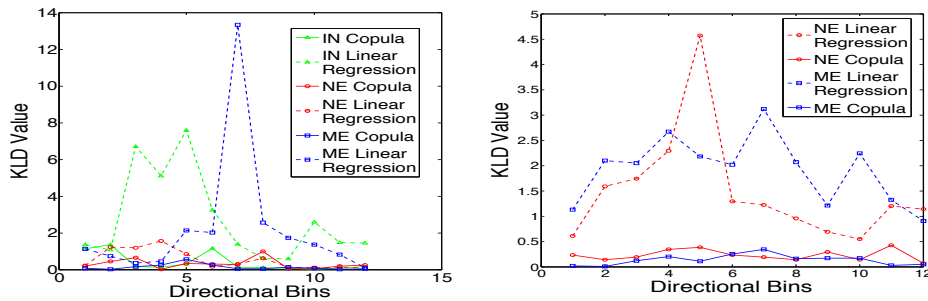


Figure 3: All farms, assessment accuracy using 3 (L) and 6 (R) months of sensing data.

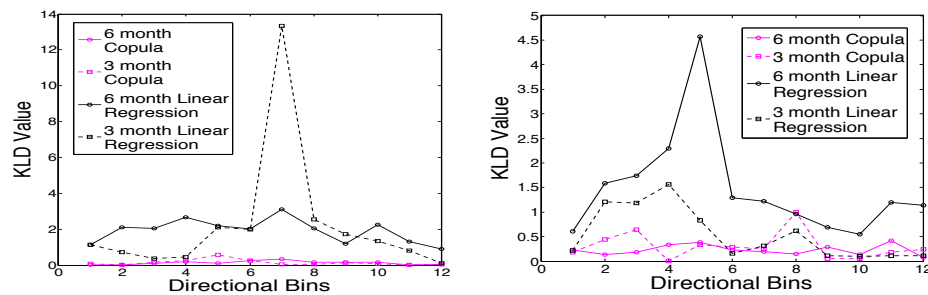


Figure 4: *Maine* (left) and *Nebraska* (right) assessment accuracy, comparing linear regression and copula with 3 and 6 months of sensing data.

## 5 Conclusions

Even with 6 months of sensing data a linear regression cannot match the accuracy of a copula trained on 3 months of sensing data. This relative gap accomplishes our goal: *using a copula it is possible to lower the cost of assessment by trimming the sensing time and concurrently a more accurate assessment is obtainable.*

## References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] S.G. Iyengar. Decision-making with heterogeneous sensors-a copula based approach. *PhD Dissertation*, 2011.
- [5] 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.
- [6] R.B. Nelsen. *An introduction to copulas*. Springer Verlag, 2006.
- [7] 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.
- [8] K. Veeramachaneni, Xiang Ye, and U.M. O’Reilly. *Computational Intelligent Data Analysis for Sustainable Development*, chapter 10, pages 303–330. Chapman & Hall/CRC, 2013.
- [9] 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.