PO.0106

Statistical Identification of Local and Regional Wind Regimes

Karen Kazor (with A. S. Hering)

Colorado School of Mines; Department of Applied Mathematics and Statistics

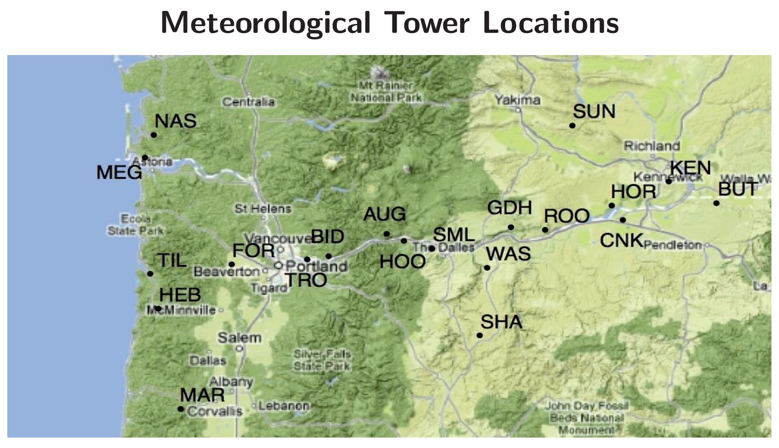
Introduction

- **Objective:**
  - Quantitatively identify predominant wind conditions, or *wind regimes*, in a given region and compare different regime identification methods.
- **Background:**
  - Wind regimes have been used successfully to:
    - Improve wind speed predictions (Gneiting et al. 2006)
    - Generate synthetic wind speed data (Shamshad et al. 2005)
    - Analyze a region's wind power potential (Burlando et al. 2008)
  - Methods for regime identification:
    - are often difficult to generalize: subjective, depend on expert knowledge, or are based on the optimization of a specific model
    - utilize only basic clustering techniques: hierarchical and k-means
  - Methods are not directly assessed for the ability to correctly classify observations, but rather indirectly on:
    - performance of a given forecasting model
    - ability to replicate certain features in synthetic data
- **Application:**
  - An more informed and objective wind regime identification process has the potential to:
    - Inform the siting of wind farms
    - Optimize turbine placement and specifications
    - Improve wind energy forecasting and prediction

Data

- One year of wind data (Dec 2010-11) obtained from the Bonneville Power Administration for twenty meteorological sites in the Pacific Northwest.

Meteorological Tower Locations



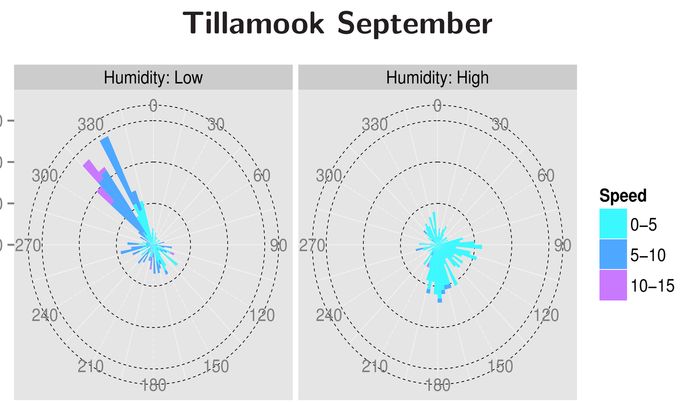
Variables:
- Barometric pressure (inHG)
- Relative humidity (%)
- Temperature (°F)
- Wind direction (Deg)
- Wind speed (MPH)

Adjustments:
- Quality corrected
- Raw 5 and 10-min averages converted to hour-ending averages
- Wind speeds adjusted to a standard height of 70m using the power law

See http://transmission.bpa.gov/Business/Operations-/Wind/MetData.aspx for more information.

Defining Regimes

Tillamook September



- Different sites generally exhibit two distinct wind behaviors
- Tillamook's September dataset selected for differences in the distribution of wind speed and wind direction
  - R1: characterized by low humidity observations
  - R2: characterized by high humidity observations

Synthetic Data Simulation

Model:
Goals in simulating data are to capture:
- distinct behaviors of the R1 and R2 datasets
- variable and temporal dependence characteristic of wind and atmospheric variables

To achieve these characteristics we use a diurnal-adjusted Markov-switching vector autoregressive (MS-VAR) model:

$$\mathbf{y}_t - \boldsymbol{\mu}_{r_t}(h) = \mathbf{A}_{r_t}(\mathbf{y}_{t-1} - \boldsymbol{\mu}_{r_{t-1}}(h-1)) + \epsilon(r_t); \qquad \epsilon(r_t) \sim N(0, \boldsymbol{\Sigma}_{r_t}),$$

- $\mathbf{y}_t = (pressure, u, v)'$
- $\{R_t\}$ a Markov chain on finite space, $\{1, 2\}$, indicating the regime at time t.
- $\boldsymbol{\mu}_{r_t}(h)$ mean vector; a function of the hour h
- $\mathbf{A}_{r_t}$ lag-one autoregressive matrix
- $\boldsymbol{\Sigma}_{r_t}$ innovation covariance matrix

$$\{\boldsymbol{\mu}_{r_t}(h), \mathbf{A}_{r_t}, \boldsymbol{\Sigma}_{r_t}\} = \begin{cases} \{\boldsymbol{\mu}_1(h), \mathbf{A}_1, \boldsymbol{\Sigma}_1\} & \text{if } r_t = 1 \\ \{\boldsymbol{\mu}_2(h), \mathbf{A}_2, \boldsymbol{\Sigma}_2\} & \text{if } r_t = 2 \end{cases}$$

Parameter Values:
- Parameters are determined based on the observations classified within the R1 and R2 regimes.
- The regime-switching process, $\{R_t\}$, is defined by a transition probability matrix $P = \{p_{jk}\}$, where $p_{jk} = P(r_{t+1} = k | r_t = j)$; $(j, k = 1, 2)$. In simulating data, observed proportions are used to switch between regimes:

$$P = \begin{array}{c c} & \begin{array}{cc} R1 & R2 \end{array} \\ \begin{array}{c} R1 \\ R2 \end{array} & \begin{pmatrix} 0.92 & 0.08 \\ 0.10 & 0.90 \end{pmatrix} \end{array}$$

- Expected proportion of time: R1 = 0.55; R2 = 0.45
- Expected duration: R1 = 12.3 hrs; R2 = 10.0 hrs

Simulation Scenarios

To better understand the performance of the regime identification methods tested, data is simulated under four scenarios:

Scenario	Variable Dependence	Observation Dependence	Constraints
S1	Independent	Independent	diag(Σ_{r_t}), A_{r_t}=0, $p_{jk} = 0.50\ \forall j, k$
S2	Independent	Dependent: AR(1)	diag(Σ_{r_t}), diag(A_{r_t})
S3	Dependent	Independent	A_{r_t}=0, $p_{jk} = 0.50\ \forall j, k$
S4	Dependent	Dependent: VAR(1)	None

For more information, email kkazor@mines.edu

Regime Identification Methods

We consider the following five methods for identifying wind regimes:

1. Switching algorithm based on the wind direction (Directional)
2. K-means clustering

3–4. Gaussian mixture model-based (GMM) clustering
   3. constrained covariance matrix, GMM(Λ_k)
   4. unconstrained covariance matrix, GMM(Σ_k)

5. Nonparametric mixture model-based (NPMM) clustering

Mixture Models

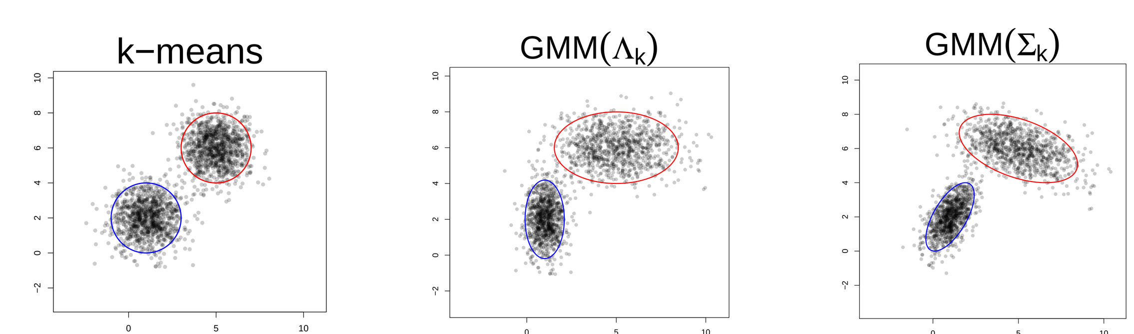
The k-means, GMM, and NPMM methods are forms of model-based clustering that involve fitting finite mixture models:
- Assume the random variable $\mathbf{X} \in \mathbb{R}^d$ is generated from K distinct random processes or regimes
- Regimes are each modeled by a density, $f_k(\mathbf{x}|\boldsymbol{\theta}_k)$
- Parameters τ_k represent the proportion of observations generated under the k^{th} regime

$$f(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \tau_k f_k(\mathbf{x}|\boldsymbol{\theta}_k),$$

where $\Theta = (\tau_1, ..., \tau_K; \boldsymbol{\theta}_1, ...\boldsymbol{\theta}_K)$ is the set of parameters.

GMM and k-means
Involve fitting a mixture of K multivariate Gaussian distributions with different covariance structures:

$$f_k(\mathbf{x}|\boldsymbol{\theta}_k) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_k)$$



$\boldsymbol{\Sigma}_k = \lambda I$ $\qquad$ $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k$ $\qquad$ Unconstrained $\boldsymbol{\Sigma}_k$

NPMM
Rather than assume a specific structure for all regimes, we also consider the use of nonparametric kernel densities.

To fit an NPMM, we use the *npEM* algorithm developed by Benaglia et al. (2009).
- Although observations may be multivariate, variables are assumed to be independently distributed within each regime.

$$f(\mathbf{x}|\Theta) = \sum_{k=1}^{K} \tau_k g_k(\mathbf{x}) = \sum_{k=1}^{K} \tau_k \prod_{j=1}^{d} f_{jk}(x_j),$$

where $\Theta' = (\boldsymbol{\tau}', \mathbf{g}') = (\tau_1, ..., \tau_K; g_1, ..., g_K)'$ denotes the parameter vector, and $f_{jk}(\cdot)$ denotes the nonparametric univariate density of the j^{th} variable under the k^{th} regime.

Simulation Outline

1. Generate 500 datasets each of length 2,000 (one season) under each of the four scenarios: S1–S4.
2. Given the number of regimes, $K = 2$, methods are applied to *non-detrended* simulated data and average misclassification rates across the 500 datasets are computed under each scenario and for each method.
3. The GMM(Λ_k) and GMM(Σ_k) methods are tested for their ability to determine the correct number of regimes using BIC.

Simulation Results: Misclassification Rates

	Average Misclassification Percentage (%) when applied to Pressure and the U and V wind components				Average Misclassification Percentage (%) when applied to the U and V wind components			
Scenario: [Var/Obs]:	S1 [Ind/Ind]	S2 [Ind/Dep]	S3 [Dep/Ind]	S4 [Dep/Dep]	S1 [Ind/Ind]	S2 [Ind/Dep]	S3 [Dep/Ind]	S4 [Dep/Dep]
Directional	37.95 (0.0482)	39.99 (0.0856)	37.98 (0.0459)	40.12 (0.0898)	37.95 (0.0482)	39.99 (0.0856)	37.98 (0.0459)	40.12 (0.0898)
k-means	35.70 (0.0626)	39.10 (0.1130)	35.19 (0.0597)	39.02 (0.1210)	35.70 (0.0628)	39.10 (0.1130)	35.18 (0.0598)	39.02 (0.1209)
NPMM	31.35 (0.1137)	37.20 (0.1310)	35.05 (0.0638)	40.15 (0.1720)	33.33 (0.0547)	37.70 (0.1177)	34.12 (0.0483)	38.26 (0.1135)
GMM($\boldsymbol{\Lambda}_k$)	25.91 (0.3583)	38.91 (0.3052)	38.03 (0.1318)	41.90 (0.1452)	29.10 (0.2593)	39.08 (0.2586)	35.09 (0.2386)	41.47 (0.1528)
GMM($\boldsymbol{\Sigma}_k$)	**21.30** (0.1896)	**30.68** (0.3421)	**22.38** (0.2670)	**31.88** (0.3529)	**25.35** (0.1818)	**31.60** (0.2700)	**26.60** (0.2144)	**32.65** (0.2643)

Simulation Results: Number of Regimes

Pressure and the U and V wind components

	Scenario [Variable/Observation]	Number of Regimes (K) 1	2	3	4	5	6	7
GMM($\boldsymbol{\Lambda}_k$)	S1 [Ind,Ind]	0.00	0.80	**97.00**	2.20	0.00	0.00	0.00
	S2 [Ind,Dep]	0.00	0.80	**83.40**	11.60	3.40	0.60	0.20
	S3 [Dep,Ind]	0.00	0.00	**92.00**	7.00	1.00	0.00	0.00
	S4 [Dep,Dep]	0.00	0.00	**55.20**	22.00	15.60	5.80	1.40
GMM($\boldsymbol{\Sigma}_k$)	S1 [Ind,Ind]	0.00	**94.40**	5.60	0.00	0.00	0.00	0.00
	S2 [Ind,Dep]	0.00	**67.60**	27.80	4.00	0.60	0.00	0.00
	S3 [Dep,Ind]	0.00	**85.80**	14.20	0.00	0.00	0.00	0.00
	S4 [Dep,Dep]	0.00	**61.20**	33.20	4.60	0.80	0.20	0.00

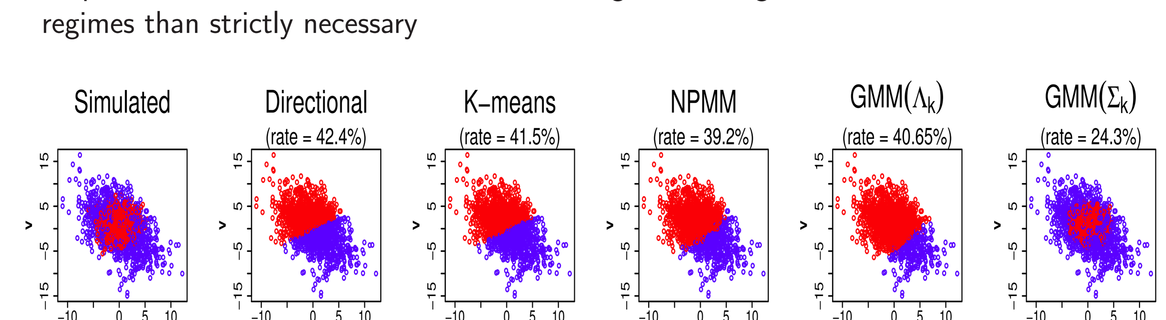
U and V wind components

	Scenario [Variable/Observation]	Number of Regimes (K) 1	2	3	4	5	6	7
GMM($\boldsymbol{\Lambda}_k$)	S1 [Ind,Ind]	0.00	0.60	**98.60**	0.80	0.00	0.00	0.00
	S2 [Ind,Dep]	0.00	6.00	**93.40**	0.60	0.00	0.00	0.00
	S3 [Dep,Ind]	0.00	0.00	**99.60**	0.40	0.00	0.00	0.00
	S4 [Dep,Dep]	0.00	0.40	**97.60**	2.00	0.00	0.00	0.00
GMM($\boldsymbol{\Sigma}_k$)	S1 [Ind,Ind]	0.00	**92.80**	7.20	0.00	0.00	0.00	0.00
	S2 [Ind,Dep]	0.00	**83.60**	16.40	0.00	0.00	0.00	0.00
	S3 [Dep,Ind]	0.00	**89.60**	10.40	0.00	0.00	0.00	0.00
	S4 [Dep,Dep]	0.20	**84.80**	15.00	0.00	0.00	0.00	0.00

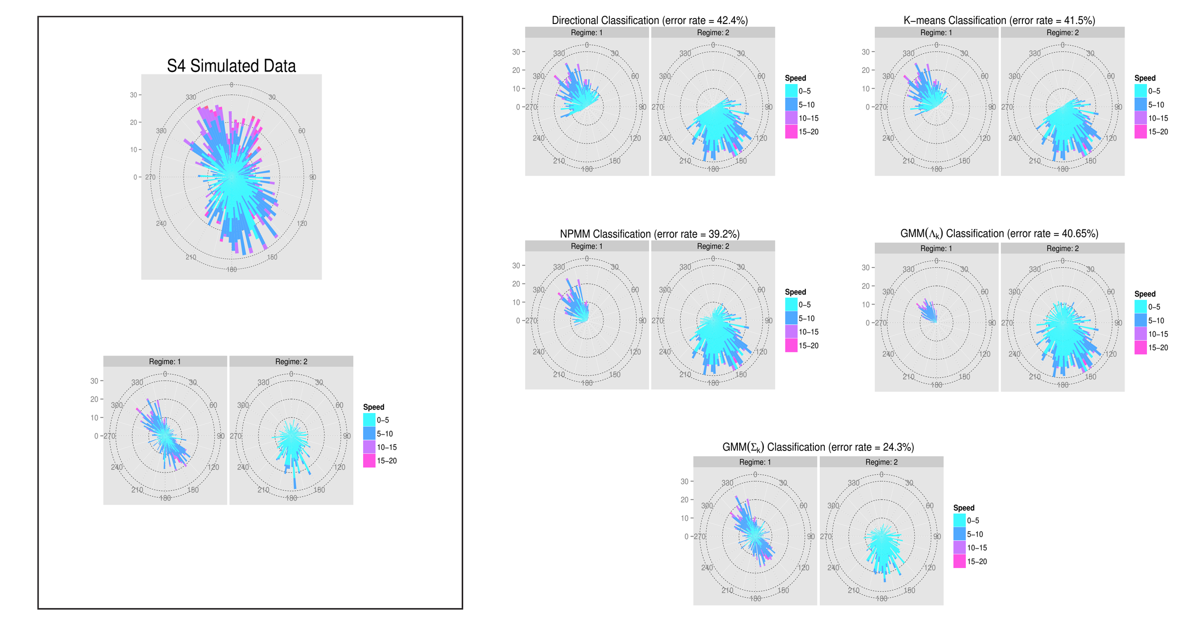
Simulation Results

Results point to GMM($\boldsymbol{\Sigma}_k$) as the best method for identifying wind regimes out of the five methods considered:

- lowest average misclassification rates
- frequent identification of correct number of regimes, though sometimes chooses more or less regimes than strictly necessary



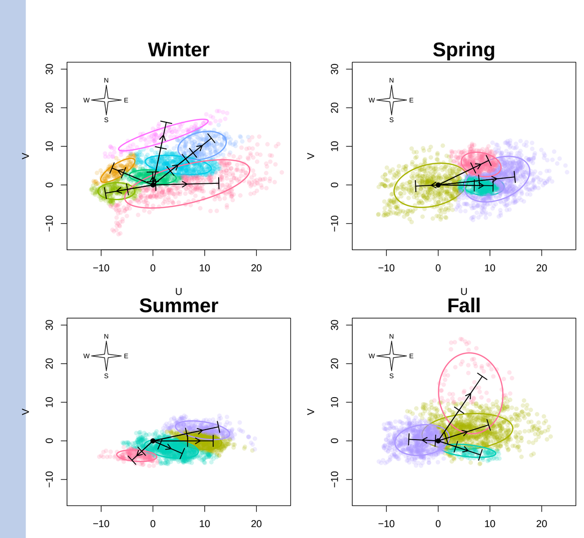
Simulation Results: Misclassification Rates



Identifying Regimes

Apply the GMM($\boldsymbol{\Sigma}_k$) method to non-detrended observed data to identify both regional and local wind regimes
- Regional: average the hourly u and v components at each time across all locations
- Local: apply to the u and v components observed at each location

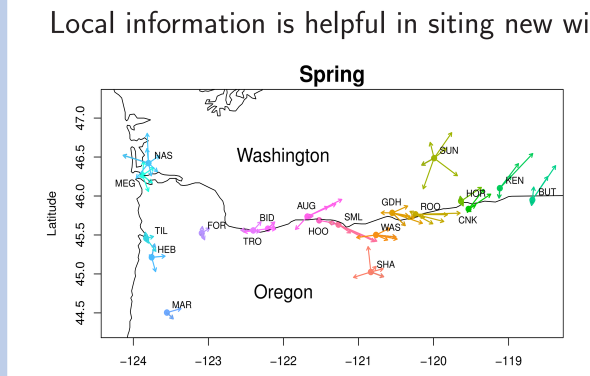
Regional Regimes



Regime	% in Each Regime	Speed Mean	Speed SD	Direction Mean	Direction SD	Cardinal Direction
Winter						
W1	17.46	2.11	1.25	179	0.92	S
W2	20.79	6.64	6.10	268	0.93	W
W3	25.80	7.11	2.16	223	0.45	SW
W4	11.99	7.24	2.18	77	0.19	E
W5	7.54	7.77	1.27	119	0.30	SE
W6	5.29	13.08	3.35	189	0.42	S
W7	11.14	13.96	2.59	223	0.15	SW
Spring						
Sp1	25.58	1.37	3.00	86	1.74	E
Sp2	20.07	8.79	1.82	271	0.16	W
Sp3	18.55	9.86	1.78	237	0.20	SW
Sp4	35.80	11.46	3.47	262	0.33	W
Summer						
Su1	40.69	4.08	2.48	300	0.80	NW
Su2	7.98	4.95	1.50	39	0.37	NE
Su3	37.15	9.17	2.47	270	0.16	W
Su4	14.18	9.98	3.20	254	0.24	W
Fall						
F1	38.58	3.12	2.57	94	0.86	E
F2	51.45	6.14	4.45	247	0.54	SW
F3	6.79	6.34	2.59	294	0.16	NW
F4	3.18	13.83	4.92	207	0.23	SW

Local Regimes

Local information is helpful in siting new wind farms or local forecasting



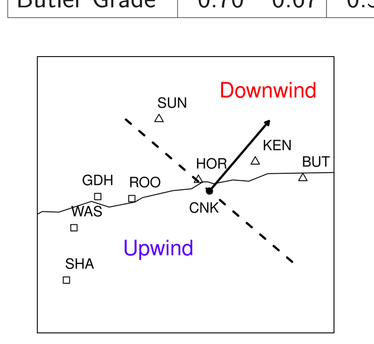
- West-east flows along the gorge
- Diversity in regime direction at sites further from the gorge
- East of the Cascades: strong easterly winds in winter and fall (TRO and BID)
- West of the Cascades: strong westerly winds in spring and summer

Case Study

In building a fully space-time model of wind for utility system planning, a common difficulty is characterizing space-time dependencies since spatial dependence among locations changes with the wind speed and wind direction.

Case study: Use regimes to distinguish between different correlation structures
- Specifically, examine changes in the correlation of the lagged wind speeds at neighboring sites and the current wind speeds at a prediction site depending on whether the neighbor is predominantly upwind or downwind in a given regime.

Prediction site: Chinook			Regional Regimes				Local Regimes			
Neighbors	Overall		Upwind		Downwind		Upwind		Downwind	
	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$	$t-1$	$t-2$
Shaniko	0.32	0.34	0.35	0.38	0.28	0.29	0.35	0.37	0.24	0.28
Wasco	0.56	0.57	0.54	0.56	0.47	0.45	0.58	0.59	0.32	0.30
Goodnoe Hills	0.78	0.78	0.76	0.76	0.42	0.41	0.79	0.79	0.37	0.35
Roosevelt	0.82	0.81	0.81	0.80	0.41	0.37	0.83	0.82	0.23	0.25
Sunnyside	0.42	0.41	0.49	0.47	0.37	0.39	0.31	0.30	0.37	0.35
Horse Heaven	0.80	0.77	0.80	0.77	0.68	0.62	0.87	0.82	0.71	0.68
Kennewick	0.62	0.62	0.36	0.33	0.58	0.57	0.38	0.33	0.62	0.61
Butler Grade	0.70	0.67	0.33	0.30	0.66	0.63	0.12	0.20	0.70	0.68



- Correlations with sites to the west are stronger when neighbors are **upwind**
- Correlations with sites to the east are stronger when neighbors are **downwind**
- Correlations with Roosevelt are much weaker when Roosevelt is **downwind**

Conclusions

- The GMM($\boldsymbol{\Sigma}_k$) method performs the best out of the five methods considered in the simulation analysis.
- Regimes identified by applying GMM($\boldsymbol{\Sigma}_k$) to regional and local data are consistent with known wind patterns in this region.
- Correlations can be considerably different when restricting computations to observations falling within a set of regimes.
- Regimes may provide a better understanding of the spatial dependence among neighboring sites under different sets of wind conditions.

References

- Benaglia, T., Chauveau, D., and Hunter, D. R. (2009) "An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures," *Journal of Computational and Graphical Statistics*, 18: 505–526.
- Burlando, M., Antonelli, M., and Ratto, C. F. (2008) "Mesoscale wind climate analysis: Identification of anemological regions and wind regimes," *International Journal of Climatology*, 28: 629–641.
- Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006) "Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space-time method," *Journal of the American Statistical Association*, 101: 968–979.
- Shamshad, A., Bawadi, M. A., Wan Hussin, W. M. A., Majid, T. A., and Sanusi, S. A. M. (2005) "First and second order Markov chain models for synthetic generation of wind speed time series," *Energy*, 30: 693–708.