Assessment of site-specific wind predictions by a hidden Markov model for multivariate circular-linear data

Alessio Pollice [1], Gianluca Mastrantonio [2], Francesca Fedele [3]

[1] *Dipartimento di Scienze Economiche e Metodi Matematici*
Università degli Studi di Bari Aldo Moro
Largo Abbazia Santa Scolastica, 70124, Bari, Italy
alessio.pollice@uniba.it
[2] *Dipartimento di Scienze Matematiche*
Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129, Torino, Italy
gianluca.mastrantonio@polito.it
[3] *ARPA Puglia*
Corso Trieste 27, 70126, Bari, Italy
f.fedele@arpa.puglia.it

Abstract. We consider a heavy industrial district close to the city of Taranto where winds from North-West quadrants and lack of precipitations are known to lead to a deterioration of urban air quality in terms of PM10 concentrations. In 2012, the Apulia Government adopted a Regional Air Quality Plan prescribing a reduction of industrial emissions by 10% every time such meteorological conditions are forecasted 72 hours in advance. In order to activate the appropriate safety measures, wind prediction is addressed by the Regional Environmental Protection Agency (ARPA Puglia) using the Weather Research and Forecasting (WRF) atmospheric simulation system. Here we investigate the ability of the WRF system to properly predict the local wind speed and direction allowing different performances for unknown weather regimes. Replicate observations of observed and WRF-predicted wind speed and direction at a relevant point location within the area of interest are jointly modeled as a multivariate 4-dimensional time series with a finite number of states (wind regimes) characterized by homogeneous distributional behavior. Observed and simulated wind data are made of two circular (direction) and two linear (speed) variables, then the 4-dimensional time series is jointly modeled by a mixture of projected-skew normal distributions with time-independent states, where the temporal evolution of the state membership follows a first order Markov process. Parameter estimates are obtained by a Bayesian MCMC-based method and results provide useful insights on wind regimes corresponding to different performances of WRF predictions.

Keywords. Environment, Mixture models

1 Observed and simulated wind data

We are concerned with the Tamburi neighborhood within the city of Taranto, located less than 1 Km away from a huge steel plant, downwind with wind directions from the North-West quadrant. Several PM10 limit value exceedances were recorded in this neighborhood mostly in presence of extreme wind conditions encouraging the pollutants transport from the industrial site to the adjacent urban area (Fedele et al., 2014). In 2012, the Apulia Government adopted a Regional Air Quality Plan prescribing a reduction of emissions by 10% every time intense winds from the North-West quadrant and lack of precipitation are forecasted 72 hours in advance. The Weather Research and Forecasting (WRF) atmospheric simulation system (Skamarock et al., 2008) is actually adopted by the local environmental protection agency (ARPA Puglia) for wind forecasting. Hourly predicted wind speed and direction data for the whole year 2014 are here compared to corresponding ground data collected at the San Vito air quality monitoring station, located in the Tamburi neighborhood. As wind direction is a directional variable, notice that observed and predicted wind direction and speed have the form of a 4-dimensional mixed circular-linear time series. We propose a modeling framework where the performance of WRF in predicting wind speed and direction is assessed, considering that atmospheric simulation systems such as WRF show different performances for unknown weather (here wind) regimes. The special topology of the support of the data complicates their modeling that implies accounting for cross-correlations between angular and linear measurements and between observed and simulated data across time and for a finite number of wind regimes (states). Indeed wind intensities are also typically negatively skewed and directional data are rarely symmetric.

2 The Projected Skew Normal distribution

As multivariate circular-linear distribution model we consider the projected skew normal (Mastrantonio, 2015). This distribution is defined constructively, starting from a $(2p+q)$-dimensional random vector $(\mathbf{W}, \mathbf{Y})'$ distributed as a multivariate skew normal (Sahu et al., 2003) with parameters $\boldsymbol{\mu} = (\boldsymbol{\mu}_w, \boldsymbol{\mu}_y)'$, $\boldsymbol{\Sigma}$ and $diag(\mathbf{0}_{2p}, \boldsymbol{\lambda})$. Vector $\mathbf{W}$ is divided into p couples each transformed into polar coordinates giving rise to p lengths R and p angular (or circular) variables Θ. The distribution of the $(p+q)$-dimensional random vector $(\boldsymbol{\Theta}, \mathbf{Y})'$ that arises transforming $(\mathbf{W}, \mathbf{Y})'$ is called the (p,q)-variate projected-skew normal ($PSN_{p,q}$) and is governed by three parameters: $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$.

One of the features that makes this distribution attractive is that the PSN is really flexible and it allows to introduce dependence between and within circular and linear variables. It is also closed under marginalization, i.e. each univariate and multivariate marginal distribution is still PSN. It follows that the marginal distribution of $\mathbf{Y}$ is skew normal while $\boldsymbol{\Theta}$ is projected normal (Wang and Gelfand, 2013). Although a closed form is not available for the PSN density, introducing the lengths $\mathbf{R} = \{R_i\}_{i=1}^p$ of the polar

representation and vector $\mathbf{D}$ of the stochastic representation of the skew normal (Li, 2005), the density of $(\boldsymbol{\Theta}, \mathbf{R}, \mathbf{Y}, \mathbf{D})'$ is given by:

$$2^q \phi_{2p+q}((\mathbf{w}, \mathbf{y})' | (\boldsymbol{\mu}_w, \boldsymbol{\mu}_y + \text{diag}(\boldsymbol{\lambda})\mathbf{d})', \boldsymbol{\Sigma}) \phi_q(\mathbf{d}|\mathbf{0}, \mathbf{I}_q) \prod_{i=1}^{p} R_i \tag{1}$$

that is more easily handled and we say that $(\boldsymbol{\Theta}, \mathbf{R}, \mathbf{Y}, \mathbf{D})'$ is distributed as an *augmented projected-skew normal*: $(\boldsymbol{\Theta}, \mathbf{R}, \mathbf{Y}, \mathbf{D})' \sim AugPSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. The PSN inherits the well known identification problem from the projected normal, successfully addressed by Mastrantonio (2015) in a Bayesian framework by the use of an MCMC parameter estimation algorithm based on a mixed slice-Gibbs sampling strategy.

Since $\mathbf{Y}$ is marginally distributed as a skew normal it is easy to interpret its associated parameters, while the meaning of those involving circular variables is less straightforward. However, Bayesian Monte Carlo approximations are obtained for the most relevant features of the PSN, such as the circular mean (α), the circular concentration (ζ), the correlation coefficient between circular variables (Fisher, 1996) and the circular-linear correlation coefficient (Mardia, 1976), thus bypassing the difficulties in parameter interpretability.

3 The Hidden Markov model

The 4-dimensional circular-linear time series is jointly modeled by a mixture of projected-skew normal distributions with time-independent states, where the temporal evolution of the state membership follows a first order Markov process, namely a hidden Markov model (HMM).

At times $t = 1, \ldots, T$, let $z_t \in \mathcal{K} \subset \mathbb{N}$ be a discrete random variable that represents the state of the HMM, $\boldsymbol{\psi}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\lambda}_k)'$ be the set of PNS parameters at state k and $\boldsymbol{\pi}_k = \{\pi_{kj}\}_{j \in \mathcal{K}}$ be a vector of probabilities. The HMM is formalized as follows:

$$f(\boldsymbol{\theta}, \mathbf{r}, \boldsymbol{y}, \mathbf{d} | \{z_t\}_{t=1}^T, \{\boldsymbol{\psi}_k\}_{k \in \mathcal{K}}) = \prod_{t=1}^{T} \prod_{k \in \mathcal{K}} f(\boldsymbol{\theta}_t, r_t, \mathbf{y}_t, d_t | \boldsymbol{\psi}_{z_t})^{I(z_t = k)}, \tag{2}$$

$$\boldsymbol{\Theta}_t, \mathbf{R}_t, \mathbf{Y}_t, \mathbf{D}_t | \boldsymbol{\psi}_k \sim AugPSN_{p,q}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\lambda}_k), \tag{3}$$

$$z_t | z_{t-1}, \{\boldsymbol{\pi}_k\}_{k \in \mathcal{K}} \sim \boldsymbol{\pi}_{z_{t-1}}. \tag{4}$$

In this work the HMM is estimated within a non-parametric Bayesian framework, relying on Dirichlet process priors for $\boldsymbol{\pi}_k$'s, thus leading to the sticky Hierarchical Dirichlet process-HMM of Fox et al. (2011). This specification allows to estimate the unknown number of latent states, along with all other model parameters.

4 Some results

Prior to modeling, circular variables were transformed from degrees to radians and the log of speed was taken, for both ground and WRF-simulated wind data. Results were back-transformed to their original units to improve the interpretation of graphical and tabular displays. Separate models were estimated for the four seasons of year 2014. All models were estimated considering 400000 iterations, with 300000 for the burn-in phase and thinning by 20, i.e. taking 5000 samples for inferential purposes. A standard weak informative prior setting was used, with $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim NIW\left(\mathbf{0}_6, 0.001, 15, \mathbf{I}_6\right)$, $\boldsymbol{\lambda}_k \sim N_2\left(\mathbf{0}_2, 100\mathbf{I}_2\right)$ for PSN parameters. HMM's allowed to estimate 5 wind regimes for all of the four seasons with $P(K = 5|\boldsymbol{\theta}, \mathbf{y}) \sim 1$. The five wind regimes were ordered from weakest to strongest, based on their ground speed HMM posterior means. As a matter of fact, the output of the model estimation process provides a wealth of information that is hard to exhaustively report in the limited space of a short paper, some highlights are then given in the following for the summer season. A remarkable bimodality of the wind directions with peaks around the SE and NW quadrants is shown in fig. 1 together with a strong asimmetry of the wind speed. Empirical and HMM-estimated marginal distributions (solid and dashed lines) substantially agree, while WRF-simulated wind speed (fig. 1, right, grey lines) overestimates ground recordings (black lines). Tab. 1 allows to investigate the main features of the five detected wind regimes, corresponding to winds blowing from the NE, W, W, SE and NW quadrants with increasing speed. On average, winds with higher speed (regimes 4 and 5) show smaller variability for circular variables, i.e. stronger winds have more focused directions. Concerning forecast verification, given the tendency of WRF to overestimate wind speed in all regimes, tab. 1 shows a good agreement of observed and WRF-simulated means for "extreme" regimes 4 and 5, while WRF seems to have more troubles in forecasting winds with low to intermediate speed. Another assessment of the predictive performance of WRF is given in fig. 2, where regimes 4 and 5 are those associated to higher circular-circular and linear-linear correlations. Observed circular-linear correlations (regimes 3 and 5) are not reproduced by WRF forecasts that show some circular-linear correlation for in regime 2. Analogous results are available for all the four seasons of year 2014, completing the picture of the proposed method in the context of distributions-oriented wind forecast verification.

References

[1] Fedele, F., Menegotto, M., Trizio, L., Angiuli, L., Guarnieri, A., Carducci, C., Bellotti, R., Giua, R., Assennato G. (2014), Meteorological effects on pm10 concentrations in an urban industrial site: a statistical analysis. *Conference Proceedings - 1st International Conference on Atmospheric DUST*, 162–167.

[2] Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, M., Duda, K.G., Huang, X.Y., Wang, W., Powers, J.G. (2008), A description of the Advanced Research

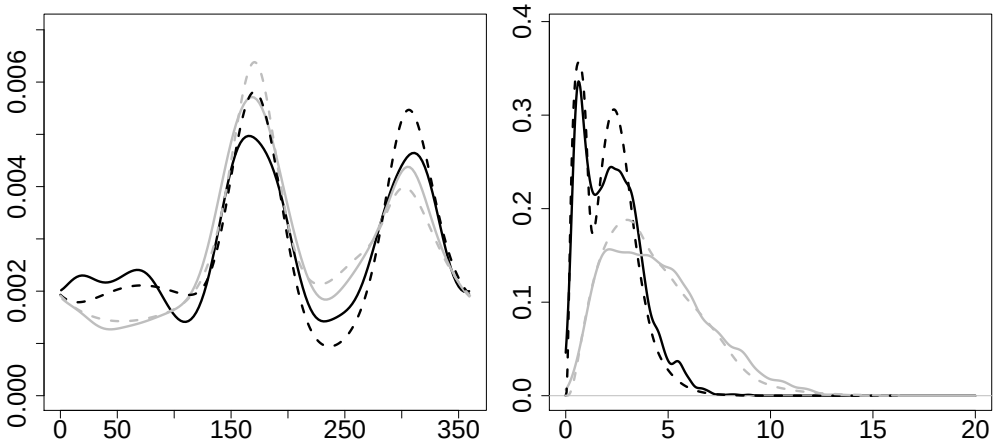

Figure 1: Marginal distributions of wind speed (right) and direction (left) in summer 2014. Black lines represent observed speed and direction, grey lines are WRF-simulated. Solid lines are smooth approximations of the empirical distribution, dashed lines are HMM-predicted distributions.

WRF Version 3, *Technical report, National Center for Atmospheric Research*.
[3] Mastrantonio, G. (2015), A Bayesian hidden Markov model for telemetry data *ArXiv e-prints*.
[4] Sahu, S.K., Dey, D.K., Branco, M.D. (2003), A new class of multivariate skew distributions with applications to Bayesian regression models, *Canadian Journal of Statistics*, 31(2), 129–150.
[5] Wang, F., Gelfand, A. E. (2013), Directional data analysis under the general projected normal distribution, *Statistical Methodology*, 10(1), 113–127.
[6] Li, J. (2005), Clustering based on a multilayer mixture model, *Journal of Computational and Graphical Statistics*, 14(3), 547–568.
[7] Fisher, N.I. (1996), *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge.
[8] Mardia, K. V. (1976), Linear-Circular Correlation Coefficients and Rhythmometry, *Biometrika*, 63(2), 403–405.
[9] Fox, E.B., Sudderth, E.B., Jordan, M.I., Willsky, A.S. (2011), A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A), 1020–1056.

	1	2	3	4	5
$\alpha^c_{k,1}$	84.871	296.234	307.542	173.961	312.629
$\alpha^c_{k,2}$	82.766	261.101	176.688	168.684	322.396
$\zeta^c_{k,1}$	0.445	0.885	0.305	0.072	0.086
$\zeta^c_{k,2}$	0.836	0.308	0.695	0.126	0.173
$\alpha^l_{k,1}$	0.7	2.033	2.337	2.573	3.795
$\alpha^l_{k,2}$	2.493	7.368	3.065	4.562	5.927
$\zeta^l_{k,1}$	0.136	1.605	1.221	0.628	1.981
$\zeta^l_{k,2}$	1.524	5.134	2.677	4.129	4.098

Table 1: MC estimates of the means (α) and variances/concentrations (ζ) of observed (1) and WRF-simulated (2) wind speed (linear l) and direction (circular c) for the 5 estimated wind regimes in the summer season.

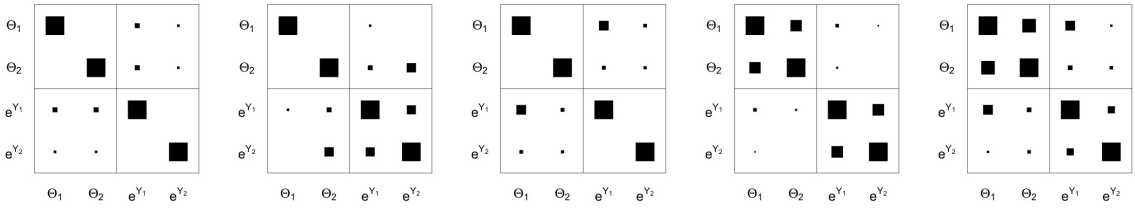

Figure 2: Correlation matrices of observed (1) and WRF-simulated (2) wind speed (e^Y) and direction (Θ) computed using circular-circular (Fisher), circular-linear (Mardia) and linear-linear (Pearson) correlation coefficients for the 5 estimated regimes in the summer season. The size of the squares is proportional to the absolute value of the correlation, empty squares indicate negative correlation.