Multivariate distributional regression forests for probabilistic nowcasting of wind profiles

Moritz N. Lang[1,2], Georg J. Mayr[2], Lisa Schlosser[1], Thorsten Simon[1,2], Reto Stauffer[1,3], Achim Zeileis[1]

[1] Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2] Department of Atmospheric and Cryospheric Science, Universität Innsbruck, Innsbruck, Austria
[3] Digital Science Center, Universität Innsbruck, Innsbruck, Austria

E-mail for correspondence: Moritz.Lang@uibk.ac.at

Abstract: This study presents statistical methods to probabilistically predict wind profiles along the approach path of an airport for one hour in advance. Accurate nowcasts of wind profiles increase safety and facilitate optimal air traffic management by timely re-routing of landing aircraft when wind direction shifts. Distributional regression trees and forests are enhanced to predict vertical wind profiles employing a multivariate normal distribution. To gain probabilistic forecasts for both wind speed and wind direction, the components of the two-dimensional Cartesian wind vector are modeled simultaneously for several height levels of a measurement tower. The resulting tree-based models can capture non-linear effects and interactions, and automatically select the relevant covariates that are associated with changes in any of the parameters of the (possibly) high-dimensional multivariate normal distribution employed. Extending the multivariate distributional regression trees to multivariate distributional regression forests can further improve the predictive performance by regularizing and smoothing the covariate effects.

Keywords: Distributional Trees; Random Forest; Multivariate Normal Distribution; Wind Profiles; Probabilistic Forecasting

1 Motivation

Statistical forecasting of numerical weather quantities has so far focused mainly on near-surface variables such as temperature, wind, and precipitation, presumably because most people are directly affected there. Accordingly, distributional regression trees and forests have already been success-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

fully applied for probabilistic rain and wind direction forecasting by accounting for appropriate univariate response distributions (Schlosser et al., 2019; Lang et al., 2020). The nowcasting task of providing vertical wind profiles for aviation forecasters and air traffic control serves as a practical real-case application to extend univariate distributional regression trees and forest to multivariate response distributions.

2 Multivariate trees and forests

Distributional regression trees (Schlosser et al., 2019) fuse distributional modeling with regression trees based on the unbiased recursive partitioning algorithms MOB (Zeileis et al., 2008) or CTree (Hothorn et al., 2006). The basic idea is to recursively partition the covariate space into (approximately) homogeneous subgroups, so that a single distributional model is sufficient to be fitted to the response in each resulting subgroup. To capture the dependence on covariates, the association between the model's scores and each available covariate is assessed using either a parameter instability test (MOB) or a permutation test (CTree). After selecting the covariate with the highest significant association as split variable (i.e., lowest significant p-value, if any), the corresponding split point is chosen within the selected covariate either by optimizing the log-likelihood (MOB) or by using a two-sample test statistic (CTree) over all possible partitions. A natural extension of (distributional) regression trees is to build ensembles or forests of such trees which can further improve the predictive performance by regularizing and stabilizing the model (Breiman, 2001).
In comparison to preceding studies using distributional regression trees and forests, this study employs distributional trees and forests for probabilistic forecasting of a multivariate response distribution. Drawing on related work for tree models of psychometric networks (Jones et al., 2019), a p-dimensional multivariate normal distribution is employed in the leaves of the trees; however, the introduced methodology is conceptually transferable to any multivariate distribution. Based on the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, the density for a single p-dimensional observation vector $\mathbf{y}_i$ is given by

$$f_{\text{MVN}}(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right).$$

In the subsequent notation we collect all parameters in a single parameter vector $\boldsymbol{\theta}$ of length $k = p + p + p(p-1)/2$. Thus, this comprises the p means from $\boldsymbol{\mu}$ and the p variances and $p(p-1)/2$ correlations, respectively, from which the covariance $\boldsymbol{\Sigma}$ can be constructed.

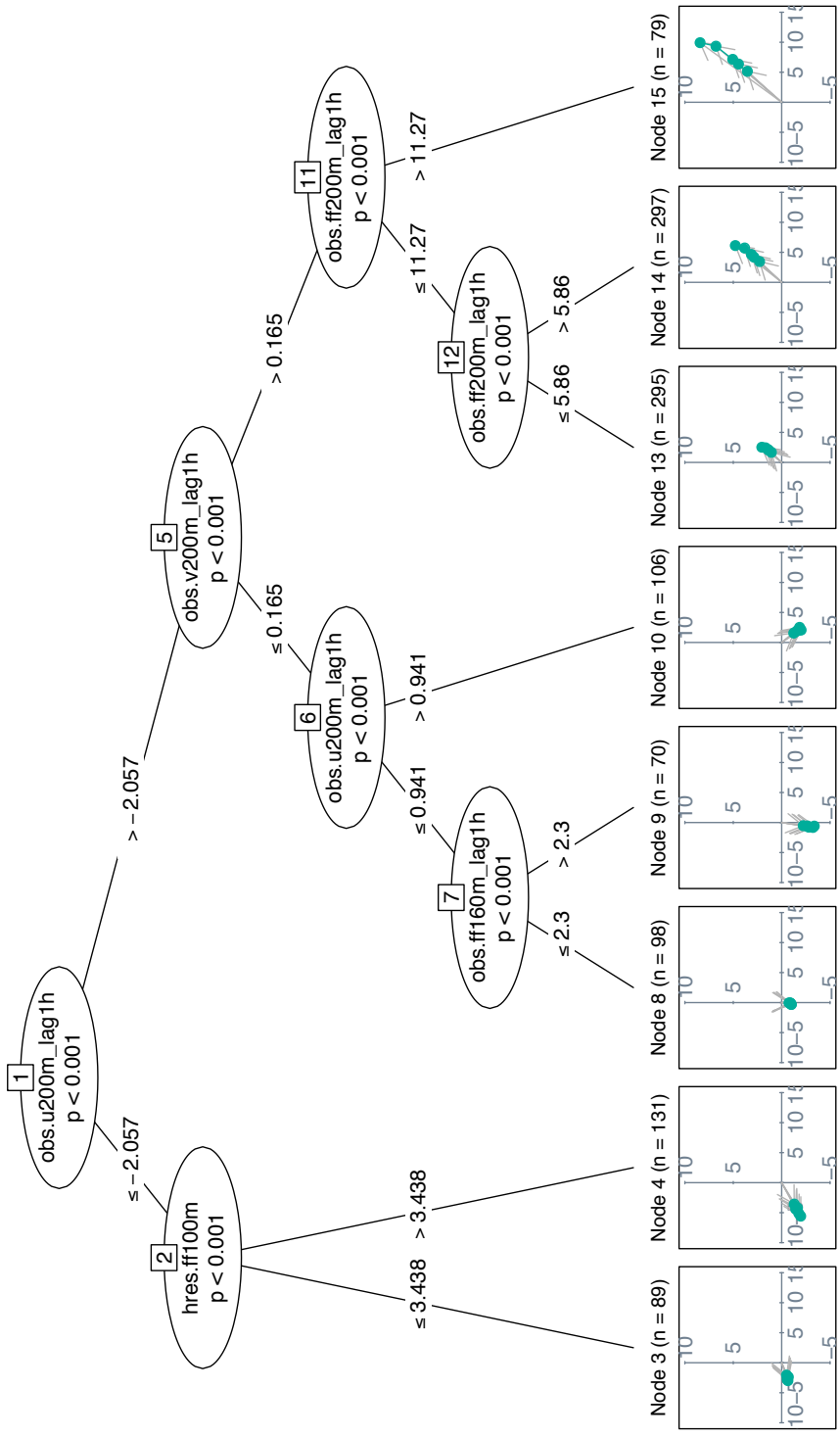

FIGURE 1. Fitted distributional regression tree based on the multivariate normal distribution for the u and v wind components at five different height levels for a measurement tower at Karlsruhe. In each terminal node, the location parameters of the wind vectors at all height levels are shown as colored points and gray arrows. The unit of the Cartesian coordinate system is in meter per second. The covariates employed are numerical high-resolution forecasts (hres), as well as 1-hourly lagged observations (obs) for wind speed (ff) and both wind vector components (u or v), all reported at different height levels (160 m, 100 m, 200 m).

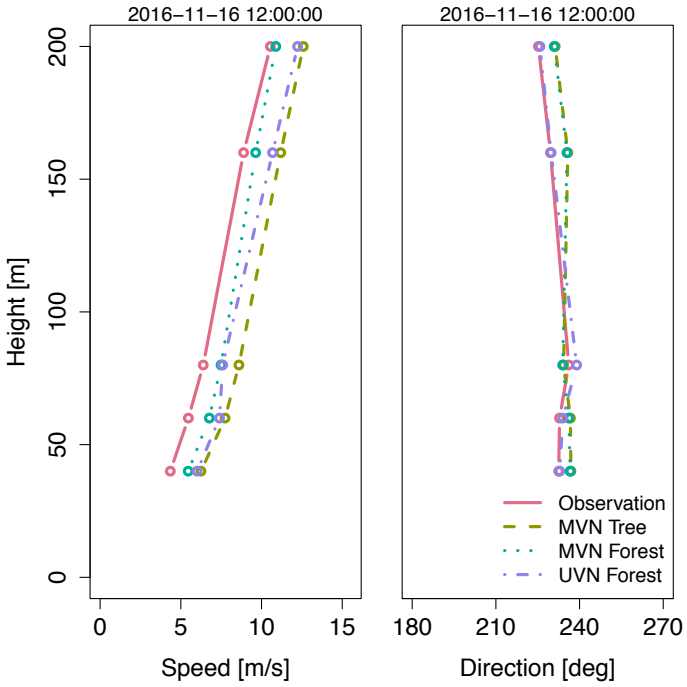

FIGURE 2. Derived wind direction and wind speed prediction at different height levels for a multivariate distributional tree and forest, as well as for an univariate distributional tree estimated per wind component and height level separately.

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the sum of the corresponding log-likelihood contributions $\ell(\boldsymbol{\theta}; \mathbf{y}_i) = \log(f_{\text{MVN}}(\mathbf{y}_i; \boldsymbol{\theta}))$ based on the n observations in a given sample. The corresponding scores $s(\boldsymbol{\theta}, \mathbf{y}_i) = (\partial_{\theta_1} \ell(\boldsymbol{\theta}; \mathbf{y}_i), \ldots, \partial_{\theta_k} \ell(\boldsymbol{\theta}; \mathbf{y}_i))$ can be employed as a general goodness-of-fit measure. Hence, evaluating the scores at the individual observations and parameter estimates $s(\hat{\boldsymbol{\theta}}, \mathbf{y}_i)$ yields an $n \times k$ matrix that assesses how well each distribution parameter estimate $\hat{\boldsymbol{\theta}}$ fits one individual observation vector $\mathbf{y}_i$. If the scores change systematically along available covariates, the parameter instabilities are incorporated into the model by maximizing a partitioned likelihood. This procedure is repeated recursively until there are no significant parameter instabilities or until another stopping criterion is met (e.g., subgroup size or tree depth).

3 Nowcasting of wind profiles

To study the performance of the novel multivariate trees and forests, 1 h predictions of vertical wind profiles for 12 UTC are issued for a measuring tower in Karlsruhe. The response has $p = 10$ dimensions, consisting of zonal (u) and meridional (v) wind components at five different height levels (40 m, 60 m, 80 m, 160 m, 200 m). Numerical weather forecasts and

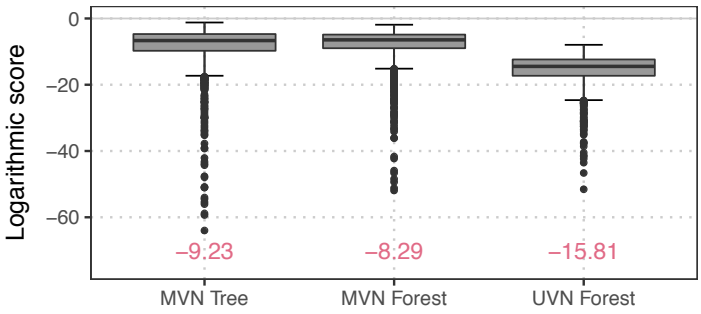

FIGURE 3. Out-of-sample predictive performance in terms of the logarithmic score based on the full predictive multivariate normal distribution for 1 h forecasts of the wind speed components at five different height levels of the measurement tower. In addition, the averaged performance over all dates is shown in red.

1-hourly lagged observations of various meteorological wind quantities are used as splitting variables, as well as derived quantities such as temporal means, minima, and maxima, or temporal and spatial differences. The terminal nodes of the tree in Fig. 1 depict the location parameters for the wind vectors at the different heights. Splits in the lagged observed u and v components (at 200 m) broadly distinguish four different regimes of wind directions: south-west (nodes 3, 4), south (nodes 8, 9), south-east (node 10), and north east (nodes 13, 14, 15). Within each regime splits in either lagged observed or predicted wind speed (ff), distinguishes low vs. high wind speeds in the same (or rather similar) directions.

To validate the estimated scale and correlation of the multivariate trees, these are compared to multivariate forests, as well as to a univariate distributional regression forest, employing the normal distribution, estimated for each wind component and height level separately. In the latter no correlation is assumed between the wind components at a single height level and between different levels. For a characteristic sample case, all three models capture the observed wind speed and direction comparably well (Fig. 2). The performance of the models, in terms of the logarithmic score, is assessed employing a yearly based four-fold cross-validation using daily data from 2014 to 2017 (Fig. 3). The box-and-whiskers show that the multivariate models outperform the univariate one, which seems to be too restrictive by the assumption of no correlation. Further, the multivariate forest is slightly superior to the multivariate tree by regularizing and smoothing the covariate effects.

The results show that the multivariate trees and forests are able to model all aspects contained in the univariate model, and further extend them by representing the correlation structure between the wind components at a single height level, as well as between the different levels. By fitting a single multivariate model for both wind components and all height levels, the

profile remains consistent and vertically coherent which allows to provide not only deterministic but rather probabilistic forecasts.

Computational details: The R package **disttree** implementing the proposed multivariate distributional regression trees and forests is available at `https://R-Forge.R-project.org/projects/partykit/`.

Acknowledgments: This project was partly funded by the Austrian Research Promotion Agency (FFG, grant no. 858537) and by the Austrian Science Fund (FWF, grant no. P31836).

References

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 1, 5 – 32.

Jones, P.J., Mair, P., Simon, T., and Zeileis, A. (2019). Network model trees. *OSF Preprints*, osf.io/ykq2a.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Stat.*, **15**, 3, 651 – 674.

Lang, M. N., Schlosser, L., Hothorn, T., Georg, J. M., Stauffer, R., and Zeileis, A. (2020). Circular Regression Trees and Forests with an Application to Probabilistic Wind Direction Forecasting. arXiv:2001.00412, *arXiv.org E-Print Archive*.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Ann. Appl. Stat.*, **13**, 1564 – 1589.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Stat.*, **17**, 2, 492 – 514.