Causal discovery among wind-related variables in a wind farm under extreme wind speed scenarios: Comparison of results using Granger causality and interactive k-means clustering

Kateřina Hlaváčková-Schindler[1], Kejsi Hoxhallari[1], Luis Caumel Morales [1], Irene Schicker[2] Claudia Plant[1]

[1] Faculty of Computer Science, University of Vienna, Vienna, [2] Geosphere Austria, Vienna, Austria

Abstract

Using the ERA5 meteorological reanalysis data from 2000 to 2020, we investigate temporal effects of ten wind related processes in time intervals of extreme wind speed values, extracted and corrected towards wind turbine locations for a wind farm in Andau, Austria. We approach the problem by two ways, by the Granger causal inference, namely by the heterogeneous Graphical Granger model (HMML) and by clustering, namely by the interactive k-means clustering (IKM). We investigate six scenarios based on the hydrological half-year, a moderate wind speed and time intervals of low or high extreme wind speed in the farm. In case of HMML, we discover causal variables and their values for each scenario. Regarding the method IKM, it is used for three clusters (clusters for a moderate wind speed and for a low and high extreme wind speed) to find coefficient representations of each interacting variable with respect to the wind speed in each of the six scenarios. We compare the results of both methods in terms of the causal variables and of the variables of the highest coefficients of representation and evaluate the interpretability of the discovered causal connections with the expert meteorological knowledge.

Heterogeneous Granger Model by Minimum Message Length

Heterogeneous Graphical Granger model (HGGM) [1], extends the multivariate Granger causality [2] for Gaussian processes to processes from exponential family. For each target variable (process) i in time t holds

$$x_i^t = \eta_i(X_{t,d}^{Lag}\beta_i')$$

where , $X_{t,d}^{Lag} = (x_1^{t-d}, .., x_1^{t-1}, .., x_p^{t-d}, .., x_p^{t-1})$ p is number of processes, and η_i an exponential link function. *One says that xj Granger–causes time series xi for a given lag d, and denote xj → xi, iff at least one of the d coefficients in j-th row of estimate of βi non-zero.* Method HMML [3] which based on the minimum message length achieves superior precision in causal inference w.r.t. baseline methods.

Wind Farm Data and HMML

Data and scenarios

Wind power production data and 10 related meteorological parameters from the ERA5 data [4] each for 38 individual wind turbines in years 2000-2020 measured hourly are considered: boundary layer height measured in m (blh), divergence at 2m in s−1 (d2m), geopotential in m²/s (z), rel. humidity in % (rel-h), temperature at 2m in K (t2m), temperature at 100m in K (t100m), temperature at 135m in K (t135m), wind direction at 100m in K (wdir100m), wind speed at 135 m in m/s (wspeed135m), wind speed at 100 m in m/s (wspeed100m). **The target variable is 96-hour wind speed time series at 135m (**hub height of the turbines). HMML was used for each turbine separately as well as the following six scenarios (at 135m) and for each year:

Table 1: Six wind speed scenarios

Hydrol. half-years in 2000	Winter	Summer
High extreme wind ≥15m/s	01-16 - 01-19	10-28 - 10-31
Low extreme wind ≤ 2m/s	01-01 - 01-04	07-22 - 07-25
Moderate (6m/s, 8m/s)	12-02 - 12-06 06	06-16 - 06-20

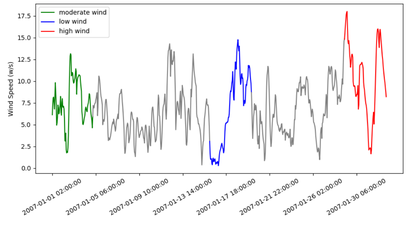

Figure 1: Wind speed in 2007. Best fitting distributions of wind speed for different scenarios best fitting exponential distribution was found by stat. fitting and used in HMML:

Experiments

To achieve a statistical validity of the causal values, for all scenarios, each turbine and each variable i, we calculated the corresponding arithmetical mean of βi-proportionality over all 21 years. βi-proportionality ~ relative strength of causality with respect to all 10 variables (all add up to 1).

The most influential variable is denoted at each turbine. Figure 3 illustrates the farm in high wind extreme scenario in summer. The values of the three most significant variables for each scenario are in Table 2.

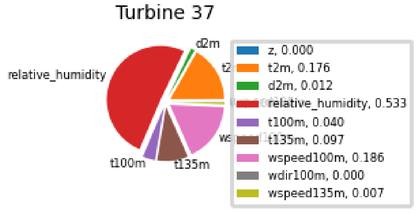

Figure 2: Interactive pie chart of one turbine

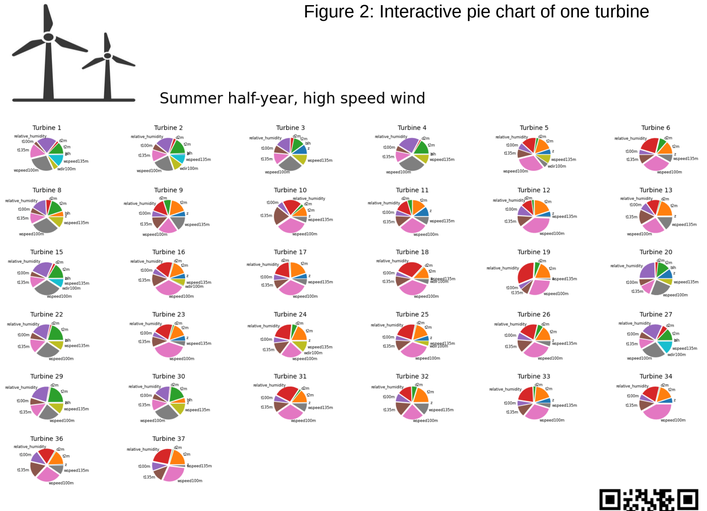

Figure 3: The farm in high wind extreme scenario in summer

Table 2: Three most significant causal variables and their average values βi-proportionality for ws135 for each scenario

Wind speed	Winter half year	Summer half year
High extreme wind	r (0.25), ws100 (0.172), t135m (0.155)	ws100m (0.302), t2m (0.203), t135m (0.123)
Low extreme wind	r (0.365), ws100 (0.206), t2m (0.105)	r (0.305), z (0.211), ws100m (0.156)
Moderate wind	ws100 (0.369), r (0.278), ws135m (0.143)	t2m (0.109), ws (0.18), t2m (0.141)

Clustering by IKM Algorithm

The Interaction K-Means algorithm, IKM [5] clusters multivariate time series. Its cluster notion is based on the interactions between the univariate time series within a data object. Objects exhibiting a similar intrinsic interaction pattern are assigned to a common cluster. For each turbine and a hydrological half-year, the knowledge of the scenario, into which wind speed belongs to, provides a label (from 1 to 3). We applied IKM for these three clusters for each hydrological half year on 10-dimensional time series of length of 96 hours. Euclidean distance was used. The obtained clustering for each hydrological half-year was assessed w.r.t to the known labeling by the clustering purity

$$CP = \frac{COUNT_M}{COUNT_O} * 100\%$$

where $Count\_M$ is the number of elements of the dominant class in a considered cluster and $Count\_O$ is the overall number of elements of that class in a dataset.

Hydrol. half-year	Cluster purity
Summer	77.40
Winter	63.78

Table 3: Three the most significant variables from coefficient matrices after clustering by IKM for each cluster for ws135 as a target variable

Wind speed at 135m	Winter: 3 most dominant variables	Summer: 3 most dominant variables
High extreme wind	d2m, z, blh	blh, widir100m, d2m
Low extreme wind	blh, z, wdir100	z, blh, wdir100
Moderate wind	z, d2m, wdir100	blh, wdir100, d2m

Table 4: Three most significant variables with the highest values of coefficients of representations for each scenario (without exact values)

Concluding, comparing the causal variables by HMML and the most dominant variables achieved by IKM clustering for all scenarios, HMML provides more plausible results.

References

[1] Behzadi, S., Hlaváčková-Schindler, K., Plant C. (2019) Granger causality for heterogeneous processes. Advances in Knowledge Discovery and Data Mining: PAKDD 2019,Springer, 463–475.
[2] Granger, C. W. (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 424–438.
[3] Hlaváčková-Schindler, K., Plant C. (2020) Heterogeneous graphical Granger causality by minimum message length. Entropy 22, 12 , 1400.
[4] https://forum.ecmwf.int/t/a-new-cds-soon-to-be-launched-expect-some-disruptions/1607
[5] Plant, C., Zherdin, A., Sorg, C., Meyer-Baese, A., Wohlschläger, A. M.(2014) Mining interaction patterns among brain regions by clustering. IEEE Transactions on Knowledge and Data Engineering, 26(9):2237–2249.

Acknowledgement: This work was funded in part within the Austrian Climate and Research Programme (ACRP) project KR19AC0K17614.