Investigating properties of statistical tests for comparing predictive performance with application to probabilistic weather forecasting

Friederike Grupe and Annette Möller

Faculty of Business Administration and Economics
Bielefeld University

26. April 2023

UNIVERSITÄT BIELEFELD
Faculty of Business Administration and Economics

Contents

1 Comparing Predictive Accuracy of Forecasts

Introduction

- **Two alternative (postprocessing) methods** are available to forecast a variable of interest (e.g. a weather quantity) over time.
- Properties of **forecast errors of both methods** can be compared after **fitting models on training data and predicting on test data**.
- **Question**: Observed difference simply due to random chance/noise in training data, or statistically significant?

$\rightsquigarrow$ Apply **formal testing procedures**

Preliminaries and Notation

- **Given**: Observations y_t and forecasts $\hat{y}_{it}$ from (postprocessing) model $i = 1, 2$, at time points $t = 1, \ldots, T$.

$\rightsquigarrow$ Resulting **forecast errors of model** i: $e_{it} = \hat{y}_{it} - y_t$.

- In many cases loss associated with forecast i **depends on forecast and observation (only) through forecast error**, that is we can conveniently write

$$L(y_t, \hat{y}_{it}) = g(\hat{y}_{it} - y_t) = g(e_{it}).$$

- General cases where loss does not collapse to $g(e_{it})$ are denoted by $L(y_t, \hat{y}_{it}) = g(\hat{y}_{it}, y_t)$.
- **Loss differential between the two forecast models** $i = 1, 2$ (resulting e.g. from two postprocessing models)

$$d_t = g(e_{1t}) - g(e_{2t}).$$

Preliminaries and Notation

- The two forecasts have **equal predictive accuracy if and only if the loss differential has zero expectation for all** t.
- Thus, we are interested in testing the **Null Hypothesis**

$$H_0 : \mathrm{E}(d_t) = 0 \quad \Leftrightarrow \quad H_0 : \mathrm{E}(g(e_{1t})) = \mathrm{E}(g(e_{2t})) \;\; \forall t$$

vs. the **Alternative Hypothesis**

$$H_1 : \mathrm{E}(d_t) \neq 0$$

$\rightsquigarrow$ Null Hypothesis of equal accuracy **equivalent to the Null Hypothesis that population mean of loss differential series is equal to 0**.

Morgan Granger Newbold (MGN, 1977) Test

- **Assumptions**:
  - Loss function is quadratic
  - Forecast errors are (a) zero mean, (b) Gaussian, (c) serially uncorrelated (no autocorrelation)
  - Additional assumption of no contemporaneous correlation of the forecast errors is supposed to be relaxed.
- **Approach**: Apply **orthogonalizing transformation to forecast errors**

$$x_t = e_{1t} + e_{2t}$$
$$y_t = e_{1t} - e_{2t}.$$

- Given above assumptions the Null Hypothesis of equal forecast accuracy is **equivalent to Null hypothesis of zero correlation between** x **and** y, that is $H_0 : \rho_{xy} = 0$ vs. $H_1 : \rho_{xy} \neq 0$, where $\rho_{xy} = Cor(x_t, y_t)$ (correlation test).

Morgan Granger Newbold (MGN, 1977) Test

- That is, with $\hat{\rho}_{xy}$ denoting the empirical Bravais-Pearson correlation coefficient the test statistic can be computed as follows:

$$MGN = \frac{\hat{\rho}_{xy}}{\sqrt{\frac{1-\hat{\rho}_{xy}^2}{T-1}}}.$$

- Under H_0 the test statistic follows a Student's t distribution with $T-1$ degrees of freedom
- **Drawback**: Application **limited to one-step ($h = 1$) ahead predictions and squared error loss**.

Diebold Mariano Test (DM, 1995)

- **Applicable to**
  - **any loss function**, that is loss can be non-quadratic, non-symmetric, non-continuous,
  - $h\,(\geq 1)$ step ahead forecasts,
  - forecast errors that are non-Gaussian, nonzero-mean, serially correlated, contemporaneously correlated
- Let $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ be the sample mean of d_t.

$$\mathrm{DM} = \sqrt{T}\,\frac{\bar{d}}{\sqrt{\sum_{\tau=-(h-1)}^{h-1} \widehat{\gamma}_d(\tau)}},$$

where the truncation lag $h-1$ refers to an h step ahead forecast, and

$$\widehat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T} (d(t) - \bar{d})(d(t-|\tau|) - \bar{d})$$

are the empirical autocovariances.

- Under H_0, test statistic approximates a standard normal distr.

Modified Diebold Mariano Test (HLN, 1997)

- Simulations in Diebold and Mariano (1995) show: **normal distribution can be poor approximation** of the DM test's finite-sample distribution under H_0.
- Results indicate **DM test tends to be oversized**, depending on degree of autocorrelation among forecast errors and sample size T.
- Harvey, Leybourne and Newbold (1997) suggest that **improved small-sample properties** can be obtained by:
  - bias correcting the DM test statistic to have approx. unbiased estimate of variance of loss differential, and
  - comparing the corrected statistic with Student-t distribution with $(T - 1)$ degrees of freedom, rather than standard normal.

Modified Diebold Mariano Test (HLN, 1997)

- **Resulting modified statistic**:

$$\mathrm{DM}_{HLN} = \left(\frac{T + 1 - 2h + T^{-1} h(h-1)}{T} \right)^{\frac{1}{2}} \mathrm{DM},$$

  where DM denotes the original statistic, h denotes the number of lags, and T is the number of time points in the sample.

- **Additional requirement**: Loss differential d_t of h-step ahead forecasts is assumed to have zero autocorrelations at lag h and beyond

Aim of this Study

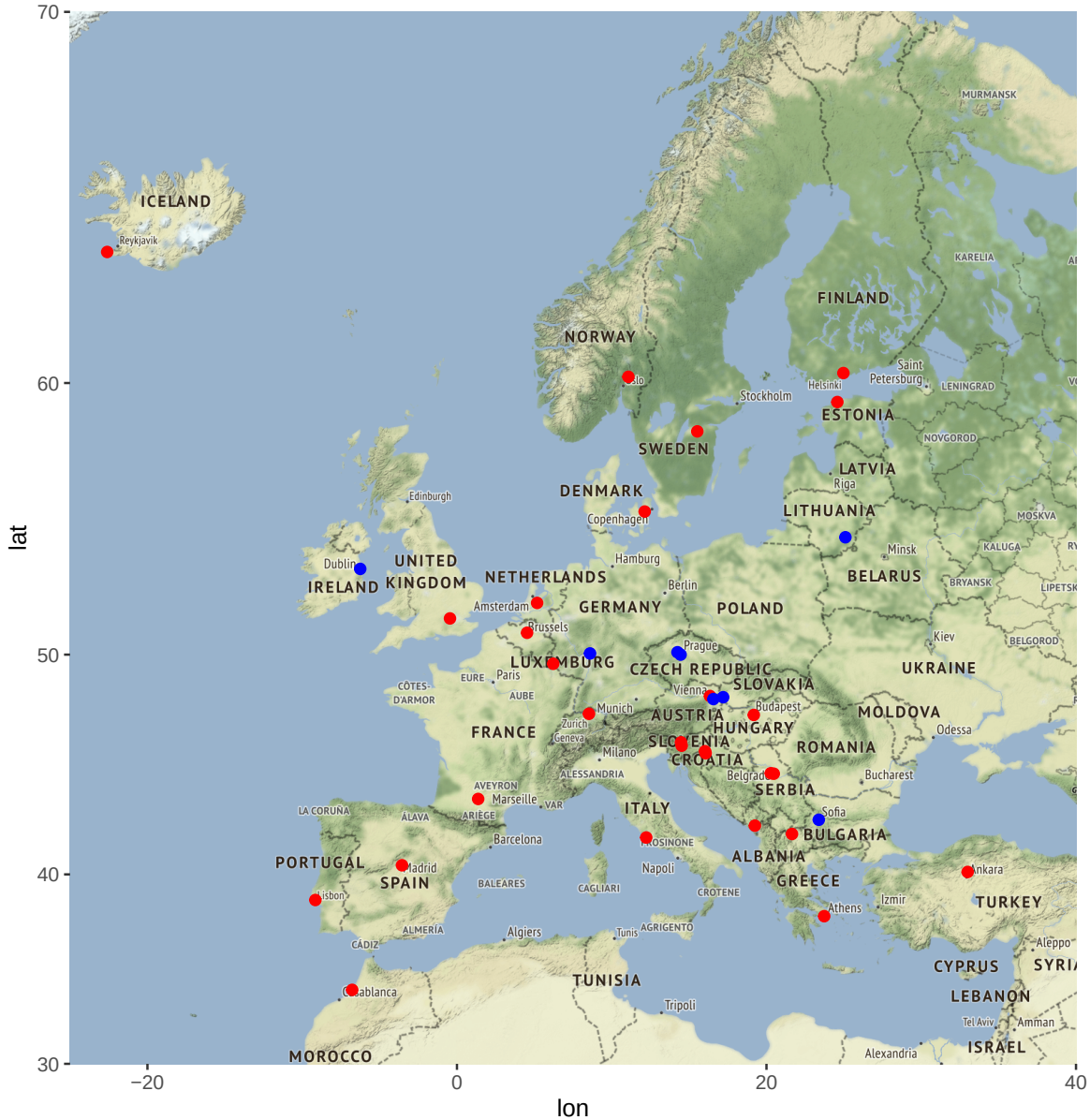
- DM test frequently used to investigate whether difference in performance of two postprocessing models is significant.
- However, up to now no **systematic investigation of behaviour of test in context of postprocessing**.
- This study (and future extensions) is interested in the following aspects:
  - use of different **loss functions** (CRPS, MSE, MAE,...)
  - use of different **forecast/postprocessing models**
  - application to different **forecasts horizons**
  - application of different **versions/types of tests**
  - **assumptions of the tests**: approx. fulfilled on real data? if not, tests robust against it?
- Some of the aspects were analysed in the current study, a few exemplary results are presented here.

2 Application to Temperature Data

Data Overview

- **ECMWF temperature forecasts** and observations over **Europe**, at **36 stations**
- Time period of 12 years, from 2002-01-01 to 2014-03-20
- Last **1000 days** fixed as test/validation data, consists of **dates between 2011-06-25 and 2014-03-20**
- Forecast ensemble with **52 members**, 50 exchangeable members, 1 control forecast, 1 high-resolution forecast
- Investigation of **24-h, 48-h, 120-h, and 240-h ahead forecasts**, initialized 12 UTC (13 Uhr (1 pm), 14 Uhr (2 pm) during daylight savings time)
- More **detailed analysis for 8 stations** (having few NA): Dublin, Frankfurt, Vienna, Prague east, Prague south, Bratislava, Sofia, Vilnius.

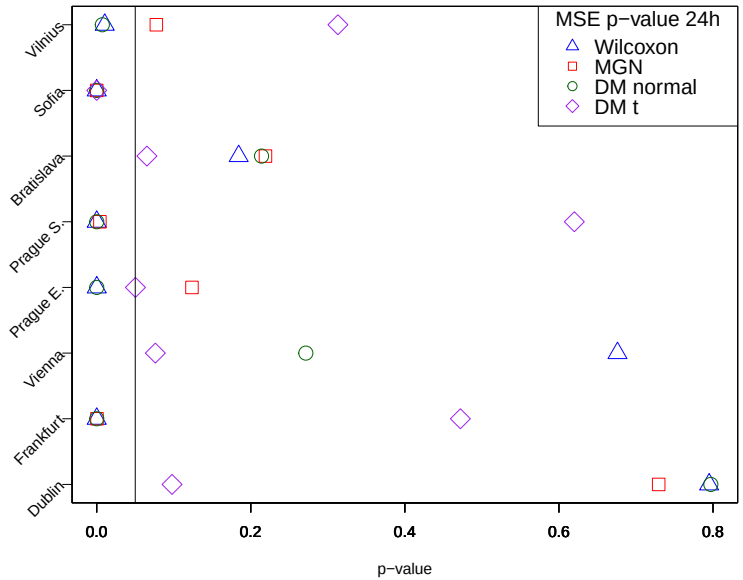
Data Overview

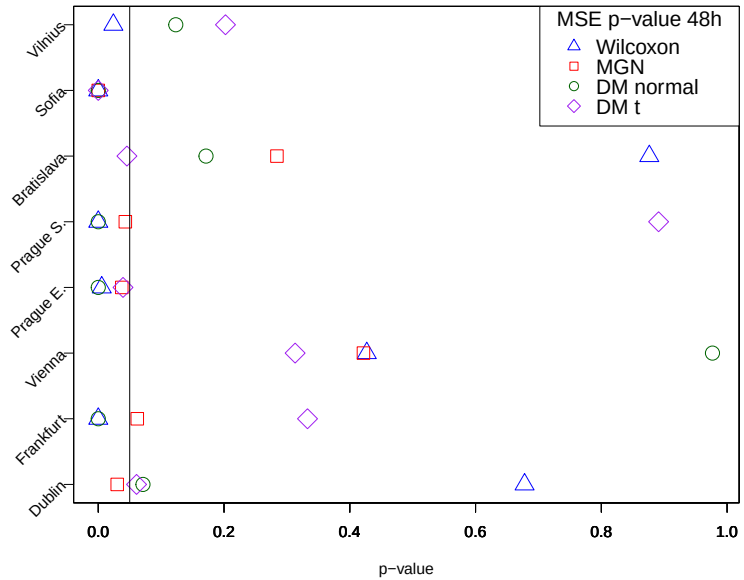


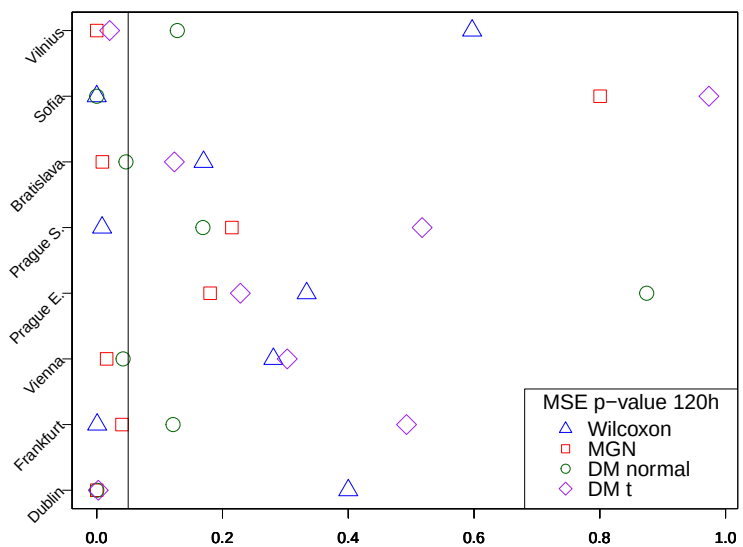
Set-up of Analysis

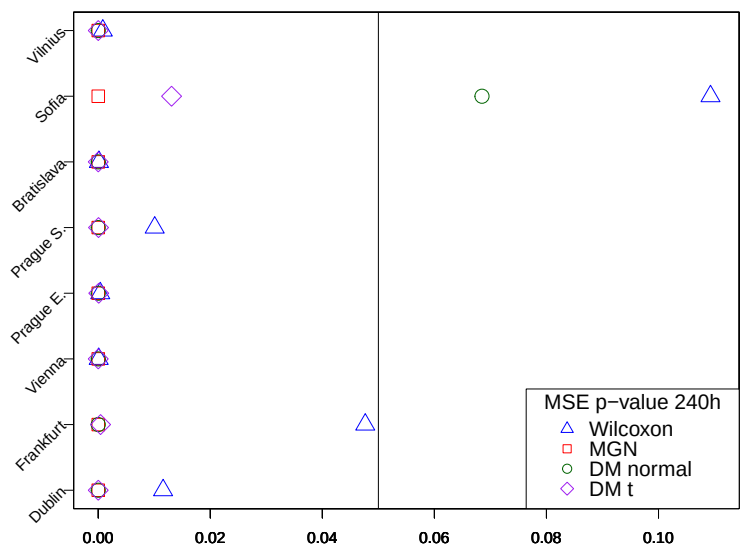
- **Benchmark forecast model is the raw ensemble**
  - computation of MAE and MSE (raw ensemble mean) and CRPS of raw ensemble for the 1000 test dates
  - results in time series of scores $s_1(t)$, $t = 1, \ldots, 1000$.
- **EMOS fitted with rolling training period on training data**
  - computation of scores MAE, MSE (EMOS predicitve mean) and CRPS for the 1000 test dates
  - results in time series of scores $s_2(t)$, $t = 1, \ldots, 1000$.
- Tests applied to $d_t = s_1(t) - s_2(t)$, thus testing $H_0 : d_t =$ Scores Raw Ens - Score Emos equal to 0 on average.

p-values MSE

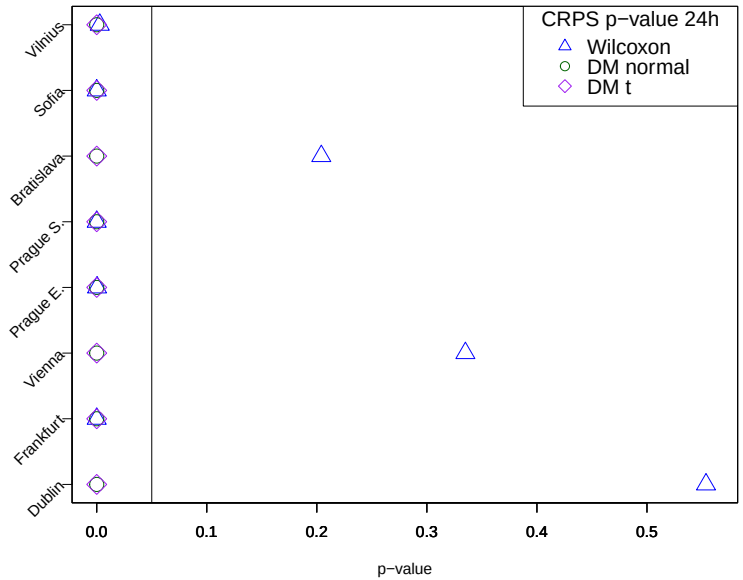

MSE p−value 24h
Wilcoxon
MGN
DM normal
DM t
Vilnius
Sofia
Bratislava
Prague S.
Prague E.
Vienna
Frankfurt
Dublin
0.0
0.2
0.4
0.6
0.8
p−value

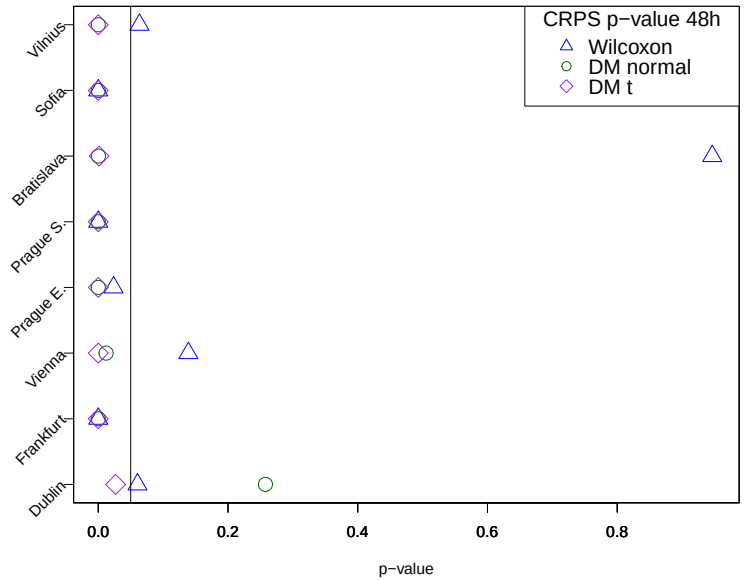

MSE p−value 48h
Wilcoxon
MGN
DM normal
DM t
Vilnius
Sofia
Bratislava
Prague S.
Prague E.
Vienna
Frankfurt
Dublin
0.0
0.2
0.4
0.6
0.8
1.0
p−value

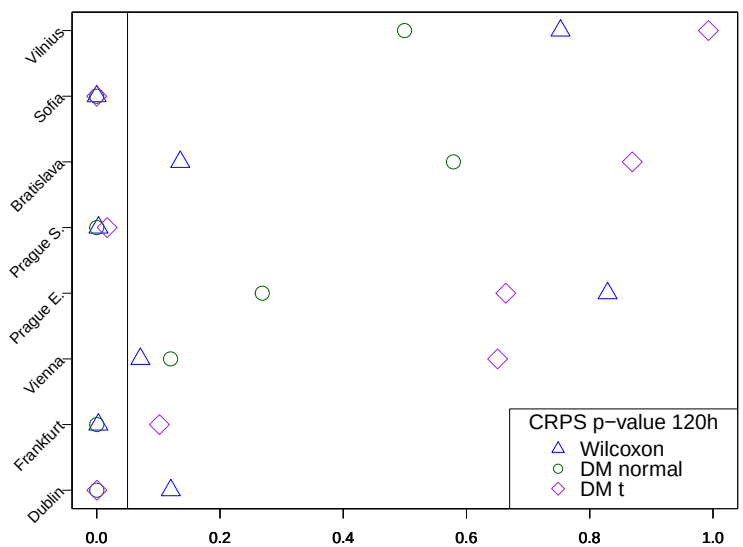

MSE p−value 120h
Wilcoxon
MGN
DM normal
DM t
Vilnius
Sofia
Bratislava
Prague S.
Prague E.
Vienna
Frankfurt
Dublin
0.0
0.2
0.4
0.6
0.8
1.0

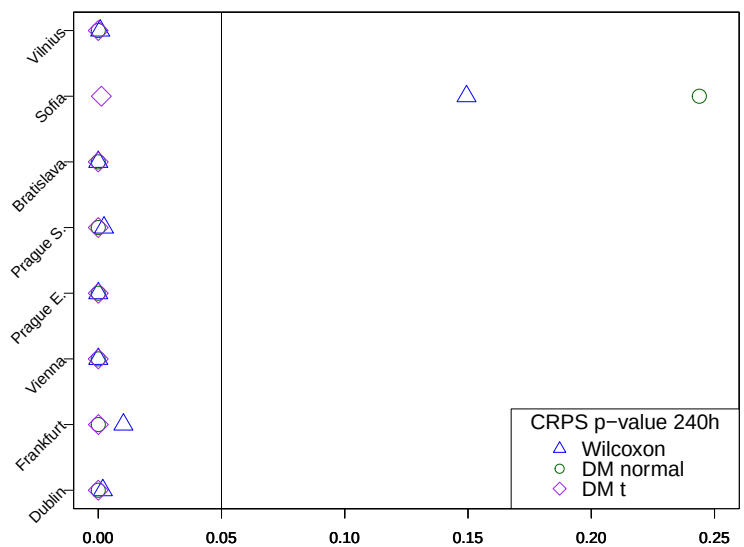

MSE p−value 240h
Wilcoxon
MGN
DM normal
DM t
Vilnius
Sofia
Bratislava
Prague S.
Prague E.
Vienna
Frankfurt
Dublin
0.00
0.02
0.04
0.06
0.08
0.10

p-values CRPS









Checking Assumptions of Tests

- **Forecast errors normal with zero mean** (MGN test)
  - EMOS: no obvious violation
  - Raw ensemble: tendency for a slight location shift to the right, otherwise roughly fine
  - For both methods no substantial differences between forecast horizons, however, differences occur between stations
- **Forecast errors show no autocorrelation** (MGN test)
  - EMOS: Autocorrelations visible in ACF (at most) up to lag $h - 1$ for h-step ahead forecasts
  - Raw ensemble: Often stronger autocorrelations visible, even for lags (way) beyond $h - 1$
  - Ljung-Box test: confirms significant autocorrelation for all stations and forecast horizons
  - Violation of assumption seems to have little effect on behaviour of MGN test.

Checking Assumptions of Tests

- **Autocorrelation of d_t zero beyond lag $h-1$ for h-step ahead forecasts** (modified DM)
  - In most cases ACF of d_t shows no substantial autocorrelation beyond lag $h-1$, across all considered scores and stations
  - In the few cases where autocorrelation beyond lag $h-1$ is present, decision of modified DM test nonetheless consistent with decisions of other tests not making this assumption
- d_t **has symmetric distribution and zero mean** (Wilcoxon)
  - Seems roughly fulfilled for all scores and forecast horizons, no instances with drastic departures from assumptions
  - MSE: tendency of heavy left tail, gets more pronounced for higher forecast horizons (and at some stations specifically)

3 Simulation Study

Simulated Data

- DM (1991, 1995) and HLN (1997) conducted simulations with respect to sample size T, distribution and correlation of forecast errors.
- Here, simulation studies designed for postprocessing setting.
- Simulated data consists of one observation and 50 ensemble members for each date, ensemble forecasts understood as one-step ahead forecasts.
- Observations drawn from a normal distribution with parameters $\mu = 0$ and $\sigma^2 = \rho_1$:

$$y \sim N(0, \rho_1)$$

Ensemble members drawn from the normal distribution:

$$X_i \sim N(\epsilon, \rho_2)$$

For the parameters, let $\epsilon \in \{0, 0.1\}$, $\rho_1 \in \{0.1, 0.25\}$, and $\rho_2 \in \{0.177, 0.3\}$ with the restriction $\rho_1 \neq \rho_2$.

Simulated Data

- $T = 2000$ **observations drawn from a normal distribution** with predefined parameters.
- For each of the $m = 50$ ensemble members, $T = 2000$ **forecasts are drawn from a normal distribution** with some parameters changed compared to the observation distribution.
- First 1500 instances used as training data, last 500 instances as test data
- Compute raw ensemble mean (Forecast Model 1), fit EMOS model based on simulated training data (Forecast Model 2).
- Again $s_1(t)$ series of scores for raw ensemble, and $s_2(t)$ series of scores for EMOS model, and tests applied to $d_t = s_1(t) - s_2(t)$.
- Data simulation, fitting of forecast models on training and prediction on test data **repeated** 100 **times**.

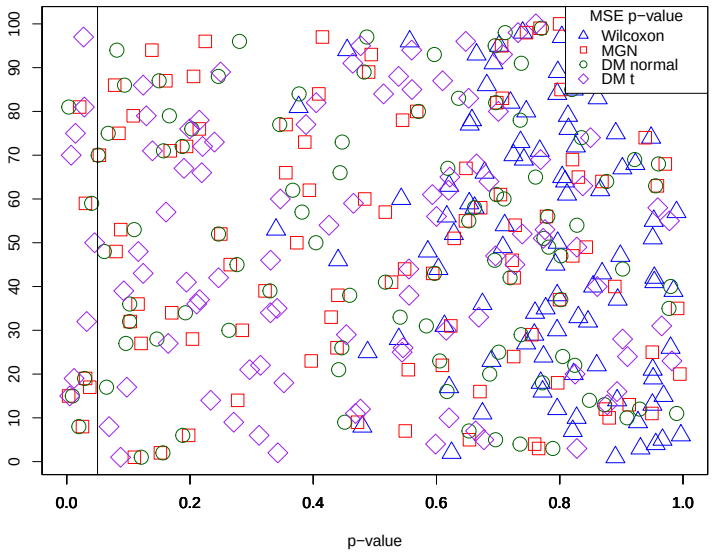
Simulation Settings

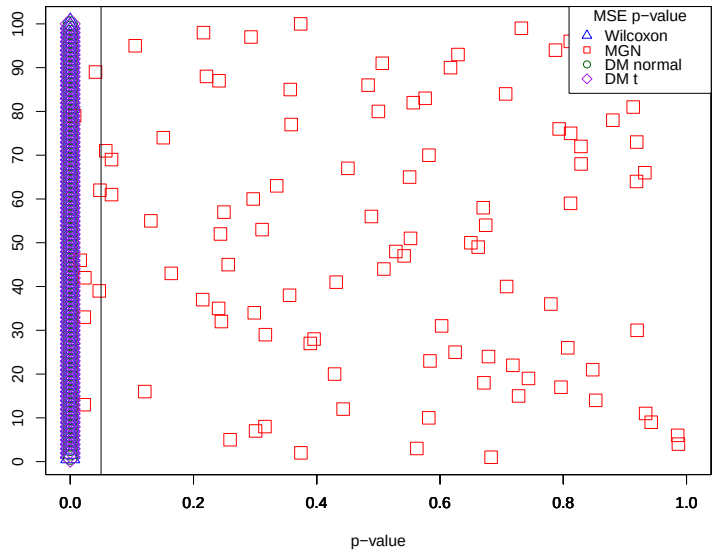
- We consider 3 different parameter combinations corresponding to 3 scenarios:
  1. **Ensemble exhibits no bias, no or little improvement expected by postprocessing.**
     Achieved with setting $\epsilon = 0$, $\rho_1 = 0.1$, and $\rho_2 = 0.177$:

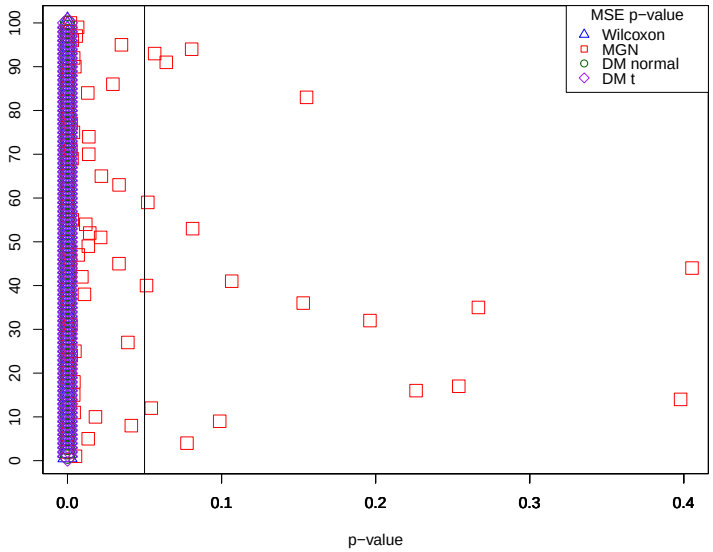
     $$y \sim N(0, 0.1) \text{ and } X_i \sim N(0, 0.177)$$

  2. **Ensemble exhibits bias, improvement by postprocessing expected.**
     Achieved by keeping distribution for simulated observations, but draw ensemble forecasts from normal distribution with parameters $\epsilon = 0.1$ and $\rho_2 = 0.177$.
  3. **Ensemble exhibits bias and dispersion errors, improvement by postprocessing expected.**
     Observations are again distributed as before, ensemble generated with a bias and a larger spread by setting parameters of the normal distribution to $\epsilon = 0.1$ and $\rho_2 = 0.3$.

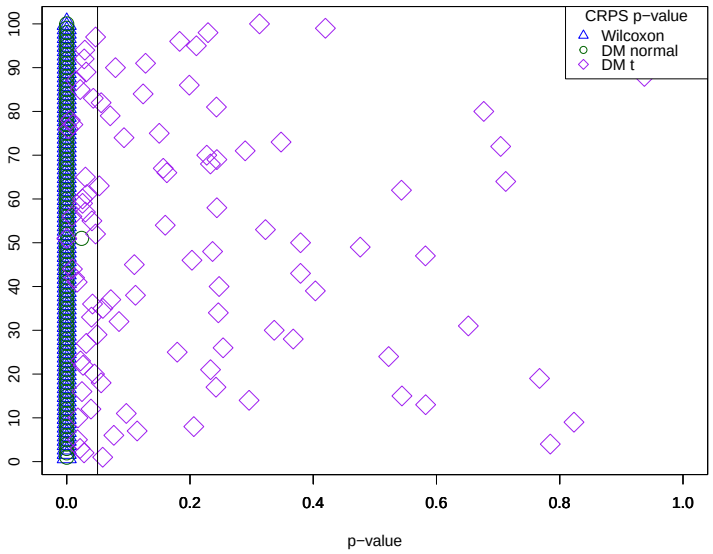
p-values of MSE Setting 1, 2, 3

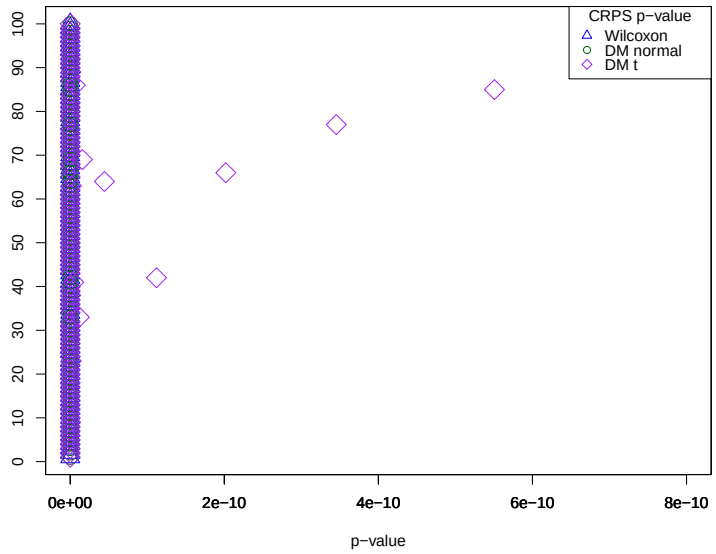


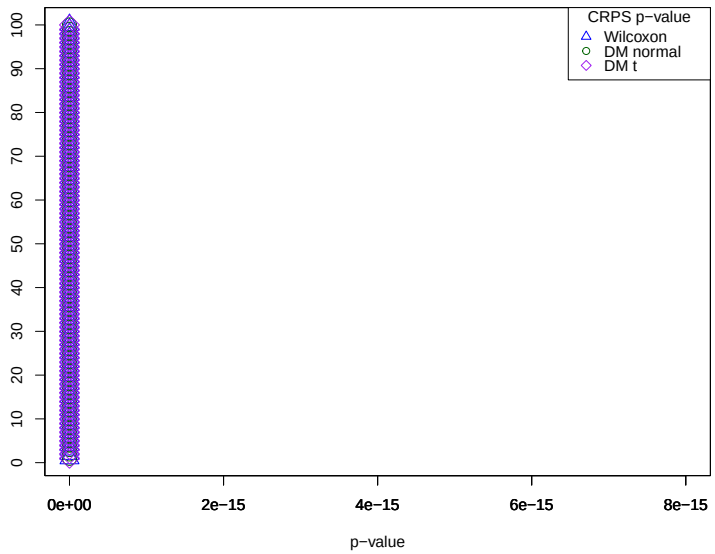




p-values of CRPS Setting 1, 2, 3







4 Outlook

Future Plans

- Investigation for other weather variables
  - assumptions of tests fulfilled as well?
  - similar performance of tests?
  - similar behaviour with respect to applied score and forecast horizon?
- Extension of simulation study
  - More systematic study of power and size of test, in conjunction with sample size
  - Simulating more/different aspects of miss-specification in postprocessing setting, of violation of assumptions for the tests

Thank you for your attention!

References

Gneiting, T. and Katzfuss, M. (2014)

Probabilistic Forecasting.

Annual Review of Statistics and Its Application 1, 125 – 151.

Diebold, F. and Mariano, R. (1991)

Comparing Predictive Accuracy I: An Asymptotic Test.

Discussion Paper 52, Institute for Empircal Macroeconomics.

Diebold, F. and Mariano, R. (1995)

Comparing Predictive Accuracy.

Journal of Business and Economic Statistics 13, 253 – 263.

Clark, T. E. (1999)

Finite-sample properties of tests for equal forecast accuracy.

Journal of Forecasting 18, 489 – 504.

Harvey, D., Leybourne, S. and Newbold. P (1997)

Testing the equality of prediction mean squared errors.

International Journal of Forecasting 13, 281 – 291.

Döhrn, R. (2019)

Comparing forecast accuracy in small samples.

Ruhr Economic Papers 833, DOI: 10.4419/86788966.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. and Graeter. M. (2020)

Simulation-based comparison of multivariate ensemble post-processing methods.

Nonlinear Processes in Geophysics 27, 349 – 371.