Barbara Casati
June 2009
FMI

Verification of continuous predictands

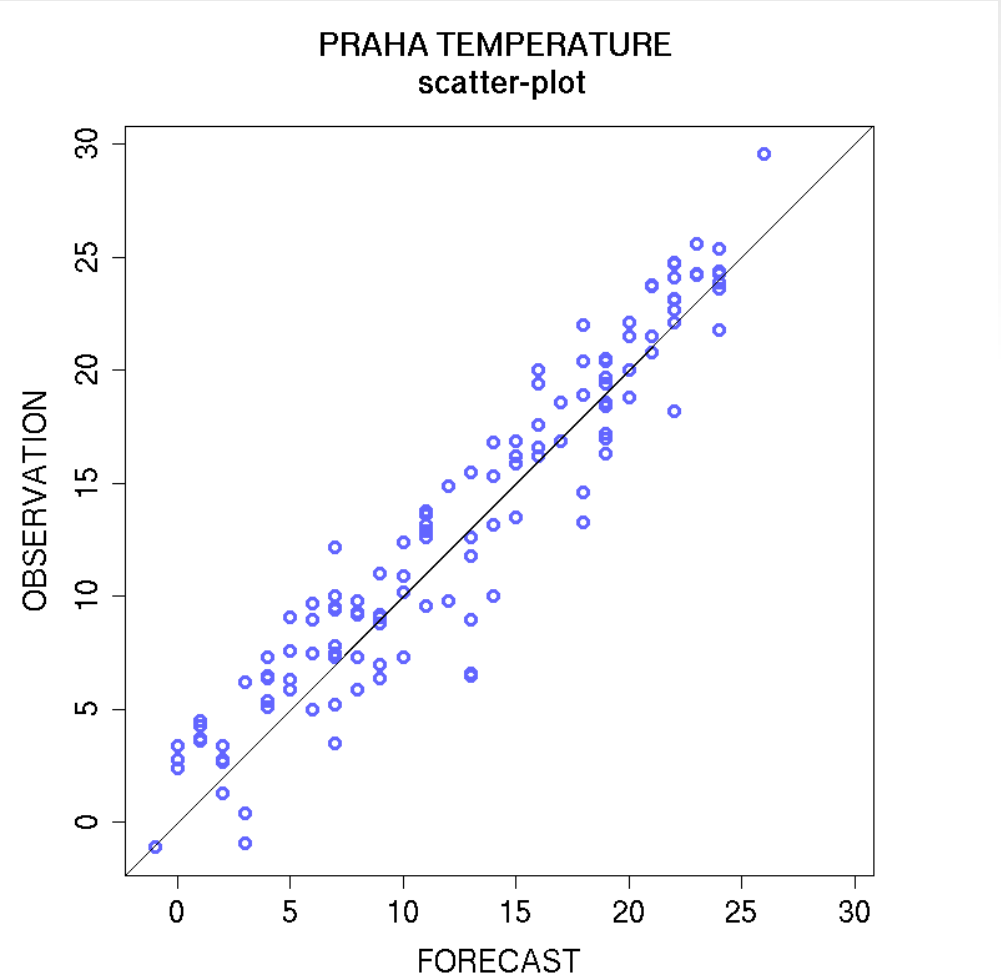
b.casati@gmail.com

Exploratory methods: joint distribution

Scatter-plot: plot of observation versus forecast values

Perfect forecast = obs, points should be on the 45^{o} diagonal

Provides information on: bias, outliers, error magnitude, linear association, peculiar behaviours in extremes, misses and false alarms (link to contingency table)

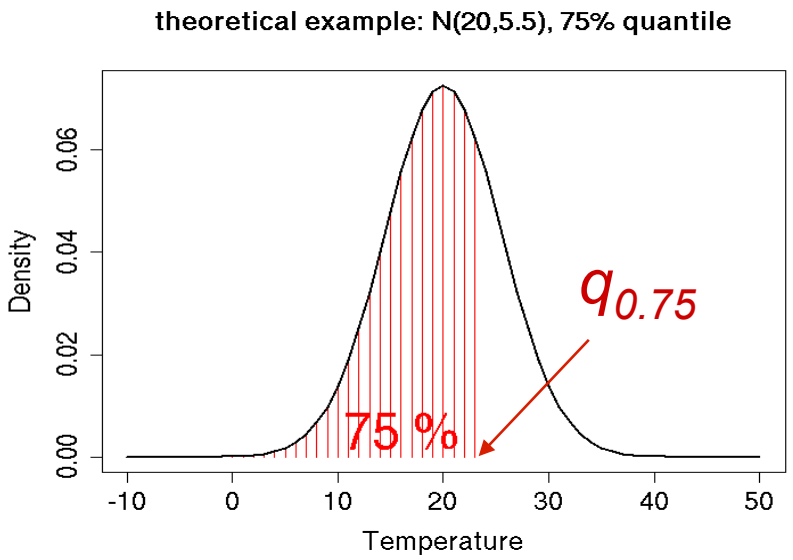


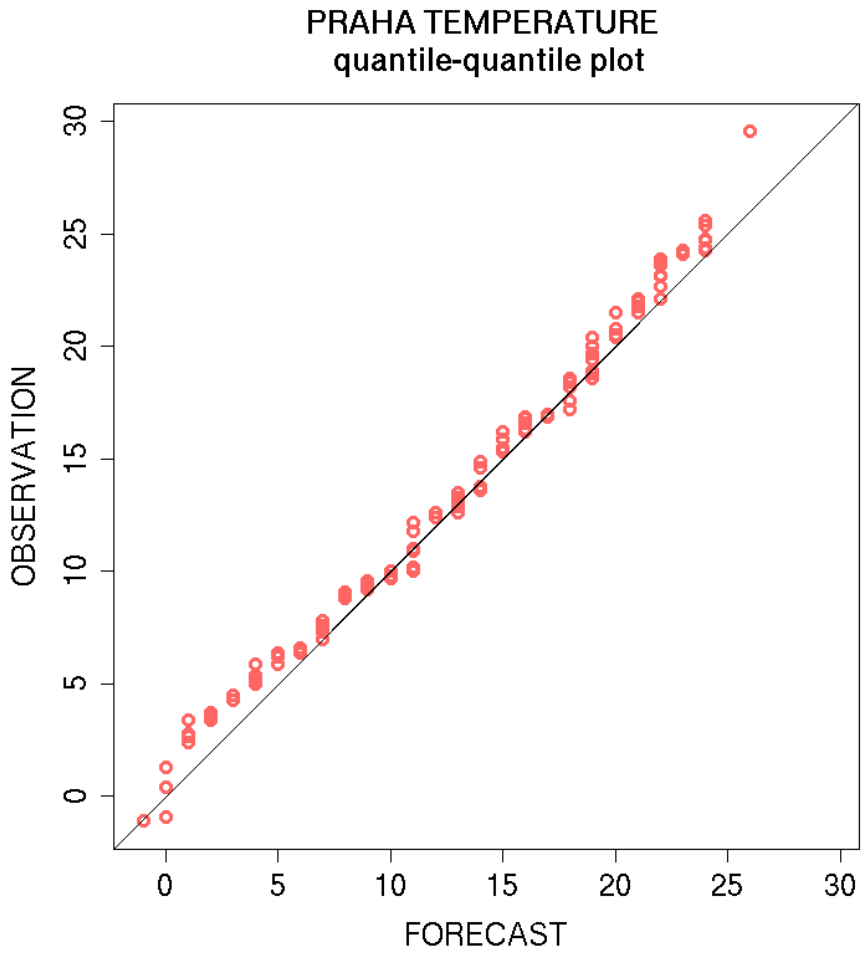
Exploratory methods: marginal distribution

Quantile-quantile plots:

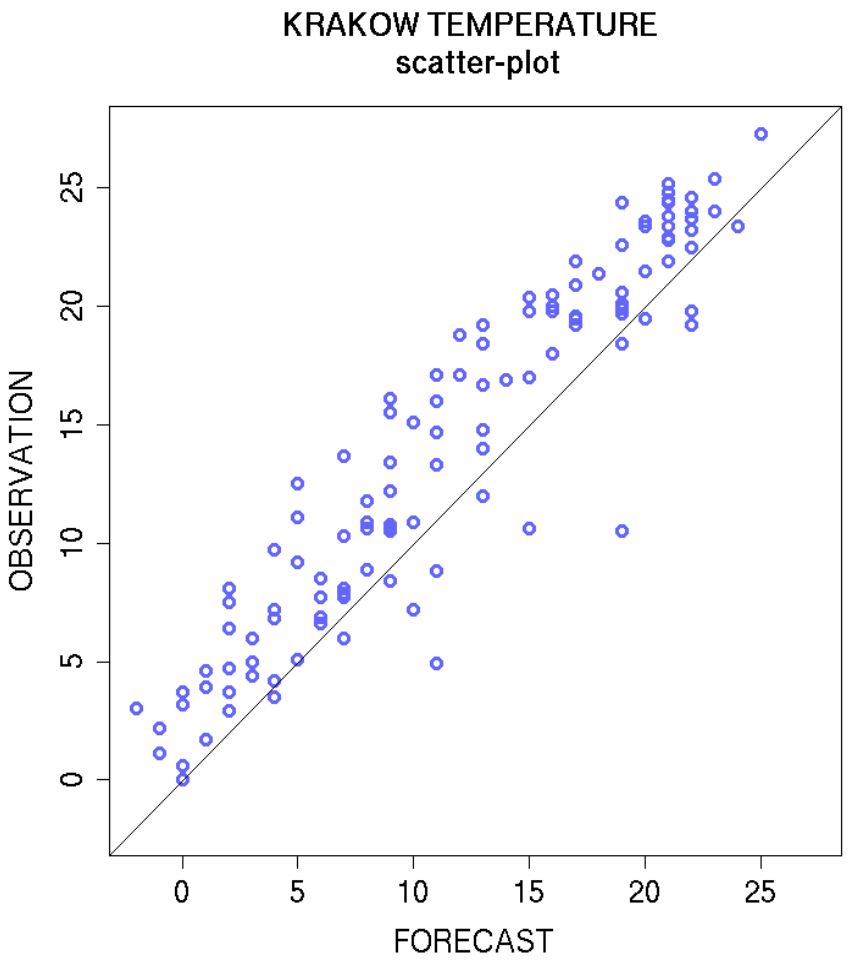
OBS quantile versus the corresponding FRCS quantile

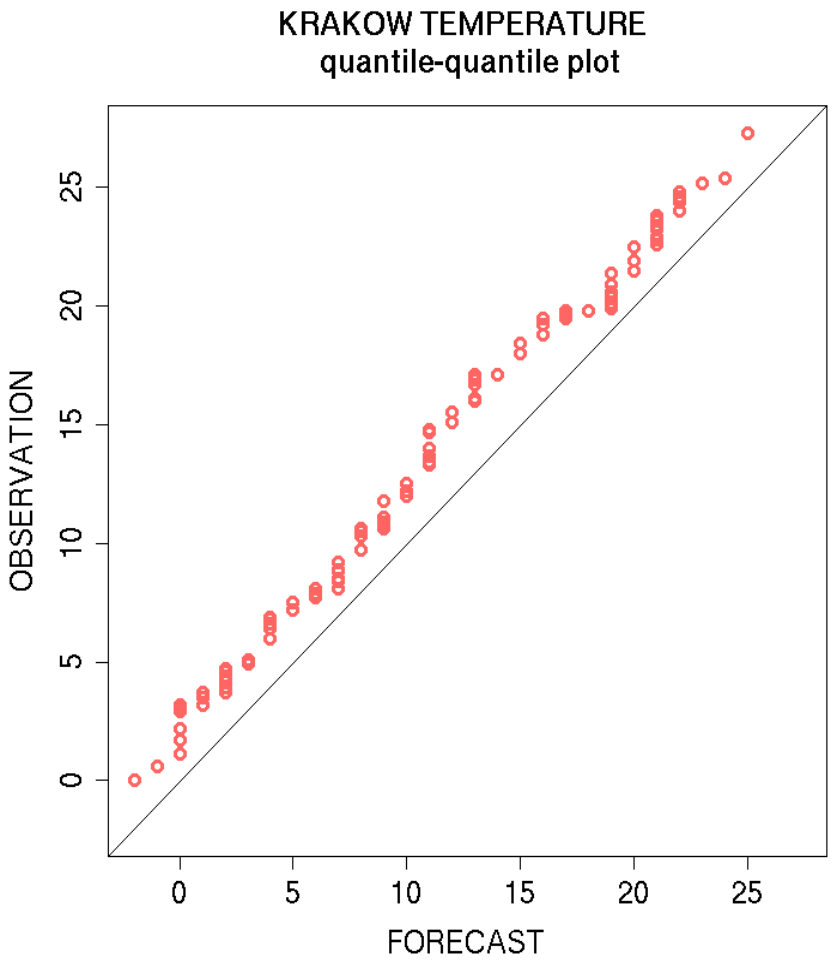
Perfect: FCST=OBS, points should be on the 45^{o} diagonal





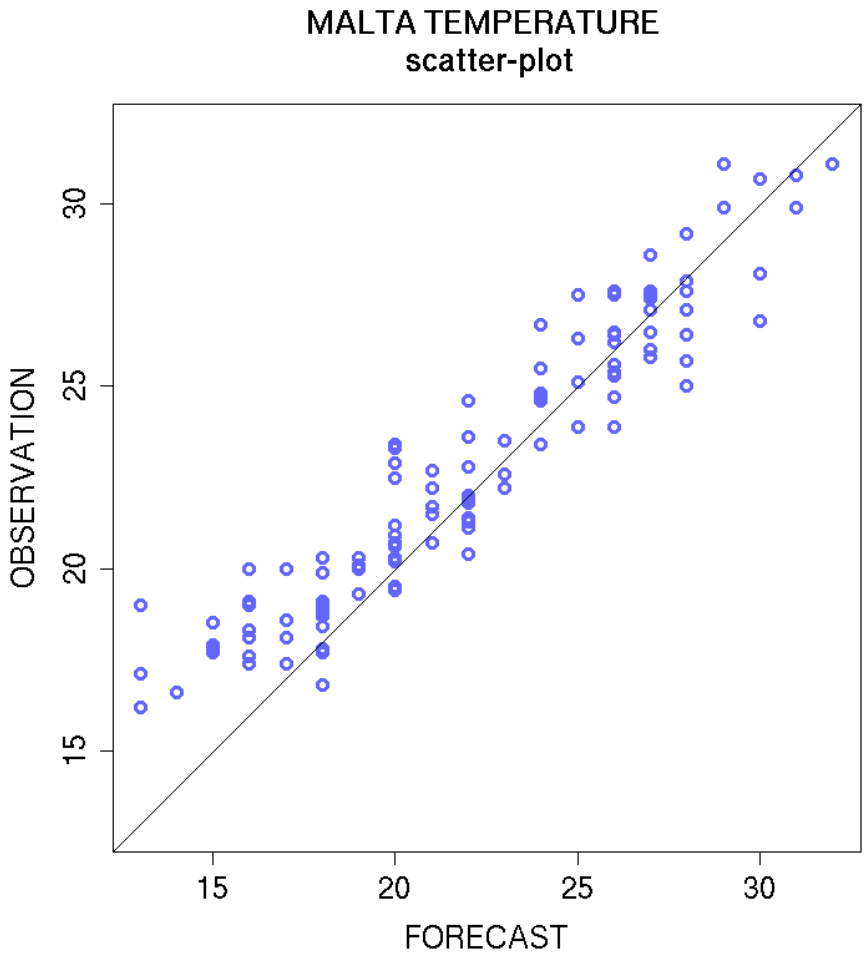
Scatter-plot and qq-plot: example 1
Q: is there any bias? Positive (over-forecast) or negative (under-forecast)?

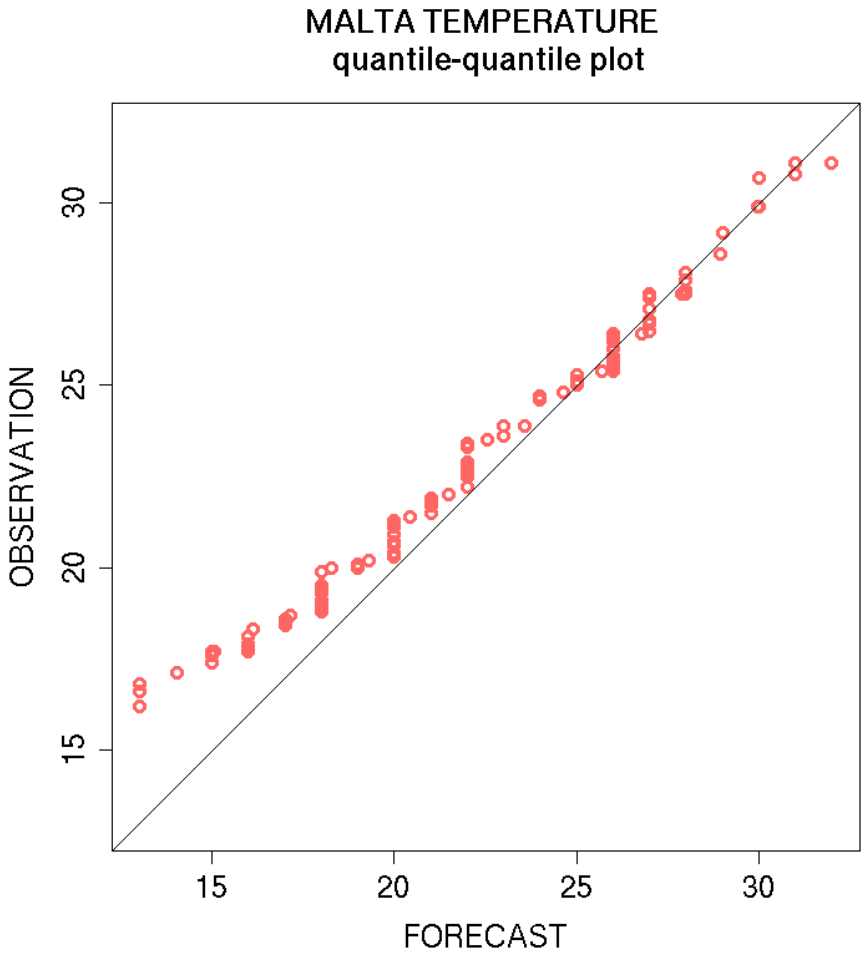




Scatter-plot and qq-plot: example 2

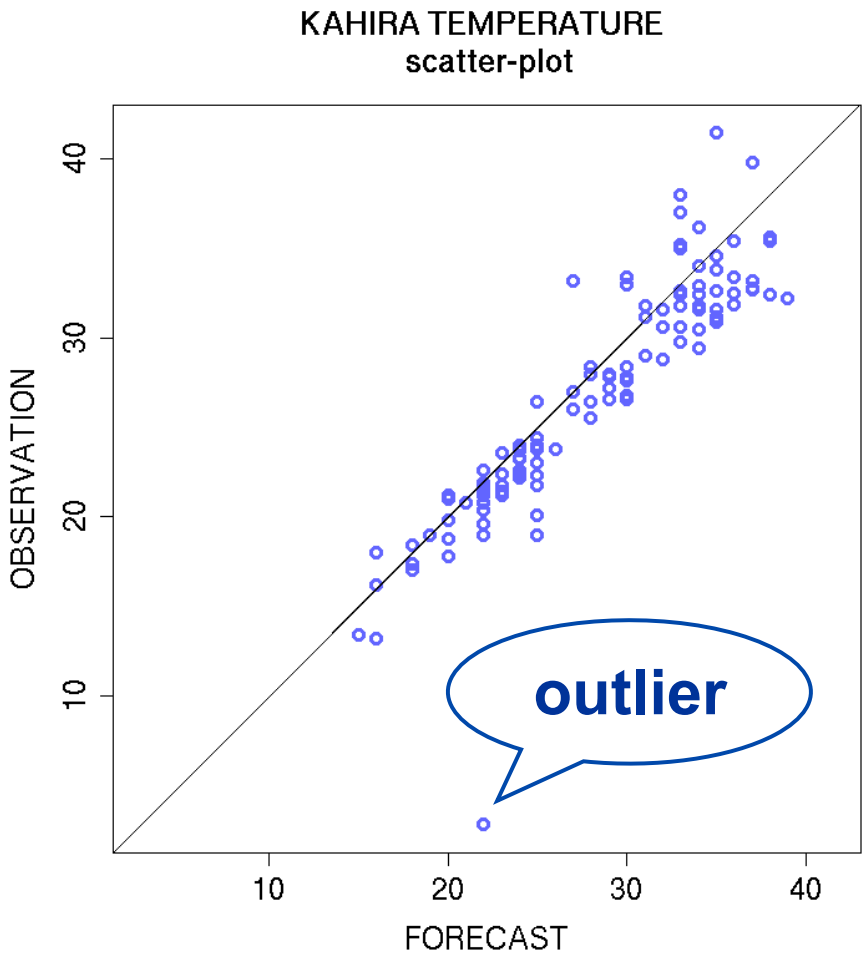
Describe the peculiar behaviour of low temperatures

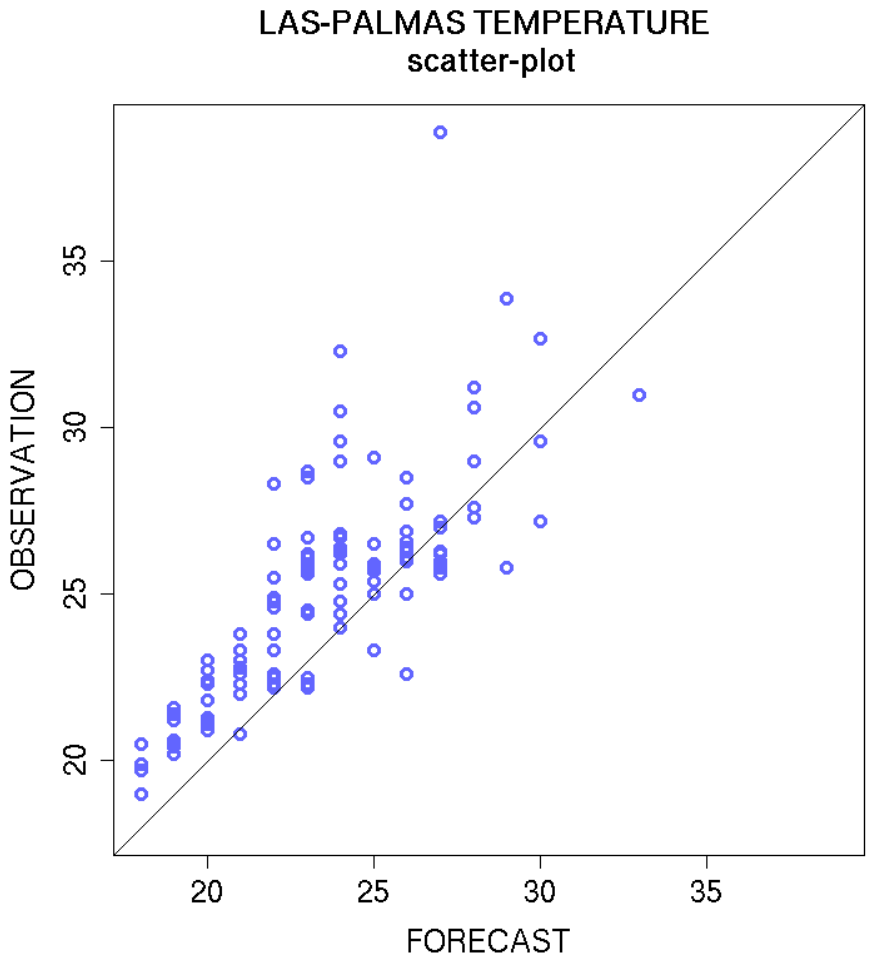




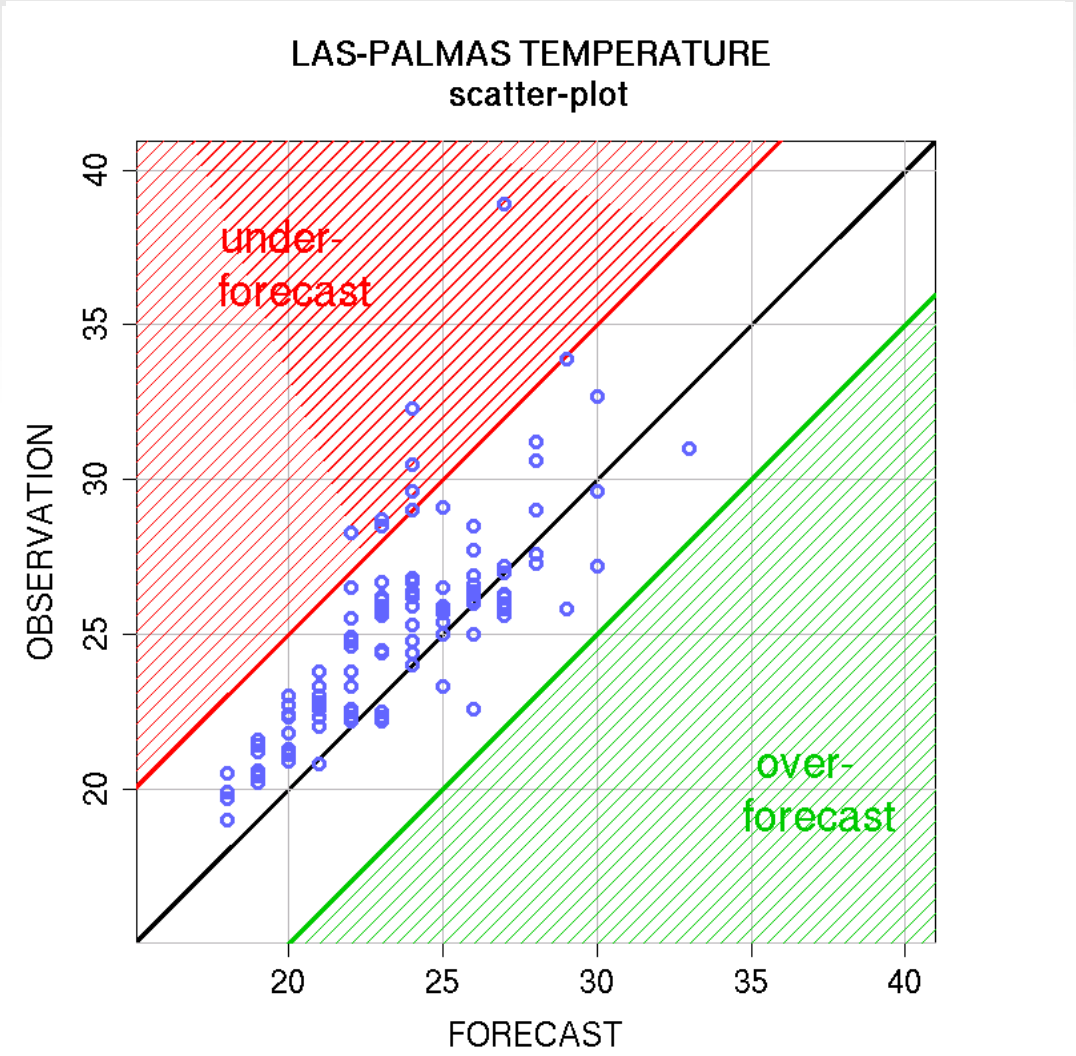
Scatter-plot: example 3

Describe how the error varies as the temperatures grow





Scatter-plot: example 4
Quantify the error



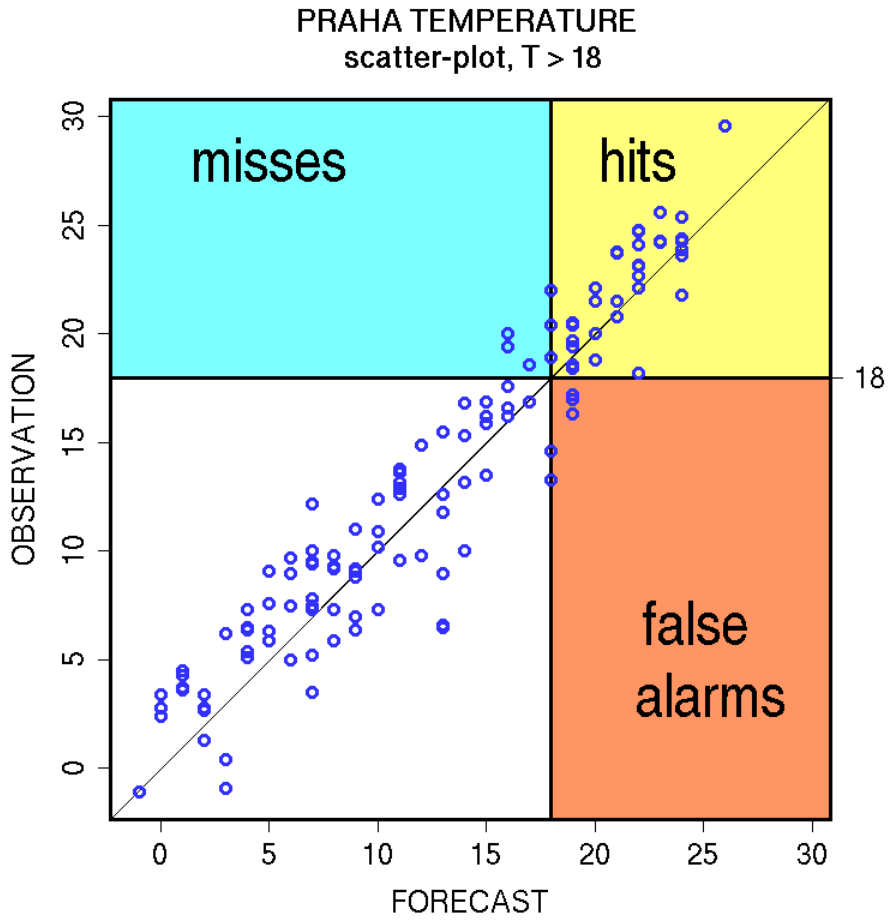
Q: how many forecasts exhibit an error larger than 10 degrees ?

Q: How many forecasts exhibit an error larger than 5 degrees ?

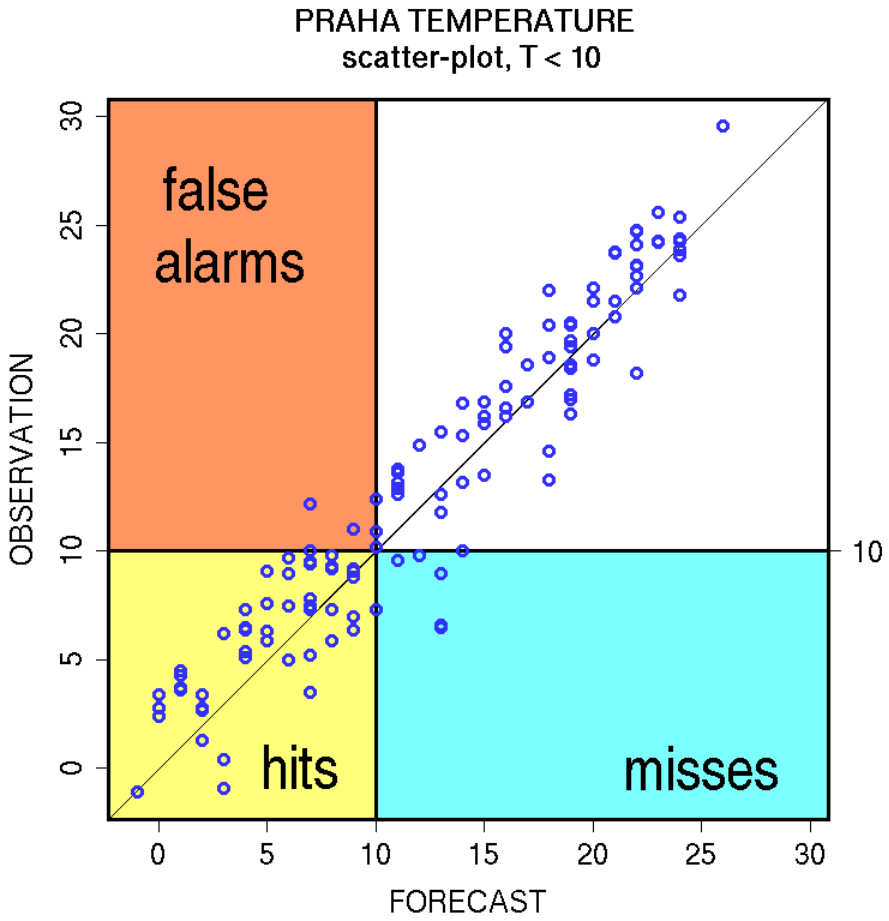
Q: Is the forecast error due mainly to an under-forecast or an over-forecast ?

Scatter-plot and Contingency Table

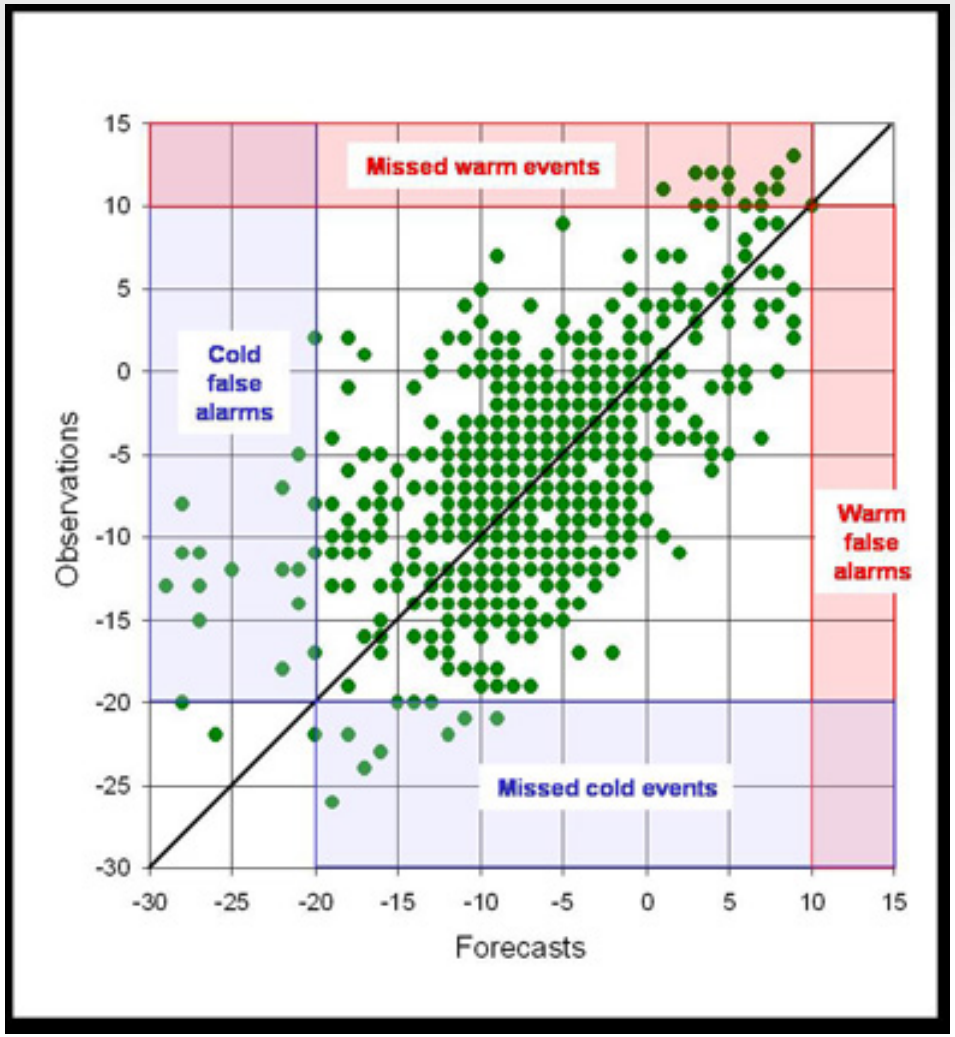
Does the forecast detect correctly temperatures above 18 degrees ?



Does the forecast detect correctly temperatures below 10 degrees ?



Scatter-plot and Cont. Table: example 5
Analysis of the extreme behavior



Q: How does the forecast handle the **temperatures above 10 degrees** ?

- How many misses ?
- How many False Alarms ?
- Is there an under- or over-forecast of temperatures larger than 10 degrees ?

Q: How does the forecast handle the **temperatures below -20 degrees** ?

- How many misses ?
- Are there more missed cold events or false alarms cold events ?
- How does the forecast minimum temperature compare with the observed minimum temperature ?

Exploratory methods: marginal distributions

Visual comparison:
Histograms, box-plots, …

Summary statistics:

- ***Location:***

$$\text{mean} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

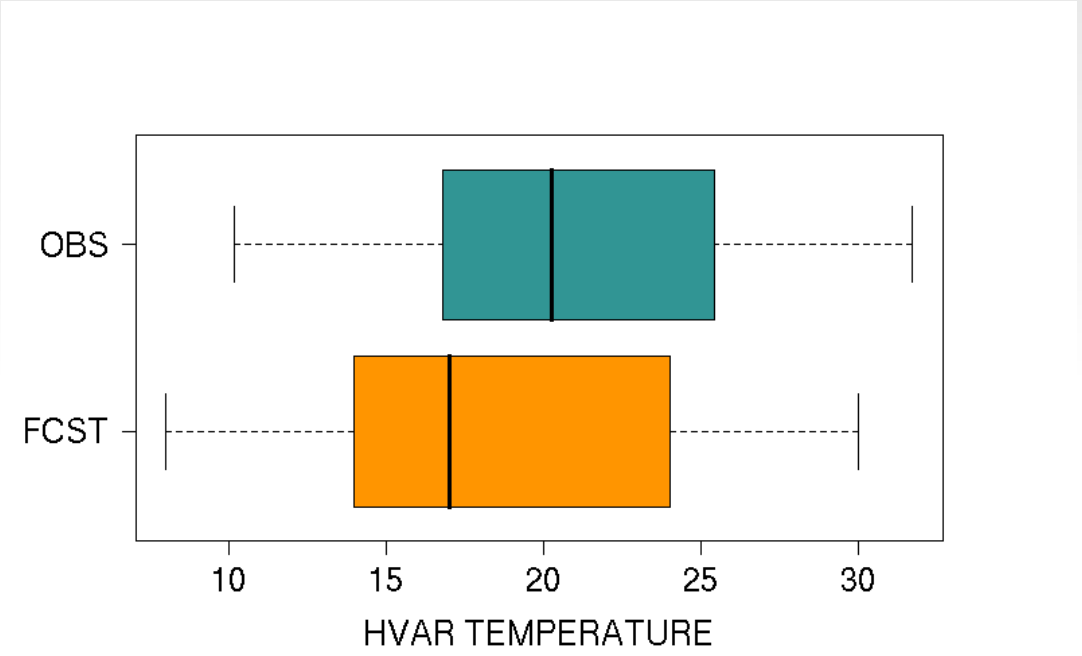
$$\text{median} = q_{0.5}$$

- ***Spread:***

$$\text{st dev} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

$$\text{Inter Quartile Range} =$$

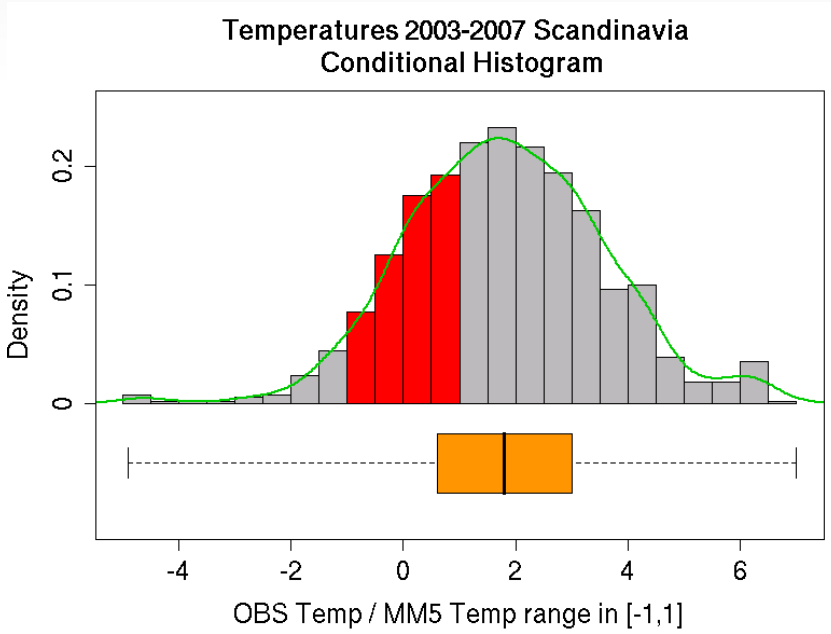
$$\text{IQR} = q_{0.75} - q_{0.25}$$

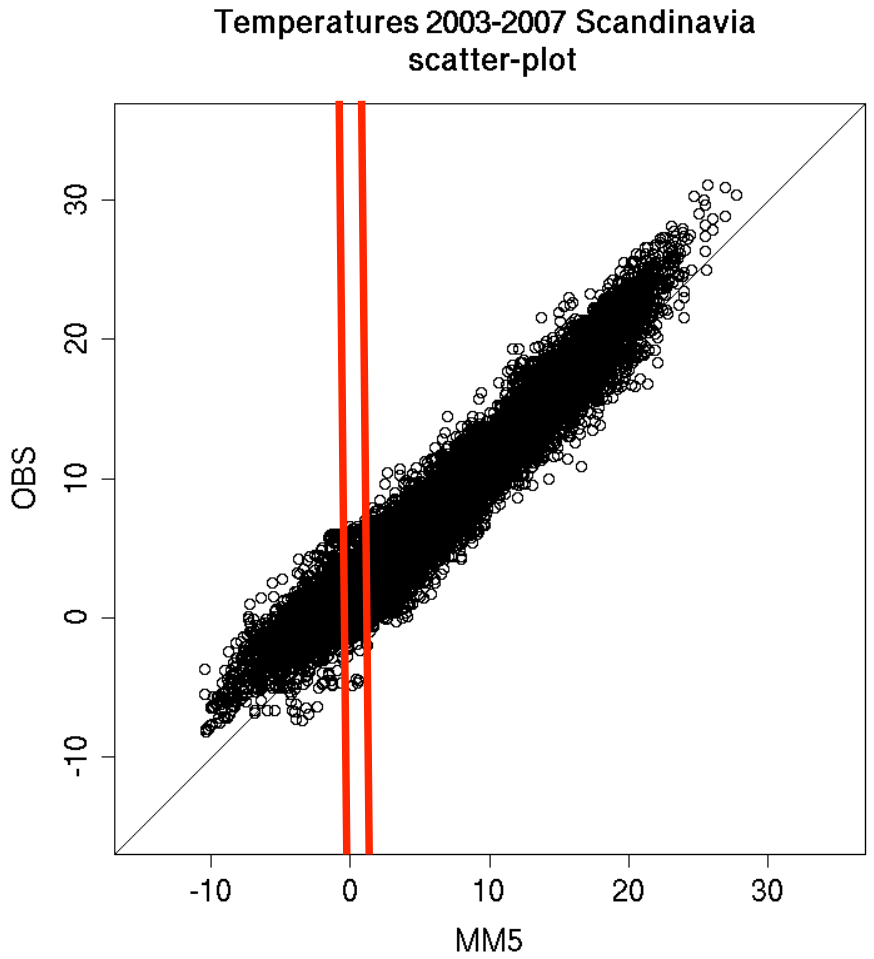


	MEAN	MEDIAN	STDEV	IQR
OBS	20.71	20.25	5.18	8.52
FRCS	18.62	17.00	5.99	9.75

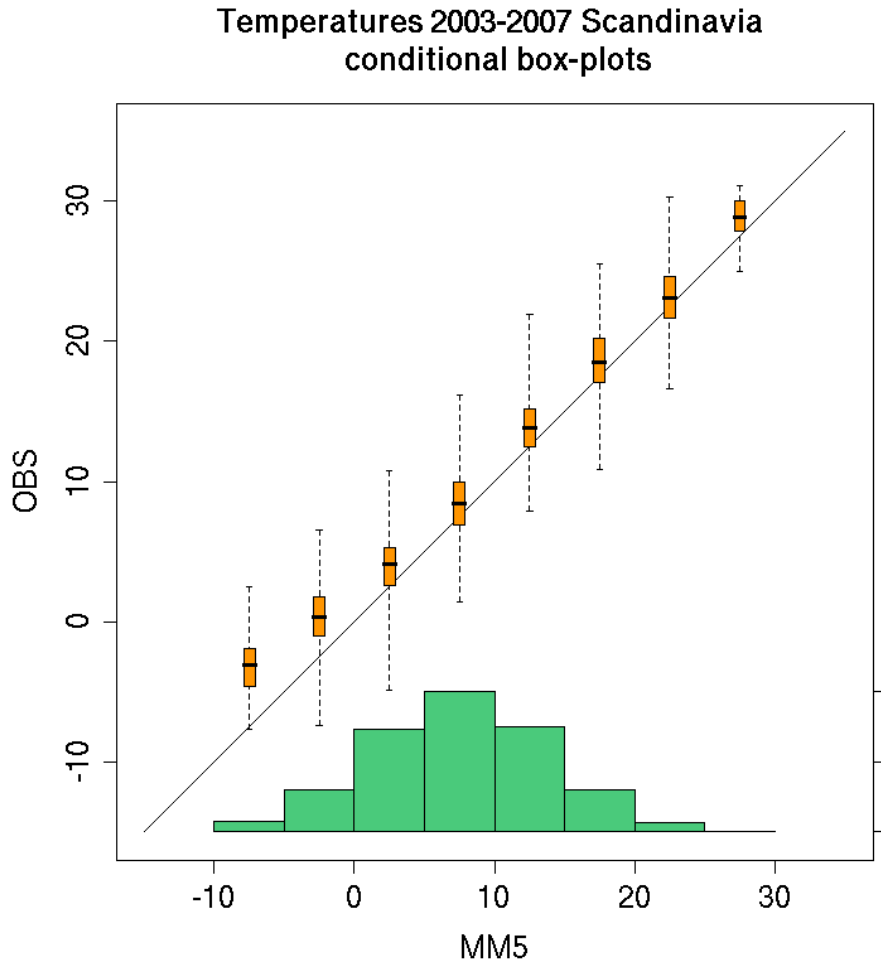
Exploratory methods: conditional distributions

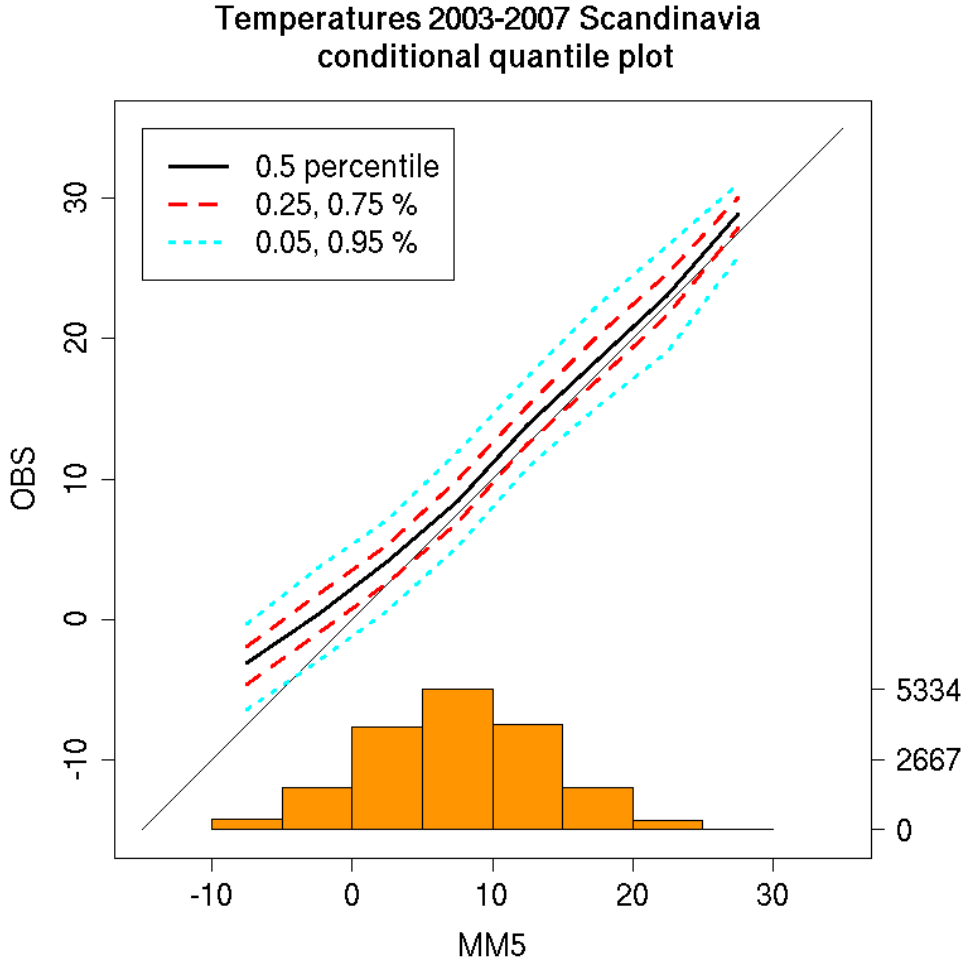
Conditional histogram and conditional box-plot





Exploratory methods: conditional qq-plot





Exploratory methods: class activity

Consider the data set of temperatures provided by Martin Benko (Benko.csv). Select a location and for the corresponding observation and forecasts:

1. Produce the scatter-plot and quantile-quantile plot: analyse visually if there is any bias, outliers, peculiar behaviours at the extremes, …
2. Produce the conditional quantile plot: are there sufficient data to produce it ? is it coherent with the scatter-plot ?
3. Produce side to side the box-plots of forecast and observation: how do the location and spread of the marginal distributions compare ?
4. Evaluate mean, median, standard deviation and Inter-Quartile-Range: do the statistics confirm what you deduced from looking at the box-plot, scatter-plot and quantile-quantile plot ?

Continuous scores: linear bias

$$\text{linear bias} = ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i) = \bar{Y} - \bar{X}$$

Attribute: measures the bias

Mean Error = average of the errors = difference between the means

It indicates the average direction of error: positive bias indicates over-forecast, negative bias indicates under-forecast (y=forecast, x=observation)

Does not indicate the magnitude of the error (positive and negative error can cancel outs)

Bias correction: misses (false alarms) improve at the expenses of false alarms (misses). Q: If I correct the bias in an over-forecast, do false alarms grow or decrease ? And the misses ?

Good practice rules: sample used for evaluating bias correction should be consistent with sample corrected (e.g. winter separated by summer); for fair validation, cross validation should be adopted for bias corrected forecasts

Continuous scores: MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i\right|$$

Attribute: measures accuracy

Average of the magnitude of the errors

Linear score = each error has same weight

It does not indicates the direction of the error, just the magnitude

Q: If the ME is similar to the MAE, performing the bias correction is safe, if MAE >> ME performing the bias correction is dangerous: why ?

A: if MAE >>ME it means that positive and negative errors cancel out in the bias evaluation …

Continuous scores: MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

Attribute: measures accuracy

Average of the squares of the errors: it measures the magnitude of the error, weighted on the squares of the errors

it does not indicate the direction of the error

Quadratic rule, therefore large weight on large errors:
→ good if you wish to penalize large error
→ sensitive to large values (e.g. precipitation) and outliers; sensitive to large variance (high resolution models); encourage conservative forecasts (e.g. climatology)

Continuous scores: RMSE

$$RMSE = \sqrt{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y - x\right)^2$$

Attribute: measures accuracy

RMSE is the squared root of the MSE: measures the magnitude of the error retaining the variable unit (e.g. ^{o}C)

Similar properties of MSE: it does not indicate the direction the error; it is defined with a quadratic rule = sensitive to large values, etc.

NOTE: RMSE is always larger or equal than the MAE

Q: if I verify two sets of data and in one I find RMSE ≫ MAE, in the other I find RMSE ≳ MAE, which set is more likely to have large outliers ? Which set has larger variance ?

Continuous scores: linear correlation

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\mathrm{cov}(Y,X)}{s_Y s_X}$$

Attribute: measures association

Measures linear association between forecast and observation

Y and X rescaled (non-dimensional) covariance: ranges in [-1,1]

It is not sensitive to the bias

The correlation coefficient alone does not provide information on the inclination of the regression line (it says only is it is positively or negatively tilted); observation and forecast variances are needed; the slope coefficient of the regression line is given by $b = (s_X/s_Y)r_{XY}$

Not robust = better if data are normally distributed
Not resistant = sensitive to large values and outliers

MSE and bias correction

$$MSE = \left(\bar{Y} - \bar{X}\right)^2 + s_Y^2 + s_X^2 - 2 s_Y s_X r_{XY}$$

$$MSE = ME^2 + \mathrm{var}\,(Y - X)$$

Q: if I correct the forecast from the bias, I will obtain a smaller MSE. If I correct the forecast by using a climatology (different from the sample climatology), will I obtain a MSE smaller or larger than the one I obtained for the forecast with the bias corrected ?

$$MSE_{bias} = \overline{\left(Y - \left(\bar{Y} - \bar{X}\right) - X\right)^2} = \mathrm{var}\,(Y - X) = MSE - ME^2$$

$$MSE_{cli} = \overline{\left(Y - c - X\right)^2} = MSE - 2\,cME + c^2$$

$$\left(ME - c\right)^2 \geq 0 \Rightarrow MSE_{cli} \geq MSE_{bias}$$

Continuous scores: class activity

5. Evaluate ME, MAE, MSE, RMSE and correlation coefficients: Compare MAE and ME, is it safe to perform a bias correction ? Compare MAE and RMSE: are there large values in the data ? Is the data variability very high ?
6. Substitute some values of your data with large (outliers) values. Re-evaluate the summary statistics and continuous scores. Which scores are the most affected ones ?
7. Add to your forecast values some fixed quantities to introduce different biases: does the correlation change ? And the regression line slope ? Multiply your observations by a constant factor: does the correlation change ? How does the observation standard deviation and the regression line slope change ? Multiply now the forecast values by a constant factor: how does this affect correlation, forecast standard deviation and regression line slope ?
8. Perform a bias correction on your data. How does this affect ME, MSE and correlation ? Then, change the variance of forecast and observation by multiplying their values by some constant factors. How does this affect the ME, MSE and correlation ?

Other suggested activities (advanced)

- Separate your data to simulate a climatology and a sample data set. **Evaluate the MSE for the forecast corrected with the sample bias and the climatology:** verify that $MSE_{cli} \geq MSE_{bias}$
- Deduce algebraically the relation between MSE and correlation if **bias is corrected and forecast rescaled** by s_X / s_Y. Does the MSE depend on the observation variance ? What happen if I rescale both forecast and observations with their corresponding standard deviations ?
- **Sensitivity of scores to spatial forecast resolution:** evaluate MSE for your spatial forecast, observation and forecast variance, ME and correlation. Then smooth the forecast and observation (e.g. averaging nearby *nxn* pixels) and re-compute the statistics. Which scores are mostly affected ?

Continuous skill scores: MAE skill score

$$SS_{MAE} = \frac{MAE - MAE_{ref}}{MAE_{perf} - MAE_{ref}} = 1 - \frac{MAE}{MAE_{ref}}$$

Attribute: measures skill

Skill score: measure the forecast accuracy with respect to the accuracy of a reference forecast: positive values = skill; negative values = no skill

Difference between the score and a reference forecast score, normalized by the score obtained for a perfect forecast minus the reference forecast score (for perfect forecasts MAE=0)

Reference forecasts:

- **persistence:** appropriate when time-correlation > 0.5
- **sample climatology:** information only a posteriori
- **actual climatology:** information a priori

Continuous skill scores: MSE skill score

$$SS_{MSE} = \frac{MSE - MSE_{ref}}{MSE_{perf} - MSE_{ref}} = 1 - \frac{MSE}{MSE_{ref}}$$

Attribute: measures skill

Same definition and properties as the MAE skill score: measure accuracy with respect to reference forecast, positive values = skill; negative values = no skill

Sensitive to sample size (for stability) and sample climatology (e.g. extremes): needs large samples

Reduction of Variance: MSE skill score with respect to climatology.
If sample climatology is considered:

$$Y = \bar{X}; \quad MSE_{cli} = s_X^2 \quad \text{and} \quad RV = 1 - \frac{MSE}{s_X^2} = r_{XY}^2 - \left(r_{XY} - \frac{s_Y}{s_X} \right)^2 - \left(\frac{\bar{Y} - \bar{X}}{s_X} \right)^2$$

linear correlation

bias

reliability: regression line slope coeff $b=(s_X/s_Y)r_{XY}$

Suggested activities: Reduction of Variance

- Show mathematically that the Reduction of Variance evaluated with respect to the sample climatology forecast is always smaller than the one evaluated by using the actual climatology as reference forecasts

- Compute the Reduction of Variance for your forecast with respect to the sample climatology, and compute each of its components (linear association, reliability and bias) as in the given equation. Modify your forecast and observation values in order to change, one at a time, each term: analyse their effect on the RV. Then, modify the forecast and observation in order to change two (or all) terms at the same time, but maintaining RV constant: analyse of how the terms balance each other

Continuous skill scores: good practice rules

- Use same climatology for the comparison of different models
- When evaluating the Reduction of Variance, **sample climatology** gives always worse skill score than **long-term climatology**: ask always which climatology is used to evaluate the skill
- If the climatology is calculated pulling together data from many different stations and times of the year, the skill score will be better than if **a different climatology for each station and month of the year are used.** In the former case the model gets credit from forecasting correctly seasonal trends and specific locations climatologies; in the latter case the specific topographic effects and long-term trends are removed and the forecast discriminating capability is better evaluated. Choose the appropriate climatology for fulfilling your verification purposes
- Persistence forecast: use same time of the day to avoid diurnal cycle effects

Continuous scores: anomaly correlation

$$y'_m = y_m - c_m$$
$$x'_m = x_m - c_m$$

Forecast and observation anomalies to evaluate forecast quality not accounting for correct forecast of climatology (e.g. driven by topography)

$$AC_{cent} = \frac{\sum_{m \in map} \left(y'_m - \overline{y}'\right)\left(x'_m - \overline{x}'\right)}{\sqrt{\sum_{m \in map} \left(y'_m - \overline{y}'\right)^2 \sum_{m \in map} \left(x'_m - \overline{x}'\right)^2}}$$

Centred and uncentred AC **for weather variables defined over a spatial domain**: c_m is the climatology at the grid-point m, over-bar denotes averaging over the field

$$AC_{unc} = \frac{\sum_{m \in map} \left(y_m - c_m\right)\left(x_m - c_m\right)}{\sqrt{\sum_{m \in map} \left(y_m - c_m\right)^2 \sum_{m \in map} \left(x_m - c_m\right)^2}} = \frac{\sum_{m \in map} \left(y'_m\right)\left(x'_m\right)}{\sqrt{\sum_{m \in map} \left(y'_m\right)^2 \sum_{m \in map} \left(x'_m\right)^2}}$$

Continuous Scores of Ranks

Continuous scores sensitive to large values or non robust (e.g. MSE or correlation coefficient) are some-times evaluated by using the ranks of the variable, rather than its actual values

Temp ^{o}C	27.4	21.7	24.2	23.1	19.8	25.5	24.6	22.3
rank	8	2	5	4	1	7	6	3

The value-to-rank transformation:

- diminish effects due to large values
- transform marginal distribution to a Uniform distribution
- remove bias

Rank correlation is the most used of these statistics

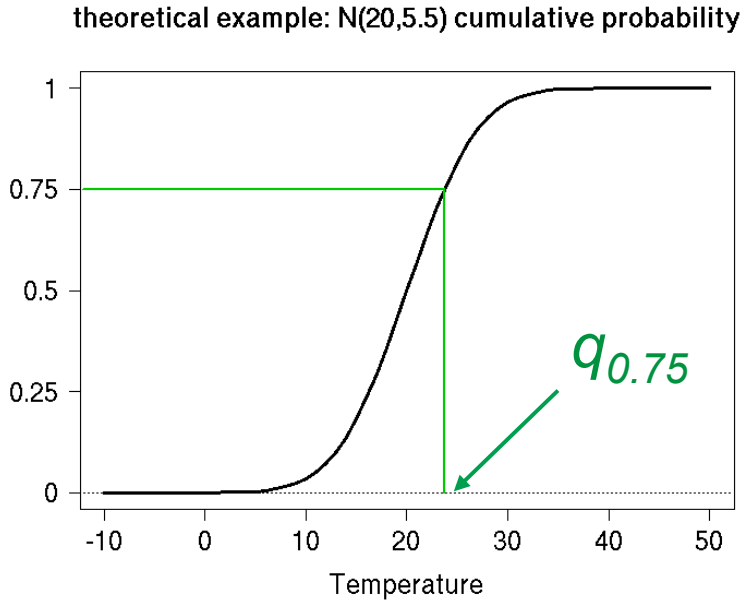
Linear Error in Probability Space

$$LEPS = \frac{1}{n}\sum_{i=1}^{n}\left|F_X(y_i) - F_X(x_i)\right|$$

The LEPS is a MAE evaluated by using the cumulative frequencies of the observation

Errors in the tail of the distribution are penalized less than errors in the centre of the distribution

MAE and LEPS are minimized by the median correction



Suggested Activities: ranks and LEPS

- Evaluate the correlation coefficient and rank correlation coefficient for your data. Substitute some values with large (outliers) values and re-calculate the scores. Which one is mostly affected ?
- Consider a precipitation data set: is it normally distributed ? Produce the observation-forecast scatter-plot and compute the MAE, MSE and correlation coefficient for
  - ➔ the actual precipitation values
  - ➔ the ranks of the values
  - ➔ the logarithm of the values, after adding 1 to all values
  - ➔ the nth root of the values (n=2,3,4, …)
  - ➔ the forecast and obs cumulative probabilities of the values

  Compare the effects of the different transformations
- If you recalibrate the forecast, so that $F_X=F_Y$, and evaluate the MAE after performing the last of the transformations above, which score do you calculate ?

Thank you!

References:

Jolliffe and Stephenson (2003): Forecast Verification: a practitioner's guide, Wiley & Sons, 240 pp.

Wilks (2005): Statistical Methods in Atmospheric Science, Academic press, 467 pp.

Stanski, Burrows, Wilson (1989) Survey of Common Verification Methods in Meteorology

http://www.eumetcal.org.uk/eumetcal/verification/www/english/courses/msgcrs/index.htm

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html