Lund University
Centre for Mathematical Sciences
Mathematical Statistics

FMS051/MASM17
2022

Project

Prediction of weather may be done globally or locally, based on more or less available information. In this project, you are to model and predict the outdoor temperature in one place in southern Sweden. The global part of the model will be fetched from SMHI (the Swedish weather bureau); together with local temperature measurements, we will then use this to do a local temperature prediction.

The measurements that form the basis for the project are taken during the end of 1993, the whole 1994, and the beginning of 1995 in Sturup and Växjö by SMHI, and in Svedala, Alvesta, and Harplinge by Sydkraft. Furthermore, a couple of times a day, SMHI produces predictions for the outdoor temperature in three hour intervals in Sturup and Växjö. In order to get hourly predictions, production personal at Sydkraft make linear interpolation between the values gotten from SMHI. Hence, the prediction of the outdoor temperature we have at our disposal for a certain hour is the last one that was produced for that particular hour. That means that we cannot give a common prediction horizon for all predictions, but the prediction that is registered is computed with a prediction horizon between 1 and approximatively 10 hours. Your model can be based solely on the measured outdoor temperature, but by using the predictions from SMHI for the same place, or a place close by, as an external signal, you can hope for a better modeling of the temperature, and possible also for a better local temperature prediction.

The task to be fulfilled in the project consists in studying parts of a set of data containing hourly measurements of outdoor air temperature and predictions thereof in Sturup, Svedala, Växjö, Harplinge, or Alvesta. You are supposed to *model* and *predict* the air temperature during some season.

A. **Modeling without an external input.**
Start with the measurements of the temperature. Choose ten weeks for the *modeling* part, clearly indicating which location and time period you have selected. Select the two following weeks (at the same location) as the *validation* data. Select two weeks about 20-30 weeks later as your *test* data. Beware that there might be outliers or missing samples, with some time periods missing more samples than other. How can you deal with such missing samples? Do you need to be concerned with outliers in your data set? Construct a model for the temperature data, without using any external inputs; this is model A.
(*20 marks*)

B. **Modeling with an external input.**
Investigate if your model and the predictions of the temperature are improved if you consider the prognosis made by SMHI as an external signal; this is model B. Remember that you need to re-design the entire model when doing this. Note that you will also need to predict this input. Present the models for your inputs and the quality of the predictions of these inputs.
(*20 marks*)

C. **Recursive model.**
Consider model B derived above, making use of an external input, and formulate a time-recursive version of this model; this is model C. Initiate the recursive estimator at the start of your modeling set (for example, using the parameters you obtained in model B), and let it run through all the available data and then extract the part corresponding to the validation data. Try to simplify the model by removing some parameters (without causing substantial loss of performance). Plot the estimated parameters for the 9-step prediction for your *validation* data. Comment on your observations.
(*20 marks*)[1]

D) **Compare with an automatic predictor (optional challenge).**
Facebook has released *Prophet*[2], a fully automatic technique for forecasting time series data, allowing for yearly, weekly, and daily seasonal trends, as well as for holiday effects. Compare your predictions on the validation data sets with that given by *Prophet*. If you want to be fair (which is of course always nice to be, but as this is an optional task, you have full artistic freedom), you should only allow *Prophet* the same amount of data that you had access to in your modeling set.
(*optional, up to 5 bonus marks*)[3]

For each models, you should present the full model, including the confidence interval for the parameters, as well as key motivating steps to obtain this model. This includes showing the ACF for the resulting model residuals (include lags up to 50), as well as commenting on the whiteness of the residuals. When using an external input, the used model for the input should be well motivated (and the quality of the input prediction shown).

Compute the 1- and 9-step predictions of the temperature of the *validation* data and present the variances of the resulting prediction residuals in the below table. Remember to compute the residuals in the measured domain. Show the ACF for each of the prediction residuals and comment on the whiteness of the residuals. Plot the 9-step predictions for each model for the validation data, comparing it to the actual validation data.

[1] In case you had problems getting your model in (B) to work properly, formulate the recursion of your model in (A) instead; in this case, (C) is worth 15 marks.

[2] https://facebook.github.io/prophet/

[3] Yea, I know, these marks should not really be here as the project is worth 60 marks, but, hey, one should get something if spending the time, right?

Model	$\sigma^2_{t+1\|t}$	$\sigma^2_{t+9\|t}$
Naive model	·	·
Model A	·	·
Model B	·	·
Model C	·	·

Here, $\sigma^2_{t+k|t}$ denotes the variance of the k-step prediction residual. The naive model should be simple, but reasonable. As a reference, give the variance of your validation data, σ^2_y.

Finally, compute the variance of the prediction residuals for your *test* data and present this in a similar table as above; remember to include the variance of the test data as well. Comment on the results.

Your report should be well motivated and clearly describe the different parts of the project. However, in the interest of conciseness, the length of the report should not exceed 30 pages. The project can be done in groups of maximally two students. Discussions on the project with anyone other than the teaching staff is prohibited and it is expected that all students refrain from this. Please state on your project that you have not collaborated with anyone when solving it and sign with your name.

During the oral presentation, which will be about 10 minutes, *one* member in the team will be asked to present the project. It is expected that each project member can *individually* motivate and explain any part in the project solution, in detail.

Note: If you have already decided that you will not hand in the take-home, please indicate this in your project report. I will then just grade it as pass/fail, which will speed up the grading procedure significantly, allowing me to report your grades quicker.

Good luck - and have fun!

Data

The data files which you are going to use can be found on the homepage for the course. The following data sets are available:

File	Comments
ptstu93.mat	Prediction of the temperature in Sturup during 1993 (SMHI)
ptstu94.mat	Prediction of the temperature in Sturup during 1994 (SMHI)
ptstu95.mat	Prediction of the temperature in Sturup during 1995 (SMHI)
ptvxo93.mat	Prediction of the temperature in Växjö during 1993 (SMHI)
ptvxo94.mat	Prediction of the temperature in Växjö during 1994 (SMHI)
ptvxo95.mat	Prediction of the temperature in Växjö during 1995 (SMHI)
tstu93.mat	Measured temperature in Sturup during 1993 (SMHI)
tstu94.mat	Measured temperature in Sturup during 1994 (SMHI)
tstu95.mat	Measured temperature in Sturup during 1995 (SMHI)
tvxo93.mat	Measured temperature in Växjö during 1993 (SMHI)
tvxo94.mat	Measured temperature in Växjö during 1994 (SMHI)
tvxo95.mat	Measured temperature in Växjö during 1995 (SMHI)
utempAva_9395.dat	Measured temperature in Alvesta during 1993-95 (Sydkraft)
utempHar_9395.dat	Measured temperature in Harplinge during 1993-95 (Sydkraft)
utempSla_9395.dat	Measured temperature in Svedala during 1993-95 (Sydkraft)
tid93.mat	Time schedule for the measurements for 1993
tid94.mat	Time schedule for the measurements for 1994
tid95.mat	Time schedule for the measurements for 1995

The data files from SMHI contain only one column with temperature data starting on the first hour of the year. As mentioned above, we have 8 values per twenty-four hours, which are interpolated linearly to hourly measurements. The accuracy is 1 ^{o}C.

The data files from Sydkraft are registered every hour apart from the last hour of the day, which is represented by 0. The accuracy is 0.3 ^{o}C. These files have three columns: day, hour, temperature.

The time schedule files have five columns: number of the hour, year, number of the week, day of the week, hour.