Project: Analysis of Trends in Temperatures

Mathematical Statistics (EBC2107)
School of Business and Economics
Maastricht University

Stephan Smeekes (s.smeekes@maastrichtuniversity.nl)
Robert Adámek (r.adamek@maastrichtuniversity.nl)
Jakob Raymaekers (j.raymaekers@maastrichtuniversity.nl)

1 Introduction

For this assignment you will analyse if there is evidence of an upward trend in temperatures recorded over the last century at several Dutch locations.

For this purpose you have to implement the statistical techniques learned during the course in the programming language R, and apply these to the analysis of historical temperature data for several locations in the Netherlands. The assignment will be graded and counts for 20% toward the final grade. The assignment is done with the same sub-group as formed for the tutorials.

2 Data

On the course page you can find an Excel file and .csv (comma separated value) files that contain average temperatures for three locations in the Netherlands: De Bilt, Eelde and Maastricht. The data are available from 1907 up to 2020.[1] Data are available on three frequencies: daily, monthly and yearly.

Daily data These represent average daily temperatures recorded at the three locations from January 1, 1907, up to December 31, 2020. As these data are rather noisy, too large to handle directly, and subject to seasonal patterns, you do not have to use them. They are simply provided as they form the basis for the other data series.

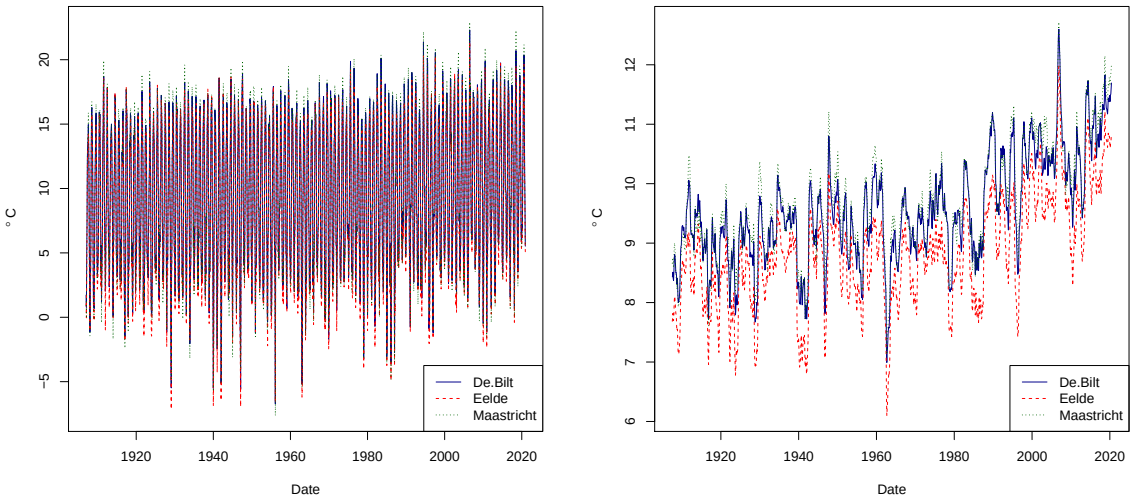
Monthly data The data are obtained by taking the average over all daily temperatures within a month. As a lot of the daily noise has been filtered out through averaging, these data are easier to handle. However, seasonal patterns still affect the data; clearly, temperatures in January will be different than in July. This means that finding trends is still complicated. Therefore you don't have to use these data either.

[1] As the website has technical difficulties, the data for 2021 cannot be downloaded currently. The dataset might be updated at a later moment.

Smoothed monthly data In order to be able to use the monthly data, they need to be "seasonally adjusted" in order to remove the seasonal patterns. On way to do this (not necessarily the best way) is to "smooth" the data, which means that for every month one takes a (weighted) average of the months around that month. Here we take a linear smoother, that is, we take an average with equal weights for all months within half a year on either side. Formally, if we let Y_i be the temperature in month i, then the smoothed temperature in month i, denoted as Y_i^s, is given by

$$Y_i^s = \frac{1}{24}Y_{i-6} + \frac{1}{12}\sum_{j=-5}^{5} Y_{i+j} + \frac{1}{24}Y_{i+6}.$$

Figure 1 plots the original and smoothed monthly data side by side. Note the - enormous - differences.



(a) Original monthly temperatures

(b) Smoothed monthly temperatures

Figure 1: Monthly Dutch temperatures 1907-2020

Annual data The annual data are simply calculated as the average over all days within one year. As these data do not contain any seasonal patterns anymore, they can directly be used for the analysis of trends. These data form the major input to the assignment. Figure 2 plots the annual data.

3 Programming in R

For the assignment you have to programme the techniques we learn in the course in the statistical software package R. R is available for (free) download on www.r-project.org. More information about R can be found on the course page.

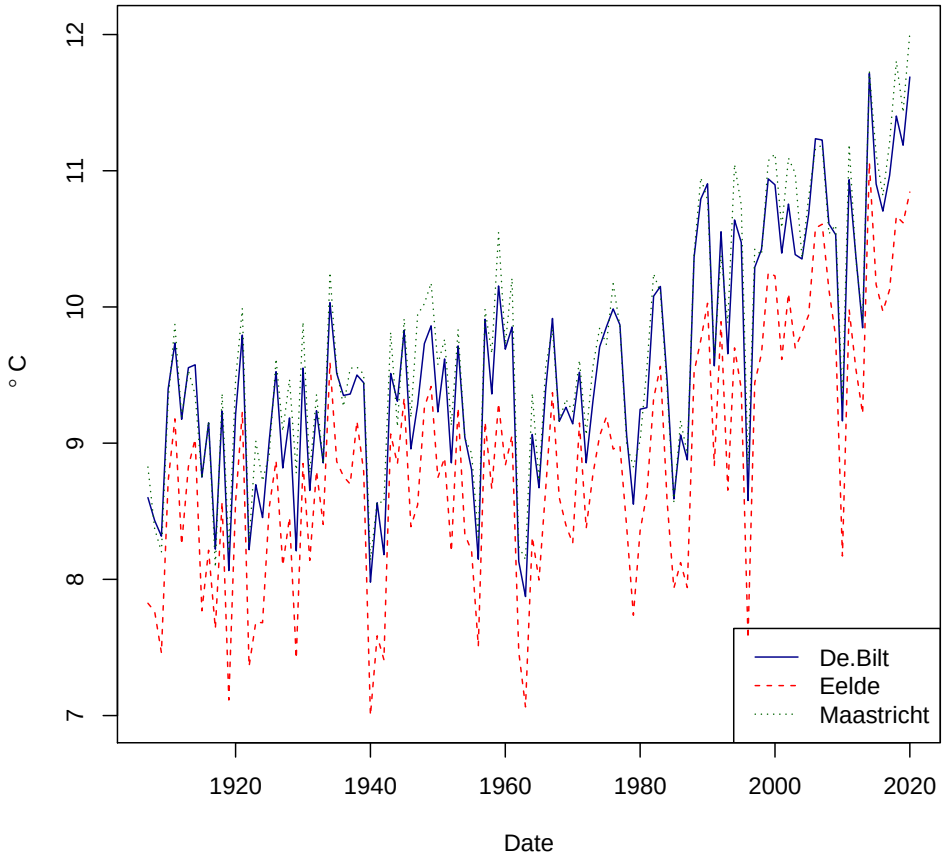


Figure 2: Annual Dutch temperatures 1907-2020

4 Assignment

For your assignment you write a paper where you should try to answer the question if there is statistical evidence of an upward trend in the temperature series. The main focus should be on the annual data. Choose one series as your main series of interest, but check if your conclusions change depending on which series you use.

To guide you in the analysis, below you can find a list with specific questions to consider in your analysis. Remember though that in the end you should provide one coherent analysis in the paper, and not a point by point answering of the questions.

Compare average temperatures in different parts of the sample

Start by analyzing average temperatures over different parts of the sample. Split the sample in a number of subsamples, and compare average temperatures across the subsamples. You can vary the way how to split your sample. Make sure estimation uncertainty is taken into account, e.g. by constructing confidence intervals. You could also consider overlapping versus

non-overlapping subsamples.

You can also consider a formal test for equality in different subsamples, for example you could split the sample in two and test whether the mean temperatures in both parts are equal.

Investigate the presence of a linear upward trend

Next we fit a linear regression model to the data. That is, if $Y_1, \ldots, Y_n$ are the temperature data, we fit the regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{1}$$

and take $x_1, \ldots, x_n$ to represent a linear trend. Estimate the model, and provide measures of estimation uncertainty such as confidence intervals. Also perform a hypothesis test to test if an upward trend is visible in the data.

Investigate the presence of a linear upward trend in part of the sample

While we have data from 1907 on, some people claim temperatures only started to rise significantly from the seventies on. Here we investigate that claim. We still consider model (1), but now we take $x_1, \ldots, x_n$ such that the linear trend only kicks in from a starting point late in the sample on. Find a reasonable point to start the trend at (e.g. somewhere in the seventies) and motivate your choice. Motivation can come from either outside sources, or from the data. In case you motivate the choice from the data, explain how this could be misleading.

Analyse the model in the same way as for the overall linear trend and contrast your findings with that case.

Implement the bootstrap

For the most interesting parts of your analyse above, construct the relevant hypothesis tests and confidence intervals using the bootstrap rather than the standard approach. Focus especially on implementing the bootstrap for the regression model. Discuss how this changes your results and how to interpret any changes.

Discuss the assumptions you need in the analysis

For all parts of your analysis, discuss carefully which assumptions you needed to make. For example, do you need to assume normality? When do you need to assume independent and/or identically distributed random variables? Can you give an asymptotic justification of your methods that avoids some of the assumptions?

Once you set up the assumptions you need, discuss how likely it is that they are satisfied for your data. You can also think about ways to check if your assumptions are satisfied, either formally or informally. For example, can we check if normality, or independence, are satisfied by the data?

Extension to monthly data

So far you could use just the annual data for the analysis. A final issue that you could consider is how using monthly data changes the picture. A straightforward extension is just

to apply the same techniques to the smoothed monthly data. You can take some of the most interesting aspects of your previous analysis and repeat them using those data. If you do so, carefully discuss how much added benefit it is to consider these smoothed monthly data relative to the annual data, especially in light of the assumptions made.

Another option would be to use the monthly data in a different way. For example, you could think of alternative ways to remove the seasonal effect. Alternatively, you could use the information in the monthly data differently; rather than looking at overall trends you could consider just summer or winter, or maybe look at the variation in temperatures within a year.

Turn your analysis into a paper and conclude

Having performed all your analyses, you need to draw meaningful conclusions from it. Make sure your paper is coherent - writing a paper is more than just ticking boxes and doing exercises. In an academic paper you tell a story that has a logical flow from start to end. Formulate research questions in the introduction, address these in the analyses, and provide answers to the questions in the conclusion.

An important aspect of the story writing is to translate your statistical findings into societal meaningful conclusions. What is the meaning of your findings? Are there limitations to the methods used, or assumptions they require, that might affect how you draw conclusions? Try to address these issues in your paper.

5 Handing in and Grading

The deadline for handing in is halfway through the resit week: **Wednesday April 7, at 23:59**. You are strongly recommended to hand in earlier though! Each group should hand in the assignment by uploading on the course page on the Student Portal. Note that this entails an automatic plagiarism check. Hand in your paper as a PDF file, and give the R code you used in your assignment as an appendix to the paper (advice on how to do so will be posted on the course page).

All group members will receive the same grade for the paper. The grading will mostly be based on the correctness and completeness of your assignment. While the paper is the major determinant of the grade, the code plays a role too. While R has built-in commands for everything, including OLS and bootstrap, you learn most if you programme the techniques yourself. Therefore, programming the techniques such as least squares or bootstrap yourself rather than using built-in commands, will have a significant positive effect on your grade.

In Table 1 the rubric with criteria is provided that are considered in the grading. This may help you to structure your analysis and paper.

Technical Component	Weight: 65%
Comparison of average temperatures	
Choice of subsamples	Was the choice of subsample varied and interesting?
Testing and quantifying uncertainty	What methods were used to test for equality? How was estimation uncertainty quantified?
Linear Regression	
Model and interpretation	Were the correct models estimated? Were the estimated parameters interpreted correctly?
Testing and quantifying uncertainty	What methods were used to test for global warming? How was estimation uncertainty quantified?
Bootstrap	
Understanding	Were the principles of bootstrapping understood and correctly applied?
Comparing with previous results	Were the bootstrap results compared to previous results, and the differences correctly interpreted?
Additional	
Extensions to monthly data	Were interesting parts of the analysis extended to the monthly data? Was this data treated appropriately (e.g. seasonality addressed)?
Self-programmed	Were the methods implemented "manually", as opposed to using in-built functions or packages?
Writing Component	**Weight: 35%**
Setup	
Research question	Formulated a clear research question that is well-positioned in research context?
Use of literature and understanding	Appropriate references made to the literature where necessary? Use of external sources?
Theory and hypothesis development	Are the hypotheses clearly developed and relevant to the research question?
Discussion	
Interpretation of findings	Are the results clearly interpreted?
Arguing on academic and societal implications	What do the results imply for existing/future research? Are managerial/societal implications discussed?
Final product	
Coherence	Is the paper coherent? Does it have a logical structure? Are the research questions addressed in the analysis and answered in the conclusion?
Form, scientific writing and correct language use	Inclusion of visuals, absence of typos, logical structure, overall clarity in writing, no direct copying of R outputs.

Table 1: Criteria for grading