Lund University
Centre for Mathematical Sciences
Mathematical Statistics

FMSN45/MASM17
2018-11-21

Project

In this project, we will model, predict, and reconstruct measurements for the vegetation index and precipitation in Sudan. We will do this using data collected by the American National Oceanic and Atmospheric Administration (NOAA), which together with NASA operates several kinds of environmental satellites[1]. We will use data collected during 1982 to 1999, from the satellites NOAA–7 to NOAA–16, which measures the reflectance from the Earth's surface in five different spectral bands (580–680 nm, 725–1000 nm, 3.55–3.93 μm, 10.3–11.3 μm, and 11.5–12.5 μm).

Using these reflectance measurements, the normalized difference vegetation index (NDVI) is calculated according to

$$\text{NDVI} = \frac{Ch_2 - Ch_1}{Ch_2 + Ch_1}$$

where Ch_1 and Ch_2 denote the reflectance in the red spectral band (580–680 nm) and in the near infrared (725–1000 nm), respectively. The NOAA and NASA distribute the resulting estimated NDVI values as a part of the *Pathfinder AVHRR Land Data Set*. These measurements are here stored as unsigned (8-bit) integers (i.e., between 0 and 255), and thus needs to be rescaled to $[-1, 1]$ before you start your modeling.

The index is a good measure of vegetation cover, as chlorophyll in healthy leaves have a clear absorption peak in the red spectral band, whereas cell structures in the leaves reflect near infrared light, as illustrated in Figure 1. Few, if any, other objects have these distinct spectral characteristics, which give large differences between Ch_1 and Ch_2 reflectances. As a result, vegetation is therefore signified by high NDVI values, whereas barren ground, water, snow, ice, etc. will yield low values.

The data set contains four measurement series, one from each of the following weather stations 1) El-Geneina, 2) El-Fasher, 3) Ed-Damazine, and 4) Kassala. In Figure 2, the four places are marked with numbers corresponding to the order above. In this study, we will work with the data from El-Geneina and Kassala.

The vegetation index data is measured three times each month, day 1–10, 11–20, and 21–to the end of the month, which gives 36 points per year. The measurement period starts 1982

[1] Further details on the environmental and meteorological satellites that NOAA operate can be found at `http://www.noaa.gov/satellites.html`.

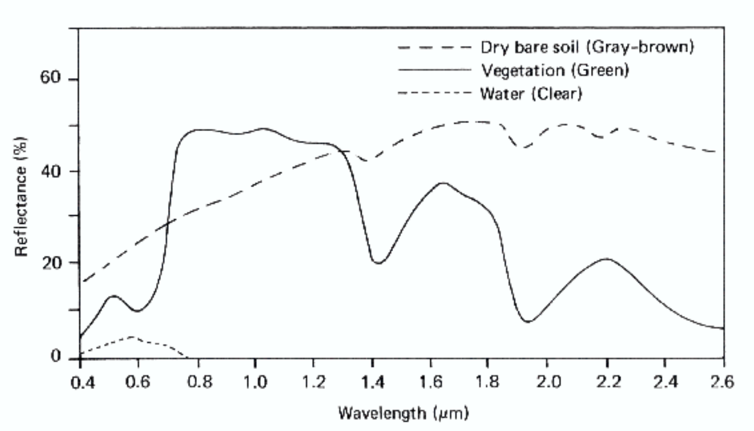

Figure 1: Green leaf reflectance for different wavelengths. The absorption peak in the red spectral band ($\sim 0.6\ \mu$m), caused by chlorophyll, is clearly seen.

and ends in 1999; in total, there are 648 measurements[2]. Once each month, the total amount of precipitation during that month is also measured. The measurement period begins 1960 and ends 1999. There are, in total, 480 data points. Since the actual rain data in `rain_org` was only measured once each month, whereas the vegetation index was measured every ten days, we have done a linear interpolation between the monthly rain measurements in order to obtain rain "measurements" for every tenth day. This interpolated data set is stored in `rain`. Both the NDVI and the rain data have missing data points, which have been estimated and filled by the MATLAB function `misdata`. The measurements are available on the course webpage, and are stored using the `struct` data type.

Your study should contain the following main parts:

(A) *Recursive reconstruction of rain data.*
As noted, the total amount of precipitation was measured once each month, whereas the vegetation index was measured three times each month (day 1–10, 11–20, and 21–end). The linear interpolation used to obtain rain "measurements" on the same time-scale as the NVDI data was done using a purely *ad hoc* approach. Since each original rain measurement is the *accumulated rain* during each period, we could instead write the measurement equations as

$$y_t = x_t + x_{t-1} + x_{t-2} + w_t, \quad \text{for} \quad t = 1,4,7,10,\ldots$$

where $x(t)$ is the accumulated rain on the denser time scale. Model the rain by an AR(1) process, and use a version of the Kalman filter to reconstruct the rain process. If needed, find a suitable data transformation of the data[3]. Plot the reconstructed rain and compare your results to the linear interpolation; note that the sum of all the rain ought to be about the same for both models.
(*20 marks*)

[2] Looking at the data, you can likely determine the year that Bob Geldof arranged the Band Aid recording of "Do they know it's Christmas?" to raise money for famine relief in Ethiopia (available on YouTube).

[3] Remember to offset the data if this is negative before taking any square root or logarithm!

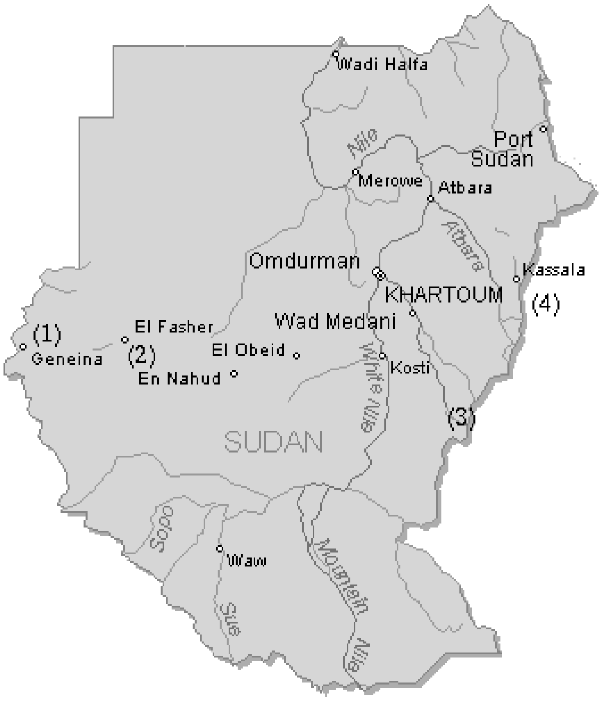

Figure 2: Weather stations in Sudan.

If you do not manage to solve part A entirely, or if the results are not to your satisfaction, simply use the supplied linearly interpolated data in the rest of the assignment.

(B) *Modeling and validation for El-Geneina.*
Construct suitable models for the NDVI data, both with and without precipitation as an external signal, for the El-Geneina location. Begin by modeling the data without taking the precipitation into account, and then examine how the precipitation may be used as an external signal[4], and how this affects the quality of your prediction. Begin by choosing a suitable amount of data for the modeling part, and then use the remaining data for validating. Present the models with confidence intervals for the estimated parameters and show the ACF of your modeling residuals. Then, use the models to predict the NDVI data and plot the 1-step as well as k-step prediction (select $4 \leq k \leq 10$). Plot the predictions as compared to true NDVI values for the validation data and compute the variance of the prediction errors[5], as well as show the ACF of the prediction errors (do this both for the model with and without input). Does the use of the precipitation to improve your prediction? Remember to compare your prediction to a naive predictor.
(*30 marks*)

[4] Remember that also simple models might work well when including an external signal.

[5] Beware that it is easy to get the predictions shifted as compared to the true data. Make sure that the true and the predicted data are aligned before you compute the residual! Recall also that the variance of the prediction error should increase with growing k; it might be worth checking if this is true.

(C) *Time-varying model for El-Geneina.*
As an optional task, consider making the better of your models recursive, such that the parameters of the model are allowed to vary over time. Does this improve the prediction ability as compared to your non-recursive model? When forming predictions using the precipitation data as an external input, you may treat future values of this data as known. Thus, even if needed, you do not need to predict the precipitation data (but you are of course welcome to do so if you wish).
(*optional, up to 5 bonus marks*)

(D) *Modeling for Kassala.*
Apply the better of your models to the data from Kassala. Are the prediction errors similar? Are the results improved if you re-estimate the model parameters based on the different measurements? Locate the two measurement stations on a map and discuss some possible reasons for the data to behave differently at the two locations.
(*10 marks*)

The project report should contain a summary of the steps you have taken throughout the study. The report should be written in a concise way with relevant figures, clearly explaining your results. The resulting models and predictors should be given in detail. Compare and discuss your results. Motivate your steps and conclusions carefully; the quality of the report will be taken into account in the marking. 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, it is expected that each project member can *individually* motivate and explain any part in the project solution, in detail.

Good luck - and have fun!