Lund University
Centre for Mathematical Sciences
Mathematical Statistics

FMSN45/MASM17
2023

Two roads diverged in a wood, and I–
I took the one less traveled by,
And that has made all the difference.
Robert Frost

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 weather stations in El-Geneina, El-Fasher, Ed-Damazine, and 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, providing 36 measurements per year. The measurement period starts 1982 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

[1] See also *http://www.noaa.gov/satellites.html*.

[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).

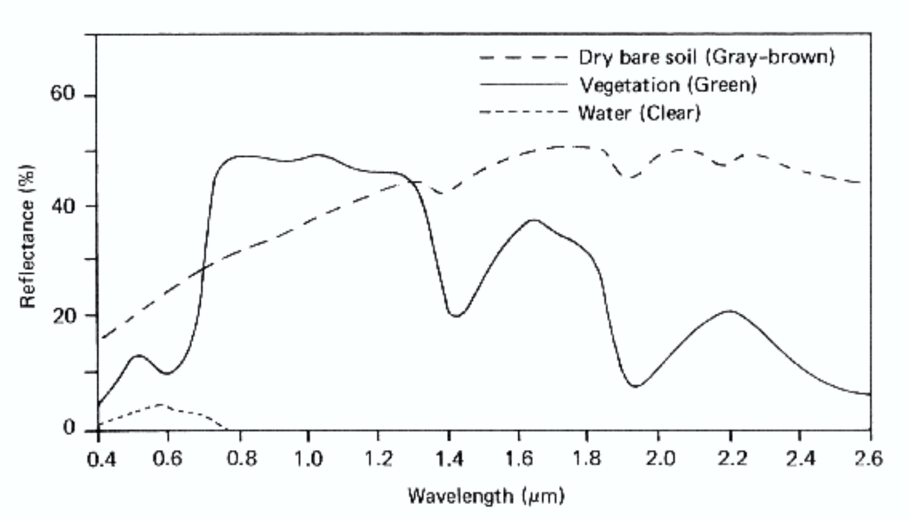

Figure 1: Green leaf reflectance for different wavelengths. The absorption peak in the red spectral band ($\sim 0.6\ \mu$m), caused by chlorophyll, is clearly seen.

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.

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 an *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 as an AR(1) process, and reconstruct the rain process using the Kalman filter. 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.

(*15 marks*)

(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, and how this affects the quality of your prediction. Begin by choosing a suitable

[3] Remember to offset the data if doing any transformation requiring non-negative data!

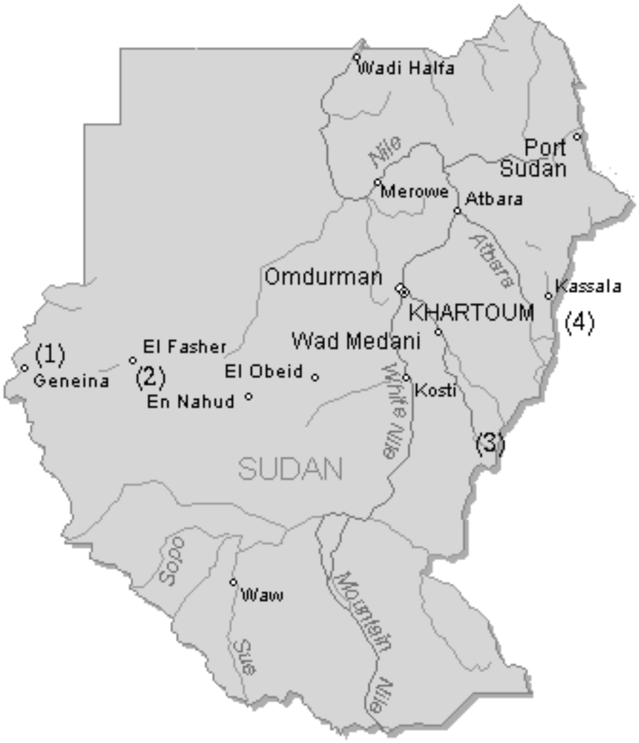

Figure 2: Weather stations in Sudan.

amount of data for the *modeling* part, and then use most of the remaining data for *validation*; retain a smaller set as *test* data. Construct your model using the modeling data and use the validation data to determine how well your model can predict this data set. Remember that you can only access the rain data up to the current time and any future rain needs to be predicted[4]. Does the use of the precipitation improve your predictions?
(*25 marks*)

(C) **Time-varying model for El-Geneina**.
Using the model with the rain as input, examine if you can improve your modeling by recursively updating you model, allowing the parameters to vary over time. Recall that you also have to recursively update your input predictions. Can you remove some parameters without losing performance? Does the recursive model improve the prediction ability as compared to your non-recursive model?
(*15 marks*)[5]

(D) **Modeling for Kassala**.
Apply your best model(s) 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.
(*5 marks*)

[4] Use the rain data from part A to form these predictions; it is allowed to use the reconstructed measurements up the current time as given.

[5] If using the model without input, this problem is worth 10 marks.

E) **Compare with an automatic predictor (optional).**
Try using *Prophet*, *NeuralProphet*, and/or *TimeGPT*[6] to form the predictions. Compare the predictions with your best model. To be fair, you should only allow the automatic technique(s) the same data used for your modeling.
(*optional, up to 5 bonus marks*)[7]

For each models, you should present the full model, including confidence intervals, as well as key motivating steps to obtain this model. This includes showing the ACF for the resulting *model residuals*, 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).

Plot the 1- and 7-step predictions of the *validation* data and compare it to the true data. Present the variances of the resulting *prediction residuals*[8], comparing these to your *naive* predictor. To allow for the variability in the data, also give the variance of the prediction residuals normalized by the variance of the corresponding validation data. Show the ACF for each of the prediction residuals and comment on these. Finally, compute the (normalized) variance of the prediction residuals for your *test* data[9]. 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 of a *random* selection of teams will be asked to present their 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 send me an email stating this*. I will then grade the project as pass/fail, which will speed up the grading procedure significantly, also allowing me to report your grades quicker.

Good luck - and have fun!

[6] *https://facebook.github.io/prophet/*, *https://neuralprophet.com/*, and *https://docs.nixtla.io/*.

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

[8] Remember to compute the residuals in the measured domain.

[9] No marks are given on the prediction quality on the test data (although your comments are fair game!).

Reflections

Some suggestions and thoughts:

- Are all coefficients significant? Can you reduce the model? Should you add some seemingly too weak component?
- Should you replace that ∇ with a $(1 - az^{-1})$ and estimate the a?
- Is the periodicity not perfectly aligned with the samples? Maybe worth testing replacing the ∇_s with

  $$(1 - a_{s-1}z^{-s+1} - a_s z^{-s} - a_{s+1}z^{-s-1})$$

  This will allow the periodicity to be close to s, but does not require it to be precisely s.
- Do not oversimplify your model; do not restrict it to be AR to allow you to use RLS. This is unlikely to work well... Do not expect complex phenomena to be well modelled with almost no parameter...
- Are your predictions ok, but lower than you expect? Did you remember to add the predicted input?
- Does it seem difficult to get rid of all dependencies? Is the model really stationary? Test using your recursive predictor.
- Always, always, always... Compare to a naive predictor. If your model is not better than that, it is not worth much!

Reflections

More suggestions and thoughts:

- Does it make sense to transform the data? Look at the Box-Cox plot. How much higher is really that peak?
- Simulate a process, with lots of data, and test your code. Test your predictions using the model you used to simulate the data. Check if your Kalman filter converge to the true parameters if initialised wrong. Everything should work really well; if it does not, something is wrong. Check your code!
- Do you suffer from outliers? Why not compute the TACF and compare it to the ACF - if you are, these estimates will likely differ.
- Remember that many of the results only holds asymptotically - and typically relies on several assumptions that might not be met. Be careful with your trust!
- It is often wise to plot your ACF without $\hat{\rho}_y(0)$ to better see the details.
- Always plot predictions in the correct domain - do not show the transformed differentiated predictions; show them in the domain you work in!
- If you add an input, do not retain your old model - use the input maximally before modelling the residual!
- Look at the predictions and residuals; do they make sense? Are they in the correct domain?
- Be careful to use Matlab commands that you do not fully understand.
- Multiple inputs? Use the one with the strongest correlation first, then use the next on the resulting residual.
- Is it good enough? Then, you are done!