Portfolio

Maxim Mahnkopf 3761412
August 31, 2023

This document outlines the workflow and my working progress in the "Spatio-temporal Data" course, which is part of the "Earth System Data Science and Remote Sensing" program. The course was divided into 13 weekly sessions (lecture + seminar) and began on April 6, 2023.

Week 1 (April 6, 2023): Introduction

During the initial lecture of 'Spatio-temporal Data,' the course content was presented. This course primarily focuses on working with data cubes. I learned that data cubes are expansive datasets encompassing both spatial and temporal dimensions. Typically, in environmental sciences, these dimensions include latitude, longitude, time, and various climate variables. Although I hadn't personally worked with data cubes before, I found the concept very useful and fascinating. Data cubes are the ideal choice for working with environmental data across vast geographical areas and extended time periods.

Week 2 (April 13, 2023): Programming with Julia

In week 2, I was able to complete the second part of the introduction to Julia script. While learning about one of Julia's key advantages – its exceptional speed – I felt slightly disappointed by the execution speed in my script. It to execute slower compared to my previous experiences with Python and R. However, during this seminar, I came to understand that Julia's apparent slowness is primarily a one-time occurrence on the first execution. This behaviour is an intentional design choice, aligning with Julia's mission to address the "two-language problem”. By offering the best of both worlds – high-level productivity and low-level performance – Julia tries to eliminate the need for developers to switch between languages. This principle also supports resource-intensive computations involving spatio-temporal datasets.

Week 3 (April 20, 2023): Data Cubes in Julia

This week I got to understand how to access data cubes with Julia. For that we used the “Zarr” and “YAXArrays” package to request the data over the “zarr” storage format. The subsets we can create from the data cubes are lazy. This refers to the concept of lazy evaluation, which means that the operation is not immediately performed after you request it. Instead, the calculation is delayed until you really need it. This is a huge advantage when working with large datasets like data cubes. The data I requested is stored on the DeepESDL data cube. I anticipate understanding this workflow to be very helpful for the project that we are going to work on.

Week 4 (April 27, 2023): Split-apply-combine

In this week we explored the split-apply-combine strategy introduced by Wickham in 2011. This widely adopted approach to data analysis offers a structured and efficient method for working with large datasets. With split-apply-combine, the dataset is initially divided into subsets. Then, a function can be applied to each subgroup. Finally, the results can be combined into a new data structure. This approach is applicable to operations on data cubes (as shown in Figure 1). I anticipate that split-apply-combine will play a pivotal role in our project's data analysis.

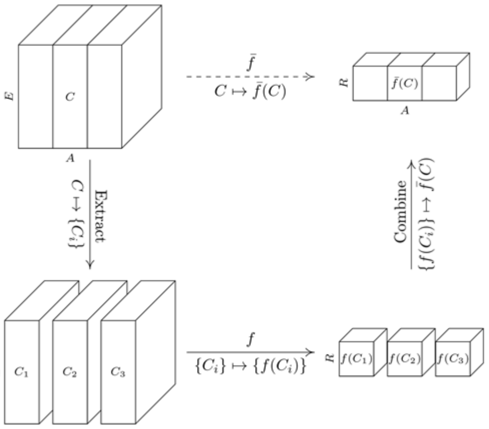

Figure 1: Schematic illustration of the “apply” functionality (Mahecha et al., 2020).

Week 5 (May 4, 2023): Memory Statistics

The most insightful part of today’s session (for me personally) was to understand how to do parallel processing in Julia. I learned that parallelization divides a task into smaller sub-tasks and executes them simultaneously. Like that I can run calculations in Julia with 8 threats on my laptop. This multi-threading function is improving the overall performance and efficiency of my code. I understand the importance of calculation speed for the work with large datasets. The differences of the calculations done in class between single-threading and multi-threading were severe. I will try to implement this technique into my future workflow.

Week 6 (May 11, 2023): PCA

In this week's session, we revisited the concept of Principal Component Analysis (PCA). Despite having delved into it thoroughly during the previous semester, revisiting this topic proved valuable due to its complexity. PCA is a vital statistical technique utilized for dimensionality reduction and data exploration. Its primary objective is to uncover a lower-dimensional representation of a dataset while preserving the most substantial part of its original variance. This is accomplished by identifying the principal components, which represent the directions of maximum variance within the dataset. Depending on the problem of our project, PCA remains a tool to consider using for tasks such as dimensionality reduction or noise mitigation.

Week 7 (May 25, 2023): Plotting in Julia

Julia offers a multitude of options for creating good visualizations. Among the diverse plotting libraries available, some were standing out to me: "CairoMakie" for versatile 2D plots, "GLMakie" for immersive interactive 3D visualizations, and "WGLMakie," a web-ready variant of "GLMakie" tailored for embedding interactive 3D plots in web applications and notebooks. In addition to the Makie ecosystem, Julia provides various other options for plotting, including "Gadfly", "Plots.jl" and "GR". Due to this versatility Julia seems to be the ideal choice for visualizing data cubes. It is notable that the “Makie” - documentation is by far better than other documentations I have seen so far. It is clear for me at this point that Julia is the preferred choice for our project, given its versatile set of plotting tools and documentation.

Week 8 (June 1, 2023): Exploring the "Global Data Set of Historical Yield" (GDHY)

I found a group for the project, and we began with the search for an intriguing topic. On our research we found a dataset that we think is very interesting. The dataset is called the "Global dataset of historical yields v1.2 and v1.3 aligned version" (GDHY), as introduced by Iizumi in 2019. It is a global dataset covering a period from 1981 to 2016. It combines national agriculture data with remote sensing data to modulate yield in t/ha (Iizumi & Sakai, 2020). We realized that there is substantial preprocessing work awaiting us for this dataset. While the data is conveniently formatted in netCDF4, each year and crop type has its dedicated data file. This means that to conduct a comprehensive time series analysis, we'll need to consolidate these yearly files into a stacked dataset. Additionally, we've recognized the necessity of rotating the data, shifting its longitudinal coordinates from 0° to 360° to the more commonly used range of-180° to 180°. We are thinking about correlating the crop yield dataset with climate variables sourced from the DeepESDL data cube.

Week 9 (June 8, 2023): Determining a Growing Season

This week we had to decide against using Julia for our preprocessing steps and data analysis. Although Julia appeared to be really feasible for the work with the DeepESDL data cube, we were deciding to continue with R and python. The reason is simply that one person in our group is not in the "Spatio-temporal data" course and therefore would prefer to not work with Julia. We accessed the data cube using the "xarray" package in python. We then chose a subset that is corresponding to the described crop yield data. Regarding the climate variables, we decided for daily mean air temperature at 2m and daily total precipitation. Since the GDHY dataset has insufficient data in the year 1981 and the precipitation data ends in 2015, we were creating a 32-year subset, spanning from 1982 to 2014. We came to the assumption that crop yield might be very depended on mean temperature and total precipitation of the crop specific growing season. Therefore, we calculated a seasonal temperature mean and seasonal total precipitation for April to September, aligning with the growing season for the selected crops and region, based on research by Sacks et al. in 2010. Finally, we merged the climate variables with the crop yield data.

Week 10 (June 15, 2023): Research Topic

We came up with a research topic: "The Impact of Temperature and Precipitation on Annual Crop Yield in North America". We decided to focus on North America for a multitude of reasons. First, it seemed very important for us to restrict our research to region with similar growing conditions. Furthermore, North America is a very relevant, well examined region regarding crop yield analysis. We decided to use the crop types maize, soybean and spring wheat, as they have a similar growing calendar. North America grew 32.88% of world's maize, 34.16% of world's soybean, and 8.7% of world's wheat in 2021 and is a major food supplier (Schlenker & Roberts, 2009).

Week 11 (June 22, 2023): Multi-language Workflow

At this point we realised, that we can’t do the project in a single programming language. While trying to achieve the programming part in a single language we ran into a few problems. Starting with accessing and preprocessing the data cube in R we tried out the just released bioconductor package called “Rarr”. Disappointingly, there was only little documentation about the package, which is why we couldn’t implement it in our code. On the other hand, we were not able to rotate the data (as described in week 8), using python. While the resampling part was easier in python, we preferred to do the statistical analysis in R. Like this we ended up with the following workflow:

Python: Accessing and preprocessing the data cube and crop yield data

R: Rotating and cropping the data

Python: Resampling and stacking the data

R: Running the statistical analysis

A more detailed description of this workflow is given in the code_merged.html file (https://github.com/mxkopf/crop_yield_climate_variables).

Week 12 (June 29, 2023): Detrending Crop Yield Data

By now we were almost done with preprocessing the data for our project and have been using the given time in the seminar to continue with the data analysis. Before starting the data analysis, we realized that we had to detrend the crop yield data. We did this by using the detrend function from the pracma package. We detrended the crop yields using linear regression, with crop yield as the dependent variable and year as the independent variable. This approach was employed to mitigate the influence of non-climatic factors, such as technological advancements, genetic improvements, and enhancements in crop and soil management practices, from the original time series of yield records. Detrending the crop yield is a crucial preprocessing step for our analysis.

Week 13 (July 13, 2023): Statistical Analysis

Last week we managed to create our first scatterplots regarding the correlation between total growing season precipitation and crop yield anomaly. The outcome was really devastating. Figure 2 shows the outcome of the first plots on the example of maize and precipitation. The plot shows very noisy data in form of a point cloud, with no correlation. We had to come up with a noise reduction method. Figure 3 shows the correlation of maize and precipitation after applying a binning technique, where the y-axis got rounded to every whole number precipitation value. In contrast to figure 2, we then calculated the mean crop yield anomaly for every rounded precipitation value. Like this we were able to find a non-linear relationship between crop yield and precipitation and a clear optimal total precipitation (over the growing period) value for each crop type and a declining crop yield above that optimum.

In the next weeks we want to improve the correlation maps we have created. We hope the maps to reveal spatial patterns regarding the correlation of the investigated crop yield and climate variables.

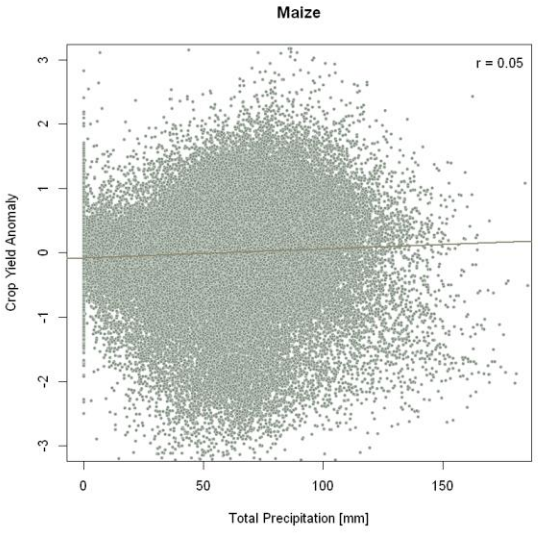

Figure 2: Maize yield anomaly and total precipitation.

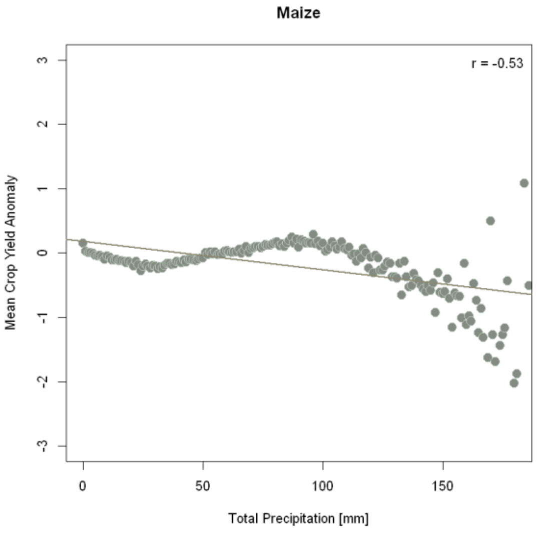

Figure 3: Average maize yield anomaly for each whole number total precipitation value.

Literature

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1–29. https://doi.org/10.18637/jss.v040.i01

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., & Reichstein, M. (2020). Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics, 11(1), 201–234. https://doi.org/10.5194/esd-11-201-2020

Sacks, W. J., Deryng, D., Foley, J. A., and Ramankutty, N. (2010). Crop planting dates: an analysis of global patterns: Global crop planting dates. Global Ecology and Biogeography. https://doi.org/10.1111/j.1466-8238.2010.00551.x

Schlenker, W., and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences 106, 37, 15594–15598. https://doi.org/10.1073/pnas.0906865106