Nick van Nuland
3774070
td85isud@studserv.uni-leipzig.de

Portfolio
Spatio-temporal Data — 12-GEO-M-DS02

2023-07-30

1. Session, Introduction (2023-04-06)

In today's session, we started by getting an overview of the spatio-temporal data class and its objectives. The lecture focused on the significance of spatio-temporal data in real-world applications. Throughout the class, we will be exploring how to work with data that combine both spatial and temporal components, so called "data cubes". To work with them in an efficient manner, we will use Julia, a high-level, high-performance programming language designed for technical and scientific computing. It is relatively simply while being very efficient. It aims to be both easy to use and to have a high computational performance, making it a popular choice for researchers and data scientists.
A significant part of today's session was dedicated to understanding big data and the challenges it poses, such as öut of memoryänd öut of core"computing. We learned about strategies to handle large data sets efficiently, considering the limitations of available computational resources.
Another topic covered was temporal and spatial auto correlation. While I have heard about both before, I found the figures really intuitive. They made it possible to grasp the concept instantly. Overall, the introductory session provided a good starting point for the class. I look forward to go deeper into the subject matter and exploring practical applications of spatio-temporal data analysis with our own projects later on.

2. Session, Introduction to Julia (2023-04-13)

Completing the Ïntroduction to Juliaëxercises has been exciting for me as someone who up until now has only worked with R and Python. While Julia's features and capabilities initially seemed very complex, the intuitive structure and beginner friendly exercises made it fun to start reading into it. I was pleasantly surprised by how quickly I grasped the language's syntax, and the concept of multiple dispatch, although new, became clearer with practice.
First I was wondering why code took so long to execute at first, but after understanding the just-in-time compilation, and seeing the speed at which my code executed left me feeling motivated to explore its potential further. The huge number of packages also presented a lot of possibilities for various projects, and I enjoyed experimenting with installing some of them during the course.
Although there are challenges, such as understanding some language nuances and finding specific packages, the learning resources provided some assistance. As of now, I am eager to continue coding in Julia, and improving my skills. While there is still much to learn, I am rather enthusiastic about the possibilities that Julia opens up for programming with earth data.
In terms of materials, I found Intro to Julia to be good. It was adequate in scope, and gave a good overview of many basic commands and applications. Especially the structure was clear, so that you could quickly read the relevant place in case of any ambiguity.

3. Session, Data Cubes in Julia (2023-04-20)

After last week's brief introduction to the data cubes, this week we took a closer look at them. The basic concept of data cubes is to make data easy to process and analyze. Stored in the form of large arrays as .zarr files, they make it possible to work with the data without time-consuming preprocessing. In Julia you need the package YAXarray to work with them, which we learned to install. As far as my knowledge of Julia goes for now, the language seems easy to learn in terms of its syntax. From previous projects I know how much effort the preprocessing of the data can be, if the data sets differ in temporal or spatial distribution, or metadata is missing. Collecting the data in one place has enormous advantages in my opinion, and working with the data in the course has been intuitive so far.

4. Session, Split-Apply-Combine (2023-04-27)

Today's lecture was about the split-apply-combine approach. During the lecture I had problems to wrap my mind around the concept. The basic idea was clear to me, but I did not understand why operations are not performed on the whole data set and also the process of how exactly to split the data sets was unclear to me. It was only during the follow-up that I understood the embarrassingly parallel problem and its relevance for

the split-apply-combine approach. Due to the structure of the data cubes in large arrays, it is computationally expensive to perform operations on them. If the same operation is performed very often, they are performed one after the other and lead to a high execution time. This is especially a problem when processing geodata, since values such as the mean often have to be calculated over the entire time period. With split-apply-combine it is possible to create multiple data sets from the data cube, and then run the computational operation in parallel over the split data sets. How the split looks depends on which dimension you want to calculate what over and of course the dimensions of the input array. After applying the command, the modified slices can be put together again in a data cube, which now has new values.
The literature references at the end of the presentation helped me the most with the material provided, as they ultimately led me to understand the principle of split-apply-combine. The paper from Wickham in particular was informative and understandable for me.

5. Session, O(1) Memory Statistics (2023-05-04)

In this lecture, we learned about the Julia package 'weighted online stats' and explored how it can help analyze spatio-temporal data. At first, I had a hard time understanding why it is useful when dealing with big data sets. However, with the help of extra resources and examples, I started to get it.
In the context of out of memory calculations, weighted stats is intended to perform calculations when the amount of data does not fit in the memory. To make this possible, an online algorithm is used. it makes it possible to process data one by one in contrast to offline algorithms, where the entire data set must be available to calculate a result. after i understood the principle of online stats in general, i also understood weighted online stats, as the packages are very similar in function, except for the weighing factor which makes it possible to put some special focus on certain variables. the documentation of OnlineStats.jl helped me the most, because it explained the basic functionality. the documentation on weightedOnlineStats.jl, on the other hand, was still expandable, and little was explained there.

6. Session, O(1) Memory PCA (2023-05-11)

In today's lecture we dealt with dimension reduction and principal component analysis. since I was already familiar with these topics and therefore had no new input, I would rather reflect on the topic selection of our group project.
currently our plan is to correlate land cover change and different climate variables on a global scale. The idea is to visualize the resulting land cover changes as the climate changes over time. We hope that our results will then show that, for example, in regions where precipitation has declined sharply over the period under review, the land cover will change accordingly. Perhaps we will even be able to identify previously undetected inter plays between climate variables and land cover change. However, there are still uncertainties in the data basis, since we would have to use a land cover product from outside the data cube and are uncertain whether the possible undercutting period is sufficient to detect large-scale changes in the expected results.

7. Session, Plotting in Julia (2023-05-25)

in today's lecture we covered plotting in Julia. A variety of packages were presented, some specific for certain use cases, others general. besides the funny and actually practical unicodePlots.jl, which allows you to plot in the terminal, I especially remembered Gadfly.jl. Since I am much more familiar with R than with Julia and thus have used ggplot2 more often, I took a closer look at the documentation.
I was surprised by the presented possibilities and the scope and did not expect to recognize not only a similar syntax and the commands of ggplot2, but also to find new functions that do not exist in ggplot2 itself. However, when trying to work with the package myself, I encountered problems, which are probably partly due to the fact that I have not yet mastered the syntax of Julia sufficiently and that there are some differences between ggplot2 and Gadfly. In summary, it seems to be a good way to plot with Julia without having to get used to it. The detailed documentation of Gadfly linked in the presentation was especially helpful for me.

8. Session, Working on Project Analysis (2023-06-01)

Today there was no new input and we are supposed to continue working on our analysis. In the meantime, we have made some changes to our initial project. due to the concern that land cover change is too "coarse"to measure changes, we now want to measure crop yield per pixel. After doing some modeling, it became apparent that the changes from climate factors to land cover change were not very significant. globally, the

focus was much more on direct human impacts such as deforestation and the like, making it difficult to analyze in terms of climate variables. instead, we now want to use crop yield, as this provides more promising data at first glance and does not have the same problems as land cover change. However, there are other challenges here, as the data set needs to be detrended to account for increasing yields due to advances in more efficient farming. We found a suitable crop yield data set called the GDHY, the global data set of historical crop yields, which spans from 1981 to 2016 and is based on remote sensing data.

9. Session, Working on Project Analysis (2023-06-08)

This week was largely spent on data preprocessing. Since the GDHY data set is not available in the data cube, we need to regrid the data and bring it to the same spatial resolution as the climate variables. Also we need to make sure that the grid cells of both data sets align, otherwise the later analysis will be erroneous.
We have now decided not to perform the analysis in Julia, but in Python. Since the analysis is to be completed in the foreseeable future and all group members do not yet feel particularly confident in using Julia, we have finally decided to do this, also in view of the fact that we have also written all the code in Python up to now, and parts of the previous code, even if it was only for tests, will also be relevant for the actual analysis. This is a pity considering that I personally found Julia very intuitive to use and think that the language would potentially be very well suited for an analysis like ours. However, due to the time and the sometimes poor documentation of the Julia packages, I don't see any other option.

10. Session, Working on Project Analysis (2023-06-15)

After taking a closer look at the data, we noticed that there are large data gaps for many regions of the world. Part of the gaps can be explained by the nature of the data, which can only be collected in regions where the crop is grown. However, there are also large gaps over several years in many regions that would be suitable for the cultivation of the respective crop. we have therefore decided to limit our analysis to North America, as there is a relatively complete data set there. In the meantime, we have a precise plan of what we want to analyze and have already started to do so in part. We want to correlate the crop yield of maize, soy and spring wheat over North America with the climate variables there, over a period of 30 years. So far, we have preprocessed the GDHY data set and detrended the data set to account for the systematic increase in crop yield due to agricultural progress. Thus, we ultimately see a time series of crop yields characterized by deviations (particularly good or bad yield years). We then compare whether the underlying climate variables also deviate strongly from the norm in these years.

11. Session, Working on Project Analysis (2023-06-22)

This week we finished a first analysis. We have now determined for each crop type how high the correlation between the climate variables and the annual crop yield is. The resulting maps thus indicate how much a variation in the environmental factors of precipitation and temperature affects crop yield. However, the values obtained show that this influence is not particularly high. For most of North America, the r^2 is about 0.1, which means either that the climate factors have little effect or that there are other weaknesses in the analysis. later, we will take a closer look at possible causes of error, but for now we will concentrate on completing the analysis.

12. Session, Working on Project Analysis (2023-06-29)

This week we have taken some final steps in the analysis. These include the creation of a map that describes which of the two climate variables, temperature and precipitation, is predominant in which region. it explains where the crop yield is influenced more by precipitation and where more by temperature. Due to the generally low r^2 value, however, the significance is limited, as in some cases no statistically relevant statements could be made.

13. Session, Working on Project Analysis (2023-07-13)

By now, we completed our analysis. We have worked out which parts of North America are dominated by which climate variable in terms of crop yield, and while our analysis has produced interesting e results, it does have some uncertainties. In hindsight, it would have been necessary to define the expression of the climate

variables differently, since a simple mean value for temperature and an accumulated value for precipitation are not sufficient to describe the course over the entire growing season. a drought at the beginning of the growing season can be offset by a wet second half of the growing season, so that a very unfavorable year for crops could appear to be average. During the analysis with Python, we sometimes encountered problems with the accessibility of the data cube, so that we switched at one point to downloading the data from the data cube and working with the data locally, even though this is not the actual intention of the whole concept of datacubes. Nevertheless, I consider the project a success in the sense of a spatio-temproal data course, as it has made me more familiar with working with this type of data.