Assignment for Spatial data scientist position 2023/05

Objective

Manipulating and combining various spatially explicit datasets into indicators of environmental impacts associated to land use activities, and making sense of those, is part of the core tasks of the advertised position. The following exercise aims at testing your capabilities in this domain, with an application to agriculture and the nitrogen cycle.

The goal of the exercise is to a) combine various datasets to generate indictors of nitrogen loss to the environment associated with wheat production at various spatial scales, b) provide graphical representations and conduct simple comparisons across a few countries and c) provide a reproducible code associated to these tasks.

At best, the tasks in this test would be performed with the R software exclusively. However, if you are not familiar enough with it and more proficient in other software, answers would still be accepted.

The exercise is expected to take up to 3 hours to complete for an experienced candidate, but no strict time limitation is imposed. We request, however, the results and related code to be submitted back to us no later than 48 hours after reception of the instructions.

Candidates will be evaluated on the following criteria:
– accuracy and quality of the presented outcome (taking account of the time given for the tasks),
– relevance of the analysis of impacts (taking account of the time given for the tasks),
– clarity of the code (candidates are requested to provide their code with their results, see instructions for use of GitHub further down),
– reproducibility of the results.

Datasets

The data you need to perform the assignment are available on the following Dropbox link
https://www.dropbox.com/sh/uv3qzqvs45j7gup/AAAfkMPVTOQnaP4JXOKuaA7Xa?dl=0

The main folder contains the following datasets in subfolders:

- the SPAM dataset[1] (SPAM_2005_v3.2 subfolder), providing at global scale in a raster format (5 arcminute spatial resolution) estimates of yield, physical area and harvested area for the year 2005 (see readme file, further documentation in SPAM 2005 Technical Documentation.pdf);

[1] You, L., U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. 2014. Spatial Production Allocation Model (SPAM) 2005 v3.2. [October 12, 2017]. Available from http://mapspam.info.

- the GAUL dataset (GAUL subfolder), providing at global scale boundaries of administrative units (polygons at sub-national resolution, country indicated by ADM0_NAME field) in a vector format;
- a country-level dataset of nitrogen use efficiency (NUE) for cropland based on Zhang et al 2015[2] (NUE_Zhang_et_al_2015 subfolder). NUE is defined as the ratio of the amount of nitrogen content in harvested product divided by the amount of nitrogen applied to the field in various forms – fertilizer, manure etc. (more information in the original article, available in the same folder);

Tasks

1) Using SPAM raster data, produce a new raster at the same resolution, containing wheat production volume (in million tons Mt). Produce a global map and export the raster in a geotif format.

2) Using the newly created raster and the GAUL shapefile of administrative borders, aggregate production to country level and export to a csv file.

3) Using the raster of wheat production generated in question 1, and assuming that 2% of harvested wheat yield consists of nitrogen (N) element, create, plot and export a raster of N output in harvested wheat yield.

4) Using the dataset of country-level nitrogen use efficiency (NUE) of wheat from Zhang et al 2015, and steps from previous questions:
a. estimate for the 10 biggest wheat producers the country-level values of N output in harvested wheat, as well as related total N inputs and N losses (i.e., surplus) and export the dataset as a csv file
b. visualize the N outputs and losses for these 10 countries in 1 summary figure (plot to be exported as pdf)
c. explain in 2-3 sentences the main patterns of N losses across countries, in relation to production volume and NUE (including any singular feature).https://iiasa.github.io/iBIOM/en/main/

5) Explain in 2-3 sentences how an analysis like the one performed in previous tasks could translate to the models within BNR's modeling suite (https://iiasa.github.io/iBIOM/en/main/), including potential limitations.

6) Please report any issues you encountered or assumptions that you had to make in order to complete the assignment.

Reproducibility of the results

In order to check the reproducibility of your results, we request you to send us an access link or a copy of your code with the exercise outcome. With the code file(s), you will provide a README file giving clear instructions on how to reproduce your results, including software dependencies (i.e. operating system and other software versions), as well as the detailed steps to reproduce and visualize your results. README file should also contain the text answer to questions 4, 5 and 6.

[2] https://doi.org/10.1038/nature15743

We strongly recommend the use of GitHub for this purpose. Here are below the few steps to use GitHub.

1) Create a public GitHub repository for your project on the website http://github.com/. We recommend you to use a name that cannot be directly linked to the exercise or recruitment process (to remain invisible to other candidates).

2) Create a local copy of the repository on your computer to develop your project. Use your common “Git” version if you are used to this versioning system. If you are not used to it, we recommend you rely on the “Quick Set-up” option [Set up in desktop app] proposed by the website.

3) In your project repository create:
– a folder called "data" (tell git to ignore all commit from this folder, using the gitignore function).
– a "README.md" file (https://help.github.com/articles/about-writing-and-formatting-on-github/)

4) Create a release version of your GitHub project (https://help.github.com/articles/creating-releases/) and send the link to the release to us.

Last important information

- Make sure that we can run your final code and reproduce your results before sending us the link, including the visualization.
- Do not include the data in your GitHub project. We already have them.
- If you need a projection system for using the GAUL shape file, WGS84 is then recommended.
- Even if no strict control of time is performed for this exercise, try not to spend more than a day performing it. If you do not succeed in completing your implementation, still submit your final state of code development and explain the difficulties encountered and missing steps.
- Finally, your analysis and code used for this technical exercise will not be utilized by IIASA or shared outside of purposes of this application.