Future Food Production project [report 1]

OPTIONAL PROJECT IN DATA SCIENCE
8 CREDITS

Supervisors:

CHARLOTTE WEIL, Stanford University

ROBERT WEST, EPFL

Student:

ROMAIN CARISTAN, EPFL

EPFL

April 2019

Table of Contents

1. Project description

a. Motivation: Exploring resilience of Food Sufficiency by 2050

This work aims to explore the resilience of the food system by the year 2050 using global statistical modeling of the crop yields production. The goal is to analyze the variability of caloric sufficiency changes, i.e. to identify where are the hotspots of resilience or vulnerability in term of food sufficiency in regard with the different changes in land use, population, climate, diet or socio-economic status. This study is done under different future climate scenarios, 5 Shared Socio-Economic Pathways (SSP), each representing a different path in both climate and society changes described in (O'Neill et al., 2014) and compared to the 2000 data. A schema summarizing these SSPs can also be found in *Figure 1*. The different scenarios are predicted using different climate models, 4 in our case: GISS-E2-R, HadGEM2 ES, MIROC-ESM and CCSM4.

In this work, we chose not to limit us to a single or few crop types as it was done in the previous works about this subject, e.g. in (White, Hoogenboom, Kimball, & Wall, 2011), the studied crops were only were wheat, maize, soybean and rice. Here we measure the resilience as a "sensitivity to change" (Turner et al., 2003) and want to study how the different parts of the world will react to those changes, thus we take into account the fact that the produced crop mix in a place will evolve with the different soil, climatic and economic parameters. That is why we chose to focus on the total calories produced and not to limit us to a few crop types.

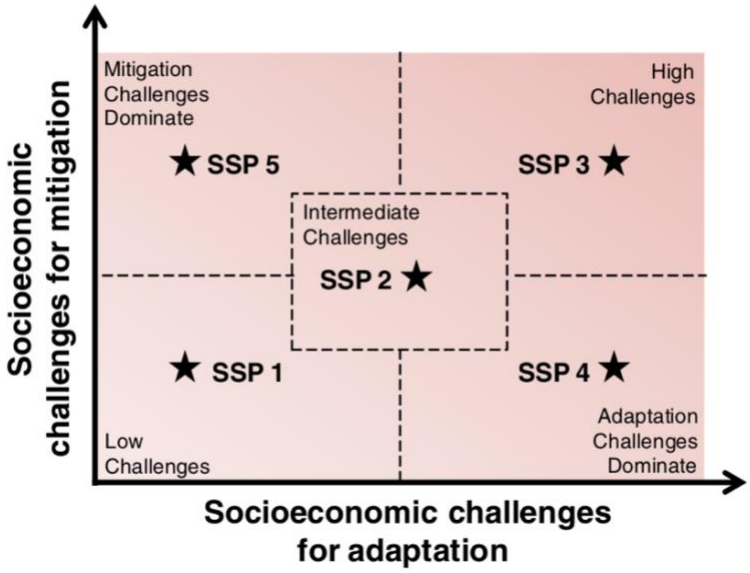

Figure 1: The "challenges space" to be spanned by SSPs

b. Project objective

The objective of this project is to improve the proposed regression approach predicting the aggregated caloric yields. The model is trained on a dataset containing crop yields in 2000 associated with multiple land and climate characteristics paired with some information about the country GDP, and is to be applied to each future scenario in 2050 made of a prediction of each of the 2000 features. As we did not have enough data to build a testing set in one of the years between 2000 and now, we need to do a certain amount of data exploration before building the model to make sure that the 2050 data is consistent with the 2000 one (i.e. that the 2050 feature vectors lie in the space spanned by the 2000 vectors).

For context (but this is beyond the scope of this student project), these regression outputs are then used to assess global food sufficiency in 2050 and analyze the results in regard to the sufficiency of the different countries and regions of the world in 2000.

c. Description of the data

The dataset is built using geographic data from different sources, the whole globe is divided into more than 9 million pixels, and around 900 000 if we discard the pixels where we have no information, e.g. points in the ocean, desert or ice floe. These pixels represent squares of 5 arc minute resolution, i.e. about 100 km^2 squares. For each of the pixels, we have different soil, climate and economic parameters as inputs. For the 2050 dataset, we get the same inputs but this time they are predicted values.

The crop yields consist of 175 of the main crops (Monfreda, Ramankutty, & A. Foley, 2008) aggregated into a single metric of crop calories per hectare to form the output of our model. This aggregation is done, as previously stated, to enable us to take into account the evolution of the crop mix.

The climate data for 2000 comes from (Fick & Hijmans, 2017) and is available in WorldClim Version 2. It consists of monthly averaged values from 1970 to 2000 for each of the climatic variables:

- Annual Mean Temperature,
- Mean Diurnal Temperature Range,
- Isothermality,
- Temperature Seasonality,
- Max Temperature of Warmest Month,
- Min Temperature of Coldest Month,
- Temperature Annual Range,
- Annual Precipitation,
- Precipitation of Wettest Month,

- Precipitation of Driest Month,
- Precipitation Seasonality (Coefficient of Variation)

We also have some topographic data within the slope and altitude columns.

The fertilizers, on the other hand are separated between the annual (ann) and perennial (per) areas with the C_3, C_4 and Nitrogen (nfx) fixing crops. It consists of the columns:

- fertl_c4ann,
- fertl_c4per,
- fertl_c3ann,
- fertl_c3per,
- fertl_c3nfx

where the unit is kg*N/ha/year/crop_season. We also add to our data some irrigation information from (Hurtt, Chini, Frolking, & Sahajpal, 2016) of the same form as the fertilizers except the unit which is here a fraction of the crop area. To do so we extract the useful data of the NetCDF files and save it to TIFF files. Then we resample these files to get them in the 5 arc minute resolution as the rest of our data. The last step is to convert them to CSV files and join them to each data file to get our final datasets for 2000 and for each climate model and scenario of 2050.

Another column represents the GPD per capita of each country in 2000 and 2050 (a predicted value), this data is used as a proxy for the mechanization of a country. A future work will be to try to get a value of the GDP per capita at a lower scale to get more precise information in the pixels.

Finally we have some soil data under the form of indices for the values considered to be influencing the most yield performance by (Fischer et al., 2012) in the columns:

- workability conditions,
- toxicity conditions,
- rooting conditions,
- oxygen availability,
- nutrient retention,
- nutrient availability,
- protected area status,
- excess salts

All the information about the geographic location (latitude, longitude) of the points is voluntarily removed from the data as this would bring bias to the model and make it use only geographic information for the prediction of the 2050 yields.

2. Data exploration

a. Features distribution

We start by looking at the data to see how the features are distributed to extract all the categorical ones to do a proper feature standardization. We can see from *Annex 1* that the categorical features to discard for the standardization are all the columns containing the word "index" in their name, i.e. all the soil data. We then keep the non-categorical ones to do a standardization of each column, useful later to compute the distance between each feature and to apply the model. Each non-categorical feature on the other hand is normalized by subtracting to each column its minimum and then dividing by its maximum.

The output distribution (i.e. the log calories per hectare produced) can be also found in *Annex 1* highlighted by the red square. We will then for the future "drop" the log and only consider the raw number of calories produced per hectare for a more precise idea of the order of magnitude of the values.

b. Aggregation assumption

The premise of this work assumes that farmers will grow "what makes sense to grow" given climate, soil and societal conditions on their landscapes. This translates, as explained above, in aggregating caloric yields of all crops, and assuming, the crop mix (thus total caloric yields of nutritionally relevant crops) will be consistent given climate and societal conditions. Unlike previous work in the field, modeling yields crop by crop, we take a coarser picture of the total caloric yields of all nutritionally-relevant crops, thus embedding the adaptation assumption (i.e. that famers will adapt the crop mix to be relevant to new climate and societal conditions). Hence we want to prove that in our 2000 data, similar land conditions will lead to similar yields in the output.

To do so, we first need to define a metric to measure the similarity between two points in our dataset, the Euclidean distance seems to be a good fit to our problem as each data points is composed of many standardized/normalized features. The number of points being extremely large (more than 900 000 if we discard all the points where we have no information, e.g. points in the ocean), it is clear that it will be inefficient to compute all the distances between all the points, that is why we decide to compute, for each of the points, its distance to another randomly selected point of the data. This gives us a consistent representation of the distances while making the computation fast. We find, taking the mean and the standard deviation of all the computed Euclidean distances, that the average distance between two points is 6.25 and the standard deviation is 2.63. The features distance distribution can be found in *Figure 2*.

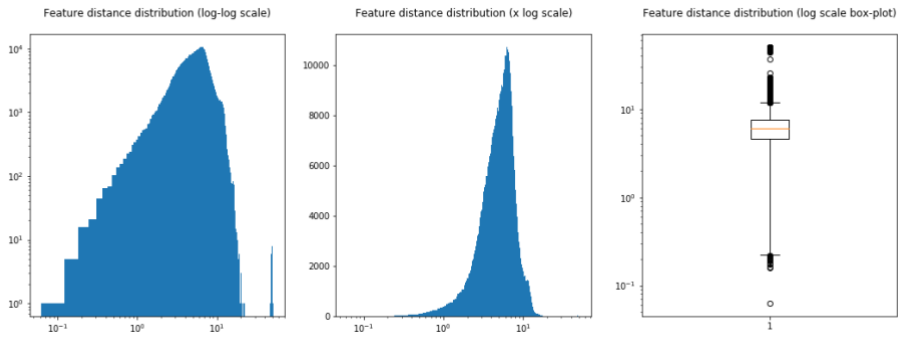

Figure 2: Features distance distribution

The next step is then to study the similarity of the outputs in term of point distances, to do so, we again sample the points for computation efficiency issues, this time taking 4 million random pairs of points with their respective feature Euclidean distance and associated output distance (i.e. the produced calories difference). We then round all the feature distances to their first decimal to create 0.1 bins of distance. The goal is to see what happens to the output when we increase/decrease the input distance (distance between feature vectors), hence increase/decrease the land condition difference. The plot of the mean calories produced per hectare difference for each created bin can be found in *Figure 3* with a 95% confidence interval. This plot comforts us in the choice of making the aggregation assumption. In fact, the more we decrease the distance between the points, i.e. the more the points have similar climate, soil or economic characteristics, the narrower is the range of the output value, i.e. the closer becomes the number of calories produced per hectare.

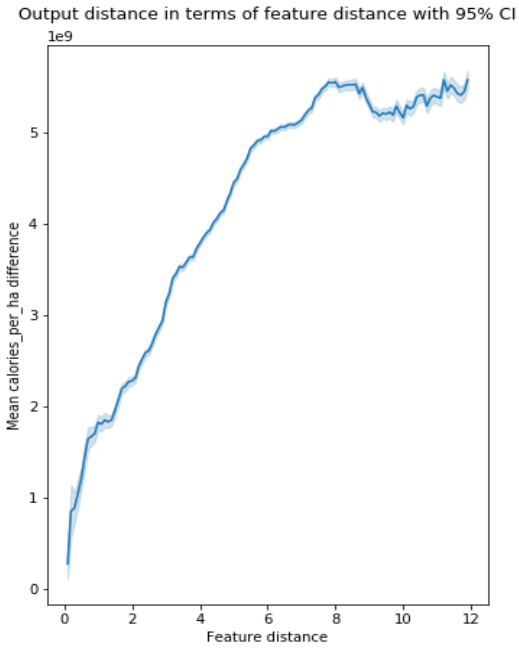

*Figure 3: Output distance in terms of input distance, average output distance: 4.7*10⁹ calories*

c. Clustering the data

Now that we are comforted in our crop aggregation assumption, we want to see if we can split the 2000 points in different categories and how the 2050 data predictions interact with these points and categories. The first step is to build clusters for our reference datasets and then show them on a map. We use a K-Means, with k=10[1] (11 counting the points where we have no information as -1 label), algorithm to create our clusters which can be found in *Figure 4*, where we observe that without any prior knowledge about the geographical position of the points, we are able to recognize the main regions of the world through the label assignment.

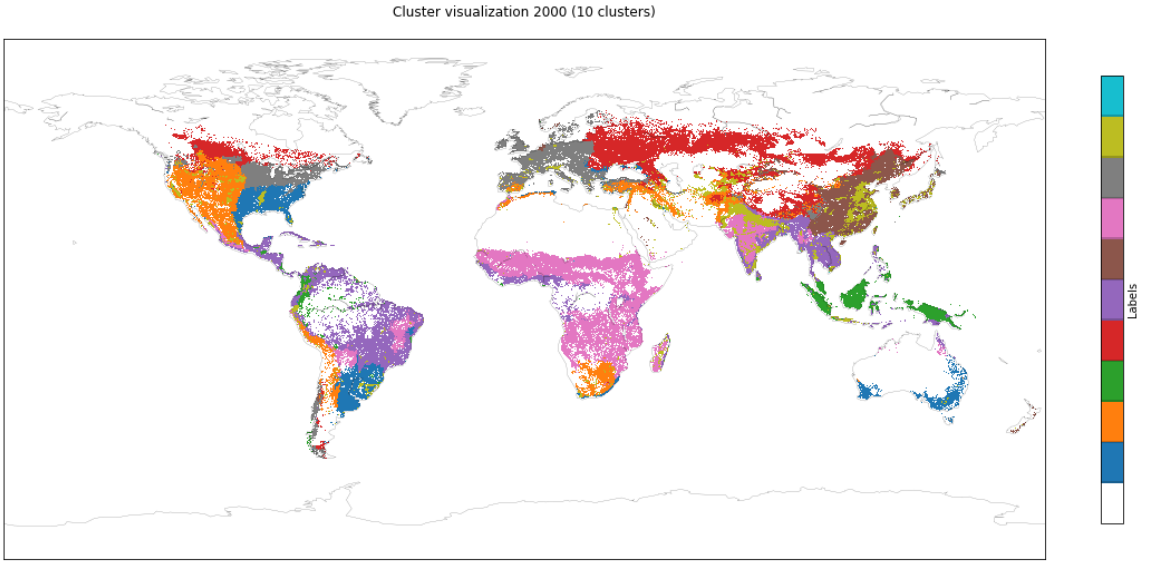

Figure 3: Map of the clusters for the 2000 data

A valuable information for our future work on the model will be to spot if some of the points of the 2050 predictions change their cluster assignment compared to 2000 and if some of the points do not even lie in the space spanned by the 2000 feature vectors. Thus for each SSP we predict, using the K-Means model used to create the 2000 clusters, their new cluster assignment. In addition, we build what we call a min-max convex-hull and check the proportion of the points for each scenario that are not in this hull. A proper convex-hull was not efficiently computable due to the number of points and features, this is why we limit us to the use of the min-max convex-hull that checks, for each 2050 point, if all of its features lie in the min-max range of the corresponding 2000 feature. If it is not the case, the point is considered outside the min-max hypercube and hence outside the convex hull. We provide

[1] Multiple parameters were tested for k, the number of clusters but k=10 was the value providing the more valuable information about the clusters when plotting on the map.

an example of one the resulting maps with the new clusters and the points outside the min-max convex-hull for the climate model HadGEM2 ES in *Figure 4*. Analyzing the points outside the hull, we find that the most common feature responsible for their eviction is the GDP per capita. The main explanation is that in some of the future scenarios (in particular SSP4 and SSP5), the GDP is planned to increase for most of the countries, thus excluding some of them of the hypercube. These points represent points that no similar points in the 2000 dataset, making their 2050 yield prediction more tedious.

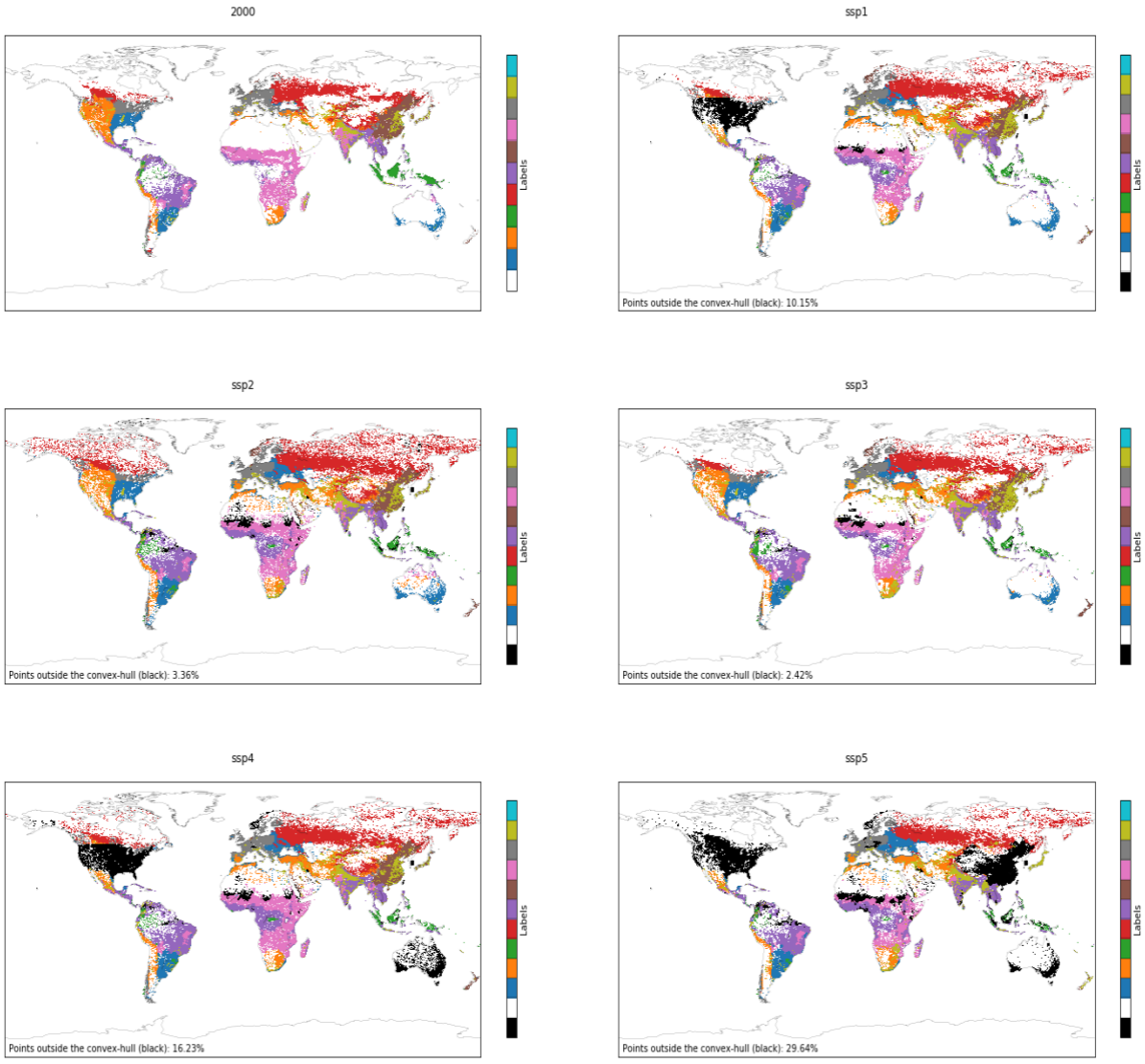

Figure 4: Maps of the 2050 points mapped to the 2000 clusters and the points outside the 2000 min-max convex-hull (HadGEM2 ES)

3. Future work (April)

The focus of this month will mostly be on the model part.

- Divide the GDP into income bins (low, medium, high income) to solve problem of many points outside of min-max convex-hull
- Retrain the model on the new data, test other models and compare outputs, sensitivity analysis on the results (which are the most influencing features)

4. References

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fischer, G., Nachtergaele, F., Prieler, S., Velthuizen, H. T. van, Verelst, L., & Wiberg, D. (2012). *Global Agro-ecological Zones Model documentation*. 179.

Hurtt, G., Chini, L., Frolking, S., & Sahajpal, R. (2016). Land Use Harmonization 2. Retrieved from http://luh.umd.edu/data.shtml

Monfreda, C., Ramankutty, N., & A. Foley, J. (2008). Farming the Planet: 2. Geographic Distribution of Crop Areas, Yields, Physiological Types, and Net Primary Production in the Year 2000. *Global Biogeochem. Cycles*, *22*, GB1022. https://doi.org/10.1029/2007GB002947

O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., … van Vuuren, D. P. (2014). A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Climatic Change*, *122*(3), 387–400. https://doi.org/10.1007/s10584-013-0905-2

Turner, B. L., Kasperson, R. E., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., … Schiller, A. (2003). A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences*, *100*(14), 8074–8079. https://doi.org/10.1073/pnas.1231335100

White, J. W., Hoogenboom, G., Kimball, B. A., & Wall, G. W. (2011). Methodologies for simulating impacts of climate change on crop production. *Field Crops Research*, *124*(3), 357–368. https://doi.org/https://doi.org/10.1016/j.fcr.2011.07.001

5. Annexes

Annex 1: Distribution of original features and output of the 2000 data