README FILE

Paper: "Geography and Agricultural Productivity: Cross-Country Evidence from Micro Plot-Level Data?"

Authors: Tasso Adamopoulos and Diego Restuccia

1. Data Availability Statement (DAS)

The datasets used and access instructions are provided below.

- **GAEZ (2000). Global Agro-Ecological Zones (GAEZ), version 3.0. Food and Agricultural Organization (FAO) and International Institute for Applied Systems Analysis (IIASA).**

The data can be accessed in ascii format from: https://www.gaez.iiasa.ac.at/
Access to the data is free but requires registration with an email and a created password, that can be used thereafter. The website recommends using Firefox, Chrome or Opera browsers for best results.

- **TM (2008). World Borders Dataset. Thematic Mapping.**

Access to the data is free without registration from:
http://thematicmapping.org/downloads/world_borders.php
The data can be downloaded in shapefile format by clicking on: TM_WORLD_BORDERS-0.3.zip
Projecting the shapefiles into maps requires a licensed software, ArcGIS 10.4 or later.

- **Alan Heston, Robert Summers and Bettina Aten, Penn World Table Version 6.3, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, August 2009.**

The data can be accessed in excel format from:
https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt-6.3?lang=en
Access to the data is free and does not require registration.

- **Harvest Choice (2012). Global Spatially-Disaggregated Crop Production Statistics Data for 2005 version 3.2. International Food Policy Research Institute (IFPRI) and International Institute for Applied Systems Analysis (IIASA).**

The data can be accessed in ascii format from Harvard Dataverse, V9, 10.7910/DVN/DHXBJX at:
https://doi.org/10.7910/DVN/DHXBJX
Access to the data is free and does not require registration.

- **UNPS (2009). Uganda bureau of statistics, national panel survey (UNPS) 2005-2009. World Bank, (Ref. UGA 2005-2009 UNPS v01 M.).**

The data can be accessed in STATA format from the World Bank's Microdata Library at:
https://microdata.worldbank.org/index.php/catalog/1001/get-microdata
Access to the data is free but requires registration with an email and a created password, that can be used thereafter.

- **FAOSTAT (2000). Production statistics. Food and Agricultural Organization (FAO), (Value of Agricultural Production, Area Harvested).**

The data can be accessed in csv format from: http://www.fao.org/faostat/en/#data
Access is free and does not require registration.

- **USDA (2015). USDA National Nutrient Database for Standard Reference, Release 28. US Department of Agriculture, Agricultural Research Service, Nutrient Data Laboratory.**

The data can be accessed in ascii format from:
https://www.ars.usda.gov/Services/docs.htm?docid=8964
Access is free and does not require registration.

2. Grid-Level Data Files

All necessary files are contained in the subfolder `/Grid_Data_and_Matlab_Files/`. Raw global grid-level data files downloaded from the Global Agro-ecological Zones (GAEZ 2000) v3.0, in ascii format. There is one grid-level file for each crop, for each of four variables. The variables are:

- **Land**: set of "lg_x.dat.txt" files, where "x" is the crop (e.g., wheat, maize, etc.)
- **Output**: set of "yg_x.dat.txt" files, where "x" is the crop (e.g., wheat, maize, etc.)
- **Potential yield – rainfed, low inputs**: set of "YLDrl_x.dat.txt" files, where "x" is the crop (e.g., wheat, maize, etc.)
- **Difference between potential and actual output – rainfed & irrigated, mixed inputs (GAEZ baseline)**: set of "YDm_x.dat.txt" files, where "x" is the crop (e.g., wheat, maize, etc.)

Below is a complete list of all the raw data files from GAEZ used in the experiments.

Crops	Land File	Output File	Potential-Actual Output Difference (mixed) Files	Potential Yield (low) Files
wheat	lg_wheat.dat.txt	yg_wheat.dat.txt	YDm_wheat.dat.txt	YLDrl_wheat.dat.txt
rice	lg_rice.dat.txt	yg_rice.dat.txt	YDm_rice.dat.txt	YLDrl_wetrice.dat.txt
maize	lg_maize.dat.txt	yg_maize.dat.txt	YDm_maize.dat.txt	YLDrl_maize.dat.txt
sorghum	lg_sorg.dat.txt	yg_sorg.dat.txt	YDm_sorg.dat.txt	YLDrl_sorghum.dat.txt
millet	lg_millet.dat.txt	yg_millet.dat.txt	YDm_millet.dat.txt	YLDrl_pmillet.dat.txt, YLDrl_fmillet.dat.txt
other cereals	lg_othcer.dat.txt	yg_othcer.dat.txt	YDm_othcer.dat.txt	YLDrl_barley.dat.txt, YLDrl_rye.dat.txt, YLDrl_oat.dat.txt, YLDrl_buckwheat.dat.txt, YLDrl_dryrice.dat.txt
potato	lg_potato.dat.txt	yg_potato.dat.txt	YDm_potato.dat.txt	YLDrl_sweetpotato.dat.txt, YLDrl_whitepotato.dat.txt
cassava, yam, cocoyam	lg_yam.dat.txt	yg_yam.dat.txt	YDm_yam.dat.txt	YLDrl_cassava.dat.txt, YLDrl_yamcoco.dat.txt
sugar beet	lg_sbeet.dat.txt	yg_sbeet.dat.txt	YDm_sbeet.dat.txt	YLDrl_sbeet.dat.txt
sugar cane	lg_scane.dat.txt	yg_scane.dat.txt	YDm_scane.dat.txt	YLDrl_scane.dat.txt
pulses	lg_pulses.dat.txt	yg_pulses.dat.txt	YDm_pulses.dat.txt	YLDrl_phbean.dat.txt, YLDrl_chickpea.dat.txt, YLDrl_cowpea.dat.txt, YLDrl_drypea.dat.txt, YLDrl_grams.dat.txt, YLDrl_pigeonpea.dat.txt
soybean	lg_soy.dat.txt	yg_soy.dat.txt	YDm_soy.dat.txt	YLDrl_soy.dat.txt
rapeseed	lg_rapese.dat.txt	yg_rapese.dat.txt	YDm_rapese.dat.txt	YLDrl_rapese.dat.txt
sunflower	lg_sun.dat.txt	yg_sun.dat.txt	YDm_sun.dat.txt	YLDrl_sun.dat.txt
groundnuts	lg_ground.dat.txt	yg_ground.dat.txt	YDm_ground.dat.txt	YLDrl_ground.dat.txt
oilpalm	lg_oilpalm.dat.txt	yg_oilpalm.dat.txt	YDm_oilpalm.dat.txt	YLDrl_oilpalm.dat.txt
olive	lg_olive.dat.txt	yg_olive.dat.txt	YDm_olive.dat.txt	YLDrl_olive.dat.txt
cotton	lg_cotton.dat.txt	yg_cotton.dat.txt	YDm_cotton.dat.txt	YLDrl_cotton.dat.txt

Cells in the grid are allocated to countries using the United Nations borders file `uncellsnew.txt`.

3. Matlab Files

All necessary files are in the subfolder `/Grid_Data_and_Matlab_Files/`and should be run from this folder. All Matlab programs are run on `Matlab R2020b`.

PPexper_low.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for all crops under the rainfed - low input scenario. The program produces an aggregate production potential output and aggregate actual land and output for each country in the sample. The default uses FAO prices to calculate actual output and production-potential output for each country. To compute actual output and production-potential output with caloric prices, comment out lines 877 & 881, and uncomment lines 876 & 880, before running the program.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns A-C.

PPexper_low_wheat.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for wheat under the rainfed - low input scenario. The program produces, for wheat only, an aggregate production-potential output and aggregate actual land and output for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns E-G.

PPexper_low_rice.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for rice under the rainfed - low input scenario. The program produces, for rice only, an aggregate production potential output and aggregate actual land and output for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns I-K.

PPexper_low_maize.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for maize under the rainfed - low input scenario. The program produces, for maize only, an aggregate production potential output and aggregate actual land and output for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns M-O.

SPexper_low.m

Uses the GAEZ grid-level data files and runs Experiment 2 on "Spatial Potential" for all crops under the rainfed - low input scenario. The program produces aggregate counterfactual spatial-potential output and land for each country in the sample, except Brazil, Canada, China, India, Kazakhstan, Russia, United

States. For the spatial-potential results for these countries see Section 3 of this README file on "Python-Gurobi Files."

Matlab toolbox required: Optimization Toolbox.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns Q-R.

TPexper_low.m

Uses the GAEZ grid-level data files and runs Experiment 3 on "Total Potential" for all crops under the rainfed - low input scenario. The program produces an aggregate total-potential output and aggregate land for each country in the sample. The default uses FAO prices to calculate the total potential output for each country. To compute the total potential output with caloric prices, comment out lines 273-275, and uncomment lines 277-279, before running the program.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns T-U.

PPexper_mixed.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for all crops under rainfed & irrigated water supply, and the mixed input scenario. The program produces an aggregate production potential output and aggregate actual land and output for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns W-Y.

SPexper_mixed.m

Uses the GAEZ grid-level data files and runs Experiment 2 on "Spatial Potential" for all crops under the rainfed & irrigated water supply, and the mixed input scenario. The program produces aggregate counterfactual spatial-potential output and land for each country in the sample, except Brazil, Canada, China, India, Kazakhstan, Russia, United States. For the spatial-potential results for these countries see Section 3 of this README file on "Python-Gurobi Files."

Matlab toolbox required: Optimization Toolbox.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns AA-AB.

TPexper_mixed.m

Uses the GAEZ grid-level data files and runs Experiment 3 on "Total Potential" for all crops under the rainfed & irrigated water supply, and the mixed input scenario. The program produces an aggregate total-potential output and aggregate actual land for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns AD-AE.

PPexper_low_EQweights.m

Uses the GAEZ grid-level data files and runs Experiment 1 on "Production Potential" for all crops under rainfed low-input scenario, but using an equal weighting of crops. The program produces an aggregate production potential output and aggregate land for each country in the sample.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns AG-AH.

PPexper_low_HCweights.m

Uses the GAEZ grid-level data on potential yields and runs Experiment 1 on "Production Potential" for all crops under rainfed low-input scenario, but using the Harvest Choice (2012) land shares for the weighting of crops. The program produces an aggregate production potential output and aggregate land for each country in the sample. The grid-level data on land by crop from Harvest Choice are of the form: `lg_x_hc.dat.txt`, where "x" is the crop, i.e., wheat, rice, maize etc. These are contained in the folder `/Grid_Data_and_Matlab_Files/`.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," columns AJ-AK.

dispersion_all.m

Uses the GAEZ grid-level data files for all crops under rainfed low-input scenario, and computes the standard deviation of log-potential yield within each country.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," column AM.

dispersion_maize.m

Uses the GAEZ grid-level data files under rainfed low-input scenario, and computes the standard deviation of log-potential yield for maize within each country.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "Matlab Output," column AO.

income.m

Calibrates and simulates the model of Section 5 to produce the results reported in Section 5 of the paper.

4. Python-Gurobi Files

For 7 out of the 162 countries, Matlab cannot handle the linear optimization problem under the Spatial-Potential counterfactual (Experiment 2), due to the large number of cells in these countries: Brazil, Canada, China, India, Kazakhstan, Russia, United States. The linear programming problems for these countries have been solved with Python, using the Gurobi 9.0.3 interface (licensed software).

The code that can be used to run the spatial-potential linear programming problem in each case is contained in: `Python_Gurobi_SPexper.txt`. This can be run in each case by opening a new window in Anaconda3/Jupyter and copying-pasting the code, with the appropriate three letter country code.

The code uses as inputs csv files for each country on (examples below for low inputs in the case of the USA):

- Amount of land allocated to each crop: `Lcrops_low_USA.csv`
- Amount of land in each cell: `Lg_low_USA.csv`
- The potential yields of all crops in all cells: `vzPOmat_low_USA.csv`

The three-letter country codes used to identify each country in the files are in the parentheses: Brazil (BRA), Canada (CAN), China (CHN), India (IND), Kazakhstan (KAZ), Russia (RUS), United States (USA).

Files under the rainfed – low input scenario have the prefix “low” in their titles. Files under the rainfed & irrigated – mixed input scenario have the prefix “mixed” in their titles.

With seven countries, three data files per country under each of the two input scenarios, there are a total of 42 csv files. These are contained in the folder `/Python_Gurobi_Files/`.

Output of code for each country and each input scenario (columns C-D for low, columns F-G for mixed) can be found in file “/Output/Program Output.xlsx,” worksheet “Gurobi Output,” lines 3-11.

5. STATA Files

All STATA programs are run on `StataMP 15` (64-bit). To run a .do file change the working directory to `…/Replication/STATA_Files` and run all programs from this folder.

Output files for Tables and Figures are in the `/Output/` folder.

PP_low_country.dta

Dataset in STATA containing the aggregate country-level results from the Production Potential counterfactual, under the rainfed low-input scenario. Includes, real GDP per capita, actual land and output and counterfactual potential output for all crops, wheat, rice, and maize. Covers all countries in the sample. The Production Potential counterfactual output in each case comes from the Matlab programs: `PPexper_low.m, PPexper_low_wheat.m, PPexper_low_rice.m, PPexper_low_maize.m.`

Variable list:

un: UN country code
cc: three-letter country abbreviation
rgdppc: Real GDP per capita (Penn World Table)
y: Actual output – all crops
l: Actual land – all crops
ypo: Potential output – all crops
y_wheat: Actual output - wheat
l_wheat: Actual land - wheat
ypo_wheat: Potential output -wheat
y_rice: Actual output - rice
l_rice: Actual land - rice
ypo_rice: Potential output - rice
y_maize: Actual output - maize
l_maize: Actual land - maize
ypo_maize: Potential output – maize

PP_low_results.do

STATA do file that uses STATA data file `PP_low_country.dta` to produce figures and summary statistics from the paper, under the rainfed low-input scenario. The list of results produced are:

- Figure 3: lines 43-59
- Figure 4: lines 62-73
- Table 2, Panel A: uncomment lines 79-85
- Table 2, Panel B: uncomment lines 88-94
- Table 2, Panel C: uncomment lines 97-103
- Table 2, Panel D: uncomment lines 106-112
- Table 4: Columns 1 & 2: uncomment lines 115-122

Running the program will produce the two figures. To obtain the summary statistics for each case the corresponding lines need to be uncommented.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "STATA Output," lines 4-24.

SPandTP_low_country.dta

Dataset in STATA containing the aggregate country-level results from the Production-Potential, Spatial-Potential and Total-Potential counterfactual, under the rainfed low-input scenario. Includes, real GDP per capita, actual land and output and counterfactual potential output and land in each case. Covers all countries in the sample. The potential counterfactual output in each case comes from the Matlab programs: `PPexper_low.m, SPexper_low.m, TPexper_low.m.`

Variable list:

un: UN country code
cc: three-letter country abbreviation
rgdppc: Real GDP per capita (Penn World Table)
y: Actual output
l: Actual land
ypo: Production potential output
Yrl_SE: Spatial potential output
Lrl_SE: Spatial potential land
Yrl_TE: Total potential output
Lrl_TE: Total potential land

SPandTP_low_results.do

STATA do file that uses STATA data file `SPandTP_low_country.dta` to produce summary statistics from the paper for Spatial Potential and Total Potential, under the rainfed low-input scenario. The list of results produced are:

- Table 3: uncomment lines 29-37
- Table 4: uncomment lines 41-50

To obtain the summary statistics for each case the corresponding lines need to be uncommented.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "STATA Output," lines 27-35.

All_mixed_country.dta

Dataset in STATA containing the aggregate country-level results from the Production-Potential, Spatial-Potential and Total-Potential counterfactual, under the rainfed & irrigated mixed-input scenario. Includes, real GDP per capita, actual land and output and counterfactual potential output and land in each case. Covers all countries in the sample. The potential counterfactual output in each case comes from the Matlab programs: `PPexper_mixed.m, SPexper_mixed.m, TPexper_mixed.m.`

Variable list:

un: UN country code
cc: three-letter country abbreviation

rgdppc: Real GDP per capita (Penn World Table)
y: Actual output
l: Actual land
ypo: Production potential output (mixed inputs)
Ym_SE: Spatial potential output (mixed inputs)
Lm_SE: Spatial potential land
Ym_TE: Total potential output (mixed inputs)
Lm_TE: Total potential land

All_mixed_results.do

STATA do file that uses STATA data file `All_mixed_country.dta` to produce summary statistics from the paper for Production Potential, Spatial Potential and Total Potential, under the rainfed & irrigated mixed-input scenario. The code reproduces Figure 5, on actual and production potential yield under the mixed input scenario against real GDP per capita (lines 33-44), and the results of Table 5 for all experiments under the mixed input scenario (lines 49-57).

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "STATA Output," lines 38-42.

caloric_country.dta

Dataset in STATA containing the aggregate country-level results from the Production-Potential and Total-Potential counterfactual, under the rainfed low-input scenario, using caloric prices. Includes, real GDP per capita, actual land and output and counterfactual potential output and land in each case. Covers all countries in the sample. The potential counterfactual output in each case comes from the Matlab programs: `PPexper_low.m, TPexper_low.m,` adjusted for caloric prices.

Variable list:

un: UN country code
cc: three-letter country abbreviation
rgdppc: Real GDP per capita (Penn World Table)
y_cp: Actual output using caloric prices
l: Actual land
ypo_pp_cp: Production potential output (low inputs) using caloric prices
ypo_tp_cp: Total potential output (low inputs) using caloric prices

caloric_results.do

STATA do file that uses STATA data file `caloric_country.dta` to produce summary statistics from the paper for Production Potential and Total Potential, under the rainfed low-input scenario, but using caloric prices. The code reproduces the results of Table 6 in the paper.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "STATA Output," lines 45-49.

alt_weights.dta

Dataset in STATA containing the aggregate country-level results from the Production-Potential and, under the rainfed low-input scenario, for two alternative weighting schemes: equal weighting of crops, and weighting of crops with Harvest Choice (2012) land shares. Includes, real GDP per capita, actual land and output and counterfactual potential output and land in each case. Covers all countries in the sample. The potential counterfactual output in each case comes from the Matlab programs: `PPexper_low_EQweights.m, PPexper_low_HCweights.m`.

Variable list:

un: UN country code
rgdppc: Real GDP per capita (Penn World Table)
y: Actual output
l: Actual land
ypo_eq: Production-potential output under equal weighting of crops
ypo_hc: Production-potential output under Harvest Choice land shares weighting of crops

weights_robustness.do

STATA do file that uses STATA data file `alt_weights.dta` to produce robustness for the Production Potential counterfactual, under the rainfed low-input scenario, using alternative weighting schemes for crops. The code reproduces Figure 6, with equal weighting of crops, and Figure D.3 in Appendix D, with Harvest Choice (2012) land shares weighting of crops.

std_data.dta

Dataset in STATA containing for each country the standard deviation of log-potential yield, for all crops and for maize alone. Country entries are produced from the Matlab programs `dispersion_all.m` and `dispersion_maize.m`.

Variable list:

un: UN country code
gdppc: Real GDP per capita (Penn World Table)
std_all: within country standard deviation of log-potential yield across all crops
std_maize: within country standard deviation of log-potential yield for maize

STD_figures.do

STATA do file that uses STATA data file `std_data.dta` to produce Panels A and B of Figure C.1 in Appendix C on the within country dispersion of rainfed, low input potential yields across all crops (Panel A) and maize (Panel B).

Subnational_Monfreda.dta

STATA data file containing data from Monfreda et al. (2018) on the percentage of data from sub-national sources.

subnational.do

Uses the STATA data file `Subnational_Monfreda.dta` to create Figure D.2 in Appendix D.

attributes.dta

STATA data file containing mean land quality attributes from GAEZ (2000) by country: fertility, depth, slope, altitude, temperature, precipitation.

Variable list:

un: UN country code
cid: three-letter country code
gdppc: Real GDP per capita (Penn World Table)
fert: soil fertility
depth: soil depth
slopein: terrain slope
altit: altitude
temp: temperature
precip: precipitation

attributes_table.do

STATA do file calculating the entries of Table 1 using the data file `attributes.dta`. Computes the mean for the 10% highest and lowest income countries, and the mean for the top and bottom 10% of the distribution of each attribute.

Output of code can be found in file "/Output/Program Output.xlsx," worksheet "STATA Output," lines 52-60.

References

FAOSTAT (2000). Production statistics. Food and Agricultural Organization (FAO), (Value of Agricultural Production, Area Harvested)**.** http://www.fao.org/faostat/en/#data (accessed January 20, 2021).

GAEZ (2000). Global Agro-Ecological Zones (GAEZ), version 3.0. Food and Agricultural Organization (FAO) and International Institute for Applied Systems Analysis (IIASA). https://www.gaez.iiasa.ac.at/ (accessed March 11, 2018)

Alan Heston, Robert Summers and Bettina Aten, Penn World Table Version 6.3, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania, August 2009. https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt-6.3?lang=en

Harvest Choice (2012). Global Spatially-Disaggregated Crop Production Statistics Data for 2005 version 3.2. International Food Policy Research Institute (IFPRI) and International Institute for Applied Systems Analysis (IIASA), (Harvard Dataverse, V9, 10.7910/DVN/DHXBJX). https://doi.org/10.7910/DVN/DHXBJX

Monfreda, C., Ramankutty, N., and Foley, J. A. (2008). Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles*, 22(1).

TM (2008). World borders dataset. Thematic Mapping. http://thematicmapping.org/downloads/world_borders.php (accessed July 25, 2014).

UNPS (2009). Uganda bureau of statistics, national panel survey (UNPS) 2005-2009. World Bank, (Ref. UGA 2005-2009 UNPS v01 M.). https://microdata.worldbank.org/index.php/catalog/1001/get-microdata (accessed June 15, 2020).

USDA (2015). USDA National Nutrient Database for Standard Reference, Release 28. US Department of Agriculture, Agricultural Research Service, Nutrient Data Laboratory. https://www.ars.usda.gov/Services/docs.htm?docid=8964 (accessed April 5, 2018).