The Effect of Climate Change on Population Growth: a Machine Learning Approach

ECON1680 Project 1

Author:

Lisa Ruixuan Li

Supervisor:

Prof. Amy Handlan

Giulia Gitti

March 2023

Contents

1 Introduction

1.1 Research Goal

For years, geologists have been publicizing the detrimental effects of climate change: with increased carbon emissions, global temperature will possibly increase, which could have important implications for human activity. Thermal fluctuations can lead to a variety of changes in different geological and climatic parameters: sea elevation, crop yield, precipitation, and soil conditions,..., just to name a few. These geological and climatic characteristics deeply influence land quality, which can contribute to human migration and population growth. Unfavorable climate change posts risks to human productivity and settlement: it worsens a variety of health outcomes (like allergies, asthma, and other moisture and thermal-sensitive diseases), influence agricultural productivity of land (such as decreasing soil quality and crop yield), and even diminishes the area of land we have access to due to increased sea elevation. Given these implications, it is important to study climate change and its effects on land quality and population growth as an important economics question.

Previous economics papers have explored similar topics and questions in two ways: 1) Cross-section approach: Regressing aggregate GDP in grid cells covering the whole world to compare economic outcomes in locations with various climates 2) Panel Data: Regressing changes in economic output on changes in climate conditions to compare climate changes over time. In this paper, I use the cross-section approach given grid-cell level data on climate, soil, and population covering more than 160 countries averaged over the 1981-2010 period. Specifically, I incorporate two major machine-learning methods, PCA and LASSO, to study the comparative weights of 44 climatic variables on population density. After reducing the dimensionality and ranking their comparative weights, I further employ a Poisson regression model on the variables of reduced dimension developed by Henderson, Storeygard, and Weil (2022),[1] with account of country fixed effects.

1.2 Methodology contribution

With machine-learning methods, this research contributes to the current development and agricultural economics literature in the following ways:

1. Variable dimensionality reduction. Traditionally, if we were to use OLS to regress population density on a set of 50 variables describing land quality, we would encounter an "overfitting" problem, inevitably making regression coefficients significant and resulting in high goodness-

1. Adam Storeygard J. Vernon Henderson and David N. Weil, "Land Quality" (Unpublished, December 2022).

of-fit (R^2). Machine learning methods such as PCA and LASSO filter and rank climate and geographical variables that are most predictive of population growth. These methods allow us to further cluster variables to sub-groups and "condense variables" within each sub-group to one dimension (for instance, condensing 50 climate variables to 5 variables each representing an aspect of climate change).

2. Cross-assessment of related papers' model specification. After dimensionality reduction, the "condensed variables" can be further put into the original model specification with diminished concerns for overfitting. Notable papers like Nordhaus (2006) and Henderson et al. (2018) have used OLS, log-linear model, and Poisson for estimating the effect of climate change on GDP. Machine learning methods allow us to assess existing results' overfitting problems.

The downside of machine learning methods is that it doesn't provide a marginal effect interpretation on the outcome variable, and it compromises in-sample accuracy for out-sample prediction. However, when projecting future population density given a future climatic scenario, we can be more confident of the comparative weights and importance of each type of climate variable contributing to population growth.

2 Dataset

2.1 Sources and Basic Information

The area of 164 countries is divided into roughly the same area, labeling each area as a tiny grid cell, which serves as one unit of observation. Each entry contains information on 44 climate and geographical variables as well as population density measures. All variables are continuous. The grid-cell level measurement is averaged over period 1981-2010. There are $237k$ observations, and a third of them have a population density of 0.

The dataset comes from two sources. Population density, the outcome variable, is from European Union's Global Human Settlements population layer (GHS-POP), which provides an estimate of population within each 30-arcsecond (approximately 1 square km) grid cell. Climate variables, a total of 44 variables, are from U.N. Food and Agricultural Organization's Global Agro-Ecological Zones version 4 dataset (i.e. FAO's GAEZv4). These variables can be classified to 5 categories:

1. **Baseline variables**, extracted from Henderson et al., that features basic geo-climate characteristics

2. **Length of Growth Period (LGP) variables**, with specifications for dry and wet weather conditions for soil growth
3. **Moisture variables**, with information on precipitation and water gain
4. **Crop-Suitability variables**, with indexes on the grid cell's suitability for the 11 largest calories production crops essential to human survival
5. **Thermal variables**, with annual average temperatures and number of days above or below extreme temperature

Table 1 summarizes all variable names within each category.

2.2 Summary statistics

To give a better sense of what the cross-sectional data looks like, I summarized the mean, standard deviation, maximum, and minimum of the most important parameters in the climate change and population density literature (Storeygard, Henderson, and Weil 2022), collapsed by continents. They are annual temperature, annual precipitation, days of growing period, and population density.

Table 2 lists the key summary statistics for each continent.

Table 1: Variable names and their categories.

Baseline	**Length of Growth Period**	**Moisture**	**Crop-suitability**	**Thermal**
Growing Period (days)	Adjusted LGP	Annual PPET ratio	Banana	Days max-temp$> 40^C$
Land Suitability (index)	Length of Longest LGP	Days with longest rain-period	Cassava	Days min-temp$< 0^C$
Ruggedness (index)	Longest Consecutive Dry Days	Rainy days	maize	Days min-temp$< 10^C$
Annual average Precipitation (mm)	Number of Dry Days	Fournier Index	Dryland Rice	Days min-temp $< 15^C$
Annual average Temperature (^{C})	Total Growth Period	Annual Precipitation	Wetland rice	Days avg-temp$> 10^C$
Distance from Sea (km)		Sum of precipitation$> 30mm$	Soybean	Days avg-temp$> 5^C$
Elevation (m)			Sweet Potato	Mean ^{C}
Malaria (index)			Sorghum	Snow-adjusted mean
			Wheat	Coolest-month mean
			White Potato	$> 5^C$ temperature sum
			Yam	$> 10^C$ temperature sum
				Days with frost
				Days with snow
				Days max-temp$> 30^C$

continent	Annual temperature (^{C})				Annual precipitation (mm)				Growing period (days)				Population density				
continent	mean	std	min	max	mean	std	min	max	mean	std	min	max	mean	std	min	max	obs
Africa	24.3	3.6	6.9	31.1	1.9	1.8	0.0	11.1	121.1	107.8	0.0	365.2	48.0	351.5	0.0	40715.4	41461
Americas	8.8	13.5	-21.0	31.8	2.5	2.2	0.0	23.1	167.7	104.9	0.0	366.0	24.8	216.7	0.0	21634.9	72810
Asia	14.0	10.2	-11.9	30.5	2.0	2.4	0.0	30.3	134.7	111.9	0.0	366.0	150.3	553.7	0.0	26727.5	50213
Europe	-2.4	8.4	-22.7	20.0	1.3	0.7	0.3	9.0	129.6	50.3	0.0	366.0	31.9	670.3	0.0	158097.7	59784
Oceania	21.4	4.7	2.6	29.5	1.7	1.9	0.1	20.9	107.0	98.4	17.3	366.0	6.6	77.0	0.0	3974.4	12755

Table 2: Summary Statistics

3 Methodology

3.1 Modeling Climate Change, Land Quality, and Population Growth

Following Henderson, Storeygard, and Weil (2022), and previous work done by Nordhaus (2006) and Henderson et al. (2018), the relationship of climate, land quality, and population density can be modeled by a Poisson regression.

Land quality, $Q_{i,c}$, is a function of a vector of climatic and geographical characteristics, where c is countries, i is grid cells:

$$Q_{i,c} = exp(X_{i,c}\beta) \tag{1}$$

Population density for a specific grid area is population counts divided by land area, $L_{i,c}/Z_{i,c}$. It can be mapped by land-quality multiplied by country fixed-effects:

$$L_{i,c}/Z_{i,c} = exp(X_{i,c}\beta)C_c \tag{2}$$

The regression coefficients, vector β, could be interpreted as "a change in X by one unit is associated with 100β percent change in population density".

Future land quality (for instance, in 2100) is a fitted value of the previous regression:

$$Q_{i,c,2100} = exp(X_{i,c,2100}\hat{\beta}) \times \frac{\sum Z_{i,c}}{exp(X_{i,c,2010}\hat{\beta})\sum Z_{i,c}}$$

In comparison to Henderson, Storeygard, and Weil (2022), which runs a Poisson regression on 53 variables with 66 indicators, I run the regression on a more narrowly selected range of 44 continuous variables, serving as a base comparison for the machine-learning regressions.

3.2 Machine Learning Methods

I employ two machine learning models in this research to study the relationship between climate change and population growth. They are Principle Component Analysis and LASSO.

3.2.1 PCA

PCA is used as a dimensionality reduction algorithm. In Henderson, Storeygard, and Weil (2022), the R^2 for equation (2) is 0.344. I first reduce the dimensionality of 44 variables to 5 dimensions and extract the most explained principal component of each dimension. A "scree plot" is produced, which maps how well each of the 5 PCAs explains their variable category. Following equation (2), I further run a regression with country fixed effects on the 5 condensed variables (most explained PCAs). The R^2 should be expected to drop for loss of variable explanation.

Of the 5 reduced-dimension variables, the higher the explained variance of the first principal component, the more correlated within their own variable category. The variables with the highest absolute β are more predictive of population growth. We should reasonably expect that the "baseline variable" dimension has lowest explained variance for its first principal component, and is the most correlated with the other 4 dimensions. This is because "baseline variables" contain the widest variety of parameters covering thermal, landscape, and moisture, and this baseline dimension simultaneously correlate with other dimensions.

3.2.2 LASSO

LASSO is used as another dimensionality reduction algorithm. The purpose of this operation is to validate the comparative importance of the additional variables of 36 characteristics (grouped to 4 dimensions) from GAEZv4 to the baseline variables studied by Henderson et al. (2018). This helps us confirm if the Henderson, Storeygard, and Weil study is econometrically effective by including 36 characteristics in addition to the 27 baseline specifications. If the LASSO algorithm decides to drop the majority of our 36 variables, or reduce their coefficients to a comparatively small absolute value, then we can reasonably conclude that the significant coefficients from Henderson, Storeygard, and Weil (2022) might be due to over-fitting. We can also calculate the percentage of independent variables dropped for each of the 4 variable categories. If the algorithm decides to drop most of the "thermal" variables, for instance, then we can infer that thermal variables might not be predictive of population growth.

To minimize in-sample mean-squared error and maximize out-of-sample prediction, I optimize the choice of penalty term α. The LASSO regression is run using the optimal penalty α.

4 Results and Expected Results

4.1 PCA results

From the scree plot, we can see that the thermal variable has the highest explained variance for the first principal component, followed by growth, moisture, and crop-suitability. The baseline variable has the lowest explained variance ratio for the first principal component. This validates my expectations – given that the baseline variable category covers the largest variety of parameters spanning the other 4 dimensions as well as indexes related to sea elevation and landscape, condensing it to one dimension demonstrates substantial loss. The first two principle components' explained variance ratio are summarized in Table 2.

	Baseline variables	Growth	Moisture	Crop Suitability	Thermal
PCA1	0.328874	0.690196	0.673110	0.590177	0.821179
PCA2	0.188281	0.243845	0.233394	0.251392	0.103807

Table 3: Explained Variance Ratio of First 2 Principal Components

I further run Poisson regression of the 5 first principal components on population density with country fixed effects. The results show significant coefficients for all of the 5 principal components. Baseline variables have the highest absolute coefficient value (0.48), which suggests this category is the most predictive of population growth. Of the remaining four categories, the variables ranked from most predictive to least predictive of population growth are thermal, crop-suitability, moisture, and soil growth. The $R2$ for this regression is 0.339, which is only slightly lower than the Poisson specification regression run on 53 variables in Henderson, Storeygard, and Weil (2022).

	coef	std err	z	P> $\|z\|$	0.025 bound	0.975 bound
const	3.3333	0.001	3004.711	0.0	3.331	3.335
Base Variables	0.4790	0.001	779.102	0.0	0.478	0.480
Growth	-0.0197	0.000	-48.007	0.0	-0.020	-0.019
Moisture	-0.0746	0.000	-417.164	0.0	-0.075	-0.074
Crop Suitability	0.1003	0.000	568.554	0.0	0.100	0.101
Thermal	-0.1389	0.000	-417.520	0.0	-0.140	-0.138

Table 4: Poisson Regression of First Principal Component for 5 Variable Categories

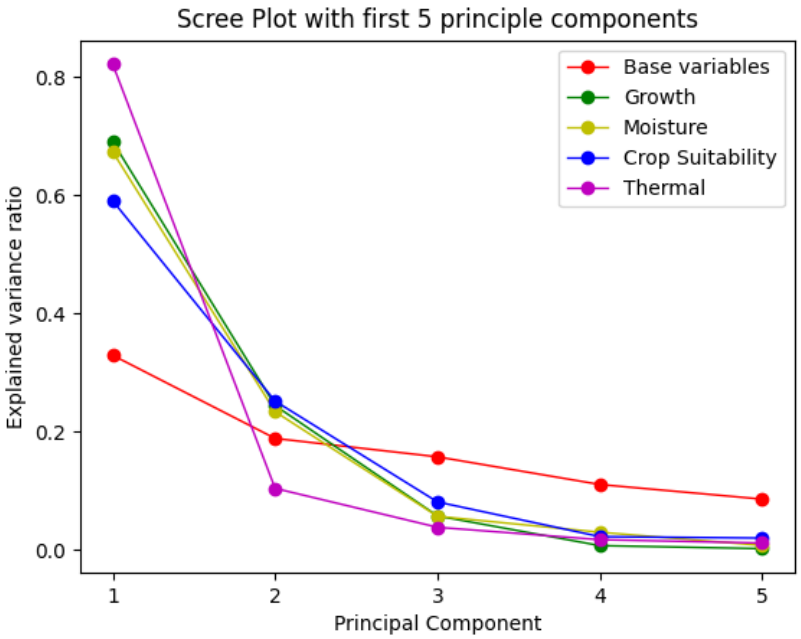

Figure 1: Explained variance ratio

4.2 LASSO results and comparison with Poisson

Compared to R^2 of 0.344 in their paper run on 53 variables, I rerun the model with 44 continuous variables on population density, all variables normalized. The R^2 for this regression is around 0.46, which is a sizable improvement. The coefficients have a marginal interpretation with respect to change in the standard deviation of the independent variables – for example, if mean annual precipitation increases by one standard deviation, there would be a 48.9% increase in population density.

Based on the Henderson, Storeygard, and Weil (2022) paper, I expect that the Lasso regressions should rarely drop any of the 8 baseline covariates. I first find the optimal penalty term α that minimizes mean-squared-error (Figure 2 plot), which is around 0.0163. Then I rank the absolute values of the LASSO coefficients, comparing them with the Poisson coefficients (Table 4).

Overall, there is a vast difference between the results of LASSO and Poisson. Of the 10 most important variables influencing population density suggested by LASSO, there are only 3 ranked as the most influencial on population density with respect to a one standard deviation change by the Poisson regression. Consistent with the PCA-Poisson regression results, thermal variables depicting

Lasso Rank	Variable	Coeff LASSO_X1	Coeff Poisson	Poisson Rank	Category
1	Mean annual temperature	332.561802	0.488821	4	Thermal
2	Coolest Month temperature	-225.177376	-0.116378	27	Thermal
3	Temperature sum	-160.905433	0.077849	34	Base
4	Annual precipitation	-146.996465	0.144227	21	Moisture
5	Days temp-min< 0	-123.436069	-0.124564	24	Thermal
6	Days temp-avg> 10	-113.387634	-0.034758	38	Thermal
7	Annual temp sum> 10	110.779433	0.108862	28	Thermal
8	Number of dry days	-105.267292	-0.330692	9	Growth
9	Fournier index	90.524663	-0.089686	32	Moisture
10	Days with frost	-84.192435	-0.492224	3	Thermal
11	Yam	-74.399511	-0.008358	43	Crop-suit
12	Days temp-avg> 5	-71.118612	-0.021382	41	Thermal
13	Dryland rice	-60.239514	0.189548	18	Crop-suit
14	Wetland_rice	58.039953	-0.290495	11	Crop-suit
15	Growing period	55.086386	0.255611	13	Base
16	Snow-adjusted mean	51.813540	-0.321193	10	Thermal
17	Cassava	43.982052	0.246345	15	Crop-suit
18	Sum of precipitation> 30	36.362937	0.121412	26	Moisture
19	Total growth period	-35.912602	-0.231598	16	Growth
20	Distance from sea	-33.055501	-0.496284	1	Base
21	Land suitability	30.948178	0.224148	17	Base
22	Sum of Precipitation> 30	-29.825713	0.104292	29	Moisture
23	Rainy days	-28.773230	-0.250066	14	Moisture
24	sweet_potato	24.729594	0.011879	42	Crop-suit
25	Annual PPET ratio	19.477271	0.056118	36	Moisture
26	Malaria	-18.970481	-0.150388	20	Base
27	Days temp-max> 35	-18.226209	-0.098464	31	Thermal
28	Elevation	14.370241	-0.285369	12	Base
29	Sorghum	14.032527	0.128831	23	Crop-suit
30	Maize	-12.313028	0.053583	37	Crop-suit
31	Soybean	10.258618	-0.384362	6	Crop-suit
32	Banana	10.180386	0.060616	35	Crop-suit
33	Length of Longest LGP	-9.668217	-0.123366	25	Growth
34	Longest consecutive dry days	7.691086	0.082688	33	Growth
35	Wheat	7.034104	0.369972	7	Crop-suit
36	Annual average precipitation	6.203225	-0.027679	40	Base
37	White potato	-6.151434	-0.387294	5	crop-suit
38	Days temp-max> 40	-2.137496	-0.001216	44	thermal
39	Adjusted LGP	-1.972812	0.361055	8	Growth
40	Days temp-min< 15	1.388781	0.164969	19	Thermal
41	Days temp-min< 10	1.239471	-0.133768	22	Thermal
42	Ruggedness	-1.236412	-0.034747	39	Base
43	$> 5^C$ temperature sum	-0.000000	0.098830	30	Thermal
44	Snow-adjusted mean	0.000000	-0.495338	2	Thermal

Table 5: LASSO vs. Poisson regression results

temperature mean and extreme weather have the highest absolute LASSO coefficients, followed by crop-suitability of common staple foods. Many of the highest-ranked variables using LASSO are ranked relatively low in the Poisson regression, displaying no specific pattern in the relative importance of climate variables belonging to a single category.

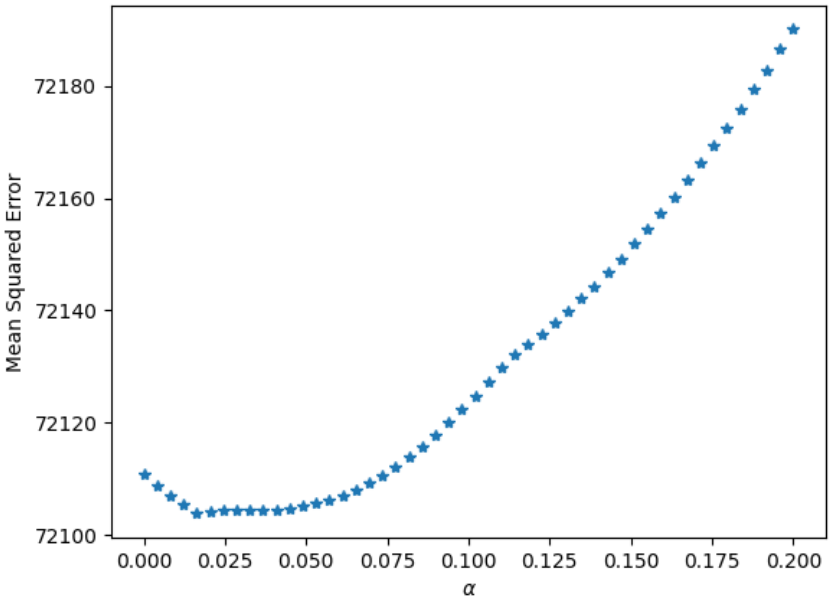

Figure 2: MSE against penalty term

5 Conclusion

This accomplishes three things: 1) Qualitatively define the comparative importance of different types of climatic and geographical characteristics on population growth, 2) Validate the methodology and results of Henderson et al.(2018) and Henderson, Storeyguard, and Weil (2022) by reduced-dimension algorithms, and 3) Offer predictive results for future population density based on future climatic scenarios. Our PCA and LASSO results, as compared to Poisson specification, gives a more specific ranking on the relative importance of climate variables on population density: thermal variables are the most influential, followed by crop-suitability and moisture. Although the LASSO algorithm does not drop the baseline variables studied by Henderson et al. (2018), they are ranked as relatively low importance in LASSO, possibly because the effect gets absorbed by variables in the other 4 dimensions.

6 References

Henderson, J. Vernon, Tim Squires, Adam Storeygard and David N. Weil. 2018. "The Global Distribution of Economic Activity: Nature, History, and the Role of Trade," The Quarterly Journal of Economics 133(1): 357–406.

Nordhaus, William. 2006. "Geography and macroeconomics: New data and new findings," Proceedings of the National Academy of Sciences, 103(10): 3510–3517.

Weil, David Nathan, Adam Storeygard, and J. Vernon Henderson. 2022. "Land Quality." Working Paper, August. https://bpb-us-w2.wpmucdn.com/sites.brown.edu/dist/1/24/files/2022/08/Land-Quality-August-2022.pdf.