STAT 230A Replication Paper Summary

Andrej Leban, andrej_leban@berkeley.edu
Isaac Schmidt, ischmidt20@berkeley.edu

`TODO: margins, font size, group name & name in upper right`

1 Paper Summary & Summary Statistics Table

- *summarize the paper's research question and its answer*

1.1 Paper Summary

The paper by Michalopoulos aims to explain ethnolinguistic diversity within and across countries by assuming that a proxy quantity - the number of languages per square kilometer - is determined by a selection of various economic, historical, and geographic variables. It determines that *variation in regional land quality* and *variation in elevation* are the most significant determinants of linguistic diversity. The hypothesis underpinning the examination is that differences in land endowments induce different levels of human capital across locations, which in turn, gives rise to localized ethnicities which are characterized by separate languages. The results of the empirical study presented are found to be consistent with this hypothesis.

1.2 Data statistics and summary table

- *describe the datasets used towards answering the question*
- *clean the dataset to replicate and interpret a summary statistics table that presents distributional characteristics (mean, median, IQR, etc.) of key variables and covariates used in the empirical analysis. The summary table need not replicate exactly as the table produced in the paper.*

`TODO: * explain squares`

The paper lacks a true summary table, and shows a couple EDA figures instead. We replicate two of those figures, and then display our own summary table of the features used in the paper's first regression.

Figure 1 shows the distribution of land suitability for agriculture across the world, at a resolution of .5-by.5 decimal degrees.

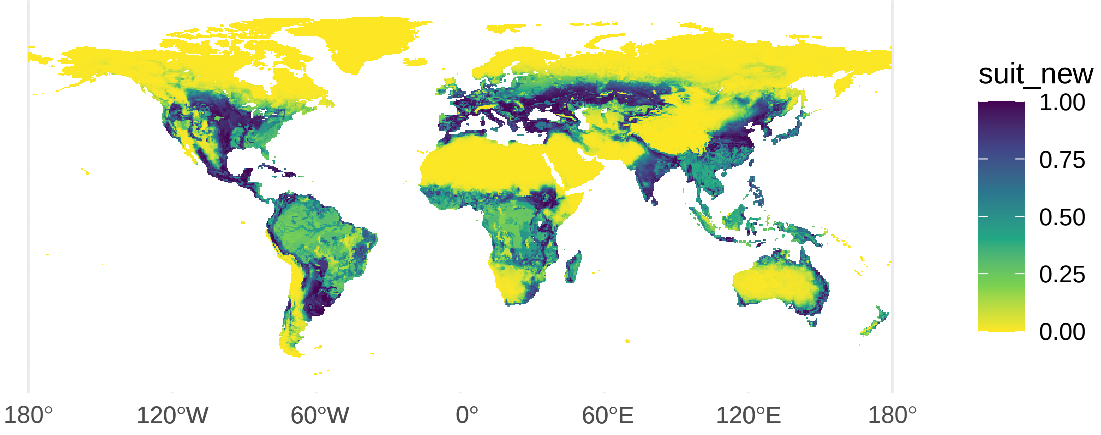

Figure 1: Land Quality Across Countries

Figure 2 shows the distribution of land quality within two countries selected in the paper—Greece and Nepal. This was done by intersecting the cells shown in the previous figure with country boundaries, and then plotting a KDE with the Epanechnikov kernel.

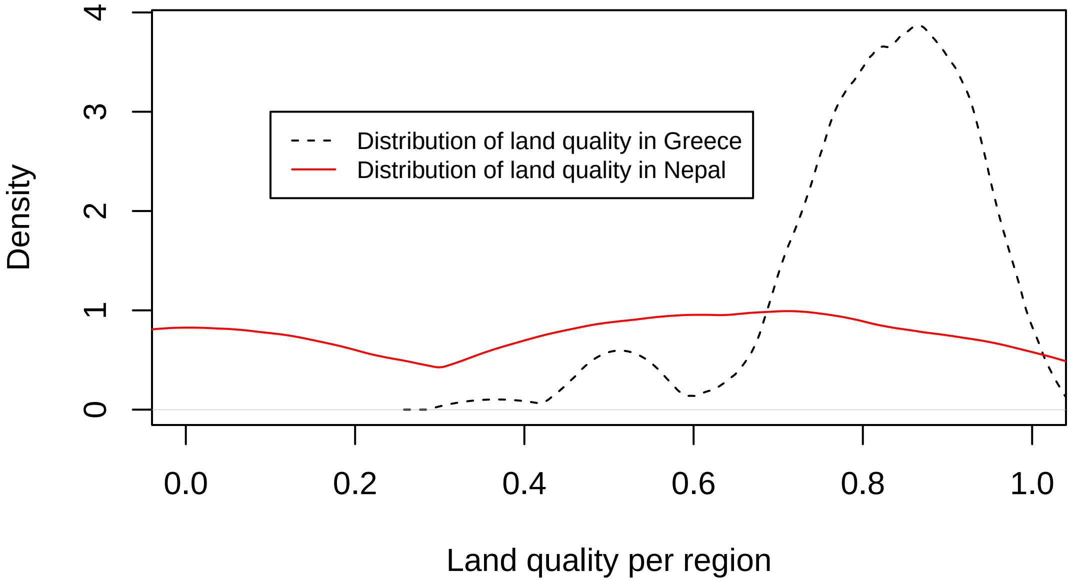

Figure 2: Kernel Density of Land Quality in Greece and Nepal

Below is our summary table of important variables. Fortunately, the data provided by the author was already processed and cleaned to the extent used in the paper. All we did was rename columns to more descriptive names. For the first regression, each unit of observation is a country. The dependent variable is `numLang`, which is the number of languages whose "traditional homeland" intersects with the country's boundary. Other variables, such as `avgSuitable` and `sdSuitable`, were aggregated from the land suitability dataset described above. Note that measures of both center and spread from the aggregation are included as features. Also included are the log of the country's 1995 population, human migration distance from Africa, and distance from a large body of water.

While some other variables in the provided dataset have missing values for some countries, note that all variables included in the first regression are known for all countries.

	numLang	sdElev	sdSuitable	avgElev	avgSuitable	absLat
min	1.00	0.01	0.00	0.03	0.00	0.64
median	10.00	0.25	0.18	0.42	0.44	24.18
max	462.00	1.95	0.41	2.52	0.96	67.79
mean	35.69	0.36	0.18	0.57	0.44	27.14
sd	73.41	0.36	0.10	0.49	0.25	17.68
n	156.00	156.00	156.00	156.00	156.00	156.00

	avgPrecip	avgTemp	lnArea	seaDist	migrationDist	lnPop95
min	4.00	-6.37	-3.24	0.01	0.10	-10.22
median	77.11	20.93	0.61	0.18	5.79	-3.07
max	278.16	28.74	4.73	1.98	26.67	-0.25
mean	91.23	17.86	0.52	0.34	8.69	-3.27
sd	63.84	8.49	1.55	0.38	6.89	1.46
n	156.00	156.00	156.00	156.00	156.00	156.00