The Origins of Ethnolinguistic Diversity

Stelios Michalopoulos

Web Appendix

Table 1a: Summary Statistics for the Real Country Analysis

_stats	Ln # of Languages	Variation in Elevation	Variation in Land Quality	Mean Elevation	Mean Land Quality	Absolute Latitude	Mean Precip.	Mean Temp.	Area	Sea Distance	Migratory Distance from East Africa	Pop. Dens in 1995
mean	2.42	0.36	0.18	0.57	0.44	27.14	91.23	17.86	5.46	0.34	8.69	0.08
sd	1.52	0.36	0.10	0.49	0.25	17.68	63.84	8.49	13.40	0.38	6.89	0.10
max	6.14	1.95	0.41	2.52	0.96	67.79	278.16	28.74	113.79	1.98	26.67	0.78
min	0.00	0.01	0.00	0.03	0.00	0.64	4.00	-6.37	0.04	0.01	0.10	0.00

See web Appendix for variables' sources and definitions

Table 1b: The Correlation Matrix for the Real Country Analysis

	Ln # of Languages	Variation in Elevation	Variation in Land Quality	Mean Elevation	Mean Land Quality	Absolute Latitude	Mean Precip.	Mean Temp.	Area	Sea Distance	Migratory Distance from East Africa	Pop. Dens in 1995
Ln # of Languages	1.00											
Variation in Elevation	0.27	1.00										
Variation in Land Quality	0.23	0.45	1.00									
Mean Elevation	0.15	0.77	0.39	1.00								
Mean Land Quality	-0.09	-0.05	0.29	-0.07	1.00							
Absolute Latitude	-0.48	0.01	0.18	-0.03	0.03	1.00						
Mean Precip.	0.25	-0.05	-0.13	-0.16	0.30	-0.51	1.00					
Mean Temp.	0.40	-0.19	-0.27	-0.22	-0.08	-0.91	0.36	1.00				
Area	0.57	0.26	0.23	0.16	-0.25	-0.03	-0.28	0.01	1.00			
Sea Distance	0.25	0.33	0.30	0.52	-0.21	-0.04	-0.31	-0.11	0.42	1.00		
Migratory Distance from East Africa	-0.05	0.21	-0.05	-0.06	0.17	-0.15	0.53	0.09	-0.16	-0.18	1.00	
Pop. Dens in 1995	-0.04	0.06	0.34	0.02	0.53	0.12	0.05	-0.08	-0.14	-0.17	-0.20	1.00

See web Appendix for variables' sources and definitions

Map 1

Land Quality Across Virtual Countries with Language Coverage from the WLMS (2006) database

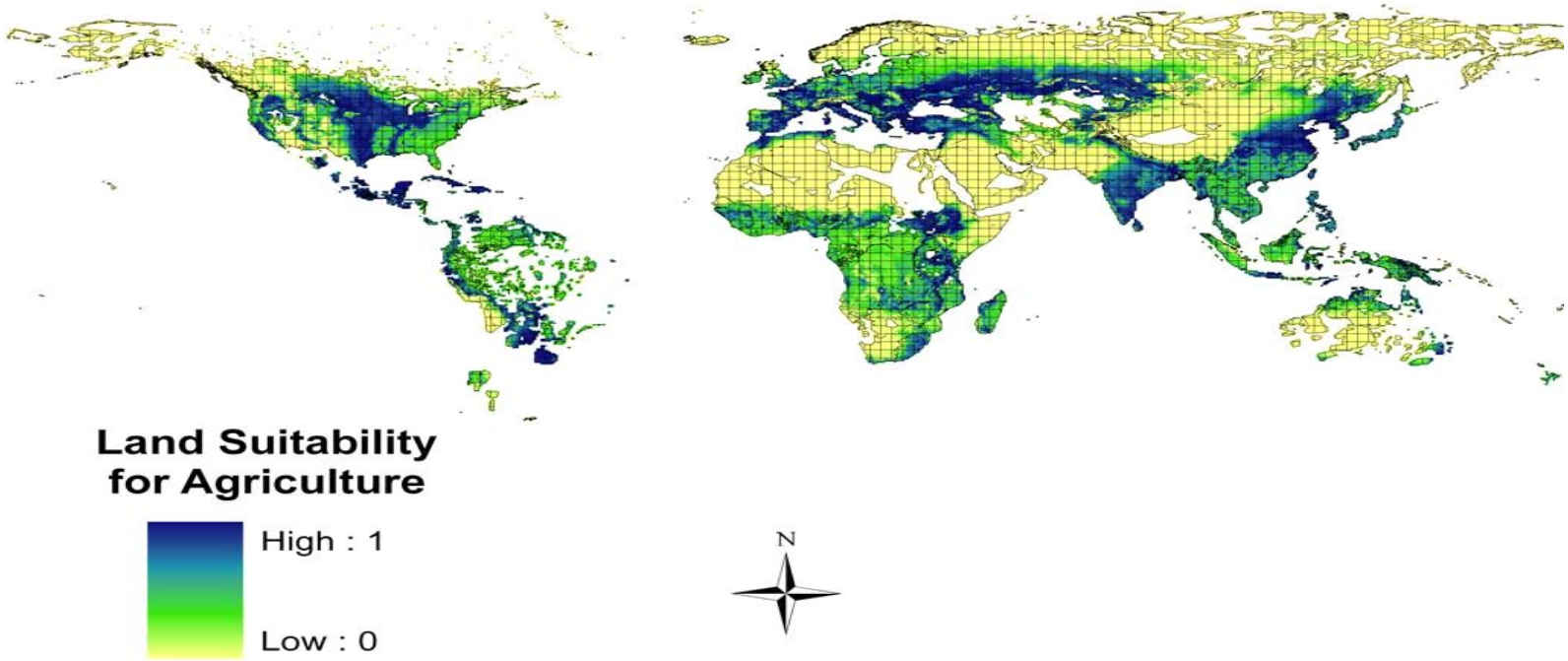

Table 2a: The Summary Statistics for the Virtual Country Analysis

_stats	Ln # of Languages	Variation in Elevation	Variation in Land Quality	Mean Elevation	Mean Land Quality	Absolute Latitude	Mean Precip.	Mean Temp.	Area	Sea Distance	Water Area	Migratory Distance from East Africa	Within Country	Pop. Dens in 1995
mean	1.21	0.19	0.11	0.35	0.70	34.35	0.07	12.76	47.92	0.63	0.87	9.41	0.60	0.06
sd	0.96	0.22	0.09	0.30	0.79	18.72	0.06	11.33	18.40	0.59	1.30	6.15	0.49	0.13
max	5.04	1.86	0.47	0.98	5.06	70.70	0.41	29.87	76.89	2.69	15.89	27.14	1.00	1.80
min	0.00	0.00	0.00	0.00	-0.02	0.52	0.00	-17.73	3.32	0.00	0.00	0.14	0.00	0.00

See web Appendix for variables' sources and definitions

Table 2b: The Correlation Matrix for the Virtual Country Analysis

	Ln # of Languages	Variation in Elevation	Variation in Land Quality	Mean Elevation	Mean Land Quality	Absolute Latitude	Mean Precip.	Mean Temp.	Area	Sea Distance	Water Area	Migratory Distance from East Africa	Within Country	Pop. Dens in 1995
Ln # of Languages	1.00													
Variation in Elevation	0.23	1.00												
Variation in Land Quality	0.27	0.32	1.00											
Mean Elevation	0.22	0.01	0.54	1.00										
Mean Land Quality	0.10	0.57	0.09	-0.18	1.00									
Absolute Latitude	-0.56	-0.13	-0.14	-0.17	-0.09	1.00								
Mean Precip.	0.43	0.04	0.20	0.34	-0.18	-0.48	1.00							
Mean Temp.	0.46	-0.02	0.10	0.20	-0.19	-0.90	0.37	1.00						
Area	0.30	0.08	0.13	0.12	0.15	-0.36	-0.05	0.36	1.00					
Sea Distance	-0.14	0.08	-0.04	-0.16	0.28	0.23	-0.35	-0.35	0.07	1.00				
Water Area	-0.04	-0.09	-0.11	-0.16	-0.01	0.04	-0.16	0.02	0.16	0.01	1.00			
Migratory Distance from East Africa	-0.19	0.12	0.11	0.18	0.04	0.02	0.34	-0.10	-0.28	-0.06	-0.01	1.00		
Within Country	-0.36	-0.10	-0.14	-0.05	-0.01	0.22	-0.07	-0.23	-0.15	0.09	-0.04	0.15	1.00	
Pop. Dens in 1995	0.34	0.14	0.42	0.59	-0.10	-0.24	0.31	0.33	0.23	-0.38	-0.04	-0.06	-0.12	1.00

See web Appendix for variables' sources and definitions

Table 3a: Summary Statistics for the Pairwise Analysis of Adjacent Regions

_stats	% of Common Languages	Difference in Land Quality	Difference in Elevation	Geodesic Distance	Migratory Distance from East Africa	Absolute Latitude	Mean Elevation	Mean Land Quality	Mean Precip.	Mean Temp.	Sea Distance	Water Area	Pop. Dens in 1995	Area	Within Country
mean	0.77	0.08	0.15	0.60	9.40	35.60	0.69	0.33	0.06	11.88	0.67	0.08	0.05	4.30	0.84
sd	0.30	0.12	0.24	0.14	5.83	19.12	0.83	0.32	0.06	12.11	0.56	0.19	0.16	1.36	0.37
max	1.00	1.00	4.00	0.79	27.55	73.75	6.07	1.00	0.62	29.99	2.75	4.95	7.64	6.16	1.00
min	0.00	0.00	0.00	0.16	0.01	0.00	-0.20	0.00	0.00	-22.37	0.00	0.00	0.00	1.00	0.00

See web Appendix for variables' sources and definitions

Table 3b: The Correlation Matrix for the Pairwise Analysis of Adjacent Regions

	% of Common Languages	Difference in Elevation	Difference in Land Quality	Geodesic Distance	Migratory Distance from East Africa	Absolute Latitude	Mean Elevation	Mean Land Quality	Mean Precip.	Mean Temp.	Sea Distance	Water Area	Pop. Dens in 1995	Area	Within Country
% of Common Languages	1.00														
Difference in Elevation	-0.16	1.00													
Difference in Land Quality	-0.16	0.27	1.00												
Geodesic Distance	-0.26	0.12	0.13	1.00											
Migratory Distance from East Africa	0.05	0.03	0.05	-0.03	1.00										
Absolute Latitude	0.39	-0.11	-0.09	-0.41	0.07	1.00									
Mean Elevation	-0.06	0.06	0.43	0.07	0.03	-0.11	1.00								
Mean Land Quality	-0.14	0.33	0.02	0.09	0.11	-0.18	-0.14	1.00							
Mean Precip.	-0.37	0.16	0.07	0.16	0.28	-0.46	-0.13	0.32	1.00						
Mean Temp.	-0.31	0.09	-0.02	0.38	-0.12	-0.90	-0.16	0.22	0.35	1.00					
Sea Distance	0.06	-0.04	0.07	-0.03	-0.09	0.10	0.32	-0.15	-0.28	-0.22	1.00				
Water Area	0.03	-0.05	-0.06	0.00	0.01	0.02	-0.04	-0.09	-0.08	0.02	-0.02	1.00			
Pop. Dens in 1995	-0.20	0.27	0.13	0.13	-0.18	-0.26	-0.01	0.56	0.29	0.31	-0.21	-0.02	1.00		
Area	-0.15	0.14	0.08	0.28	-0.19	-0.61	0.18	0.22	0.12	0.61	0.04	0.05	0.36	1.00	
Within Country	0.26	-0.10	-0.09	-0.09	0.12	0.15	-0.05	-0.07	-0.09	-0.14	0.02	-0.02	-0.13	-0.10	1.00

See web Appendix for variables' sources and definitions

Data Sources

Geographic Variables

Absolute Latitude: Absolute latitudinal distance from the equator in decimal degrees.
Source: At all levels of aggregation the distance from the equator is calculated from the centroid of the respective unit of analysis using ArcGIS.
Surface Area: surface area in 1000's of sq. km.
Source: At all levels of analysis the area is constructed using ArcGIS. Unless otherwise noted, I include only areas over which both language and land quality data are available.
Geodesic Distance: geodesic distance between the centroids of adjacent regions in 100's of km's.
Source: Calculated using the Haversine formula
Difference in Elevation: difference in elevation within pairs of adjacent regions in km.
Source: Constructed using information on elevation above the sea level at 0.5 dd x 0.5 dd. National Oceanic and Atmospheric Administration (NOAA) and U.S. National Geophysical Data Center, TerrainBase, release 1.0, Boulder, Colo. http://www.sage.wisc.edu/atlas/data.php?incdataset=Topography
Mean Elevation: average elevation within the unit of analysis in km.
Source: For the pair-wise analysis see Difference in Elevation. For the cross-country and cross-virtual country regressions the G-Econ database is used and the average elevation is constructed as the mean elevation across grids within the respective units of analysis. Each grid corresponds to an area of 1 dd x 1dd.
Dispersion of Elevation: the difference between the maximum and the minimum elevation.
Source: See Mean Elevation.
Variation in Elevation: standard deviation of elevation within real and virtual countries in km.
Source: G-Econ database. Variation in elevation across the 1 dd x 1 dd cells within a real or virtual country
Ln(Area): natural log of surface area.
Ln(Population Density in 1995): natural log of population density in 1995 in 1000's people per sq. km.
Source: Center for International Earth Science Information Network (CIESIN). Available at: http://www.sage.wisc.edu/iamdata/grid_data_sel.php
Mean Land Quality: average land quality within the respective unit of analysis.
Source: The raw dataset is available at the Atlas of the Biosphere. http://www.sage.wisc.edu/iamdata/
The mean land quality index, lq, is the product of two components capturing the climatic suitability for cultivation, lq_clim, and the soil suitability, lq_soil. Hence, lq=lq_clim*lq_soil. Each component is constructed in the following way: lq_clim=f_1(GDD)*f_2(m), where GDD denotes growing degree days and m is a moisture index capturing the availability of water to plants. Regarding soil characteristics, lq_soil=g_1(C_soil)*g_2(pH_Soil), where C_soil stands for soil carbon density and pH_soil captures the acidity or alkalinity of soil. Each functional form is derived from the probability density function of actual cropland area versus each component. For example, in the case of f_1(GDD) and f_2(m) according to Ramankutty et al (2002) a sigmoidal function best fits the observed empirical relationship between the fraction of a cell that was cultivated in 1990 and the GDD and m respectively. Specifically, $f_1(GDD)=1/(1+\exp(a(b-GDD)))$ and $f_2(m)=1/(1+\exp(c(d-m)))$ with a=0.0052, b=1334, c=14.705 and d=0.3295. The functional forms of g_1(C_soil) and g_2(pH_soil) are the following:

$g_1(C\_soil)=(a/(1+\exp(b(c-C\_soil))))*(a/(1+\exp(d(e-C\_soil))))$ with a=3.9157, b=1.3766, c=3.468 and d=-0.0791

$$g_2(pH\_soil) = \begin{cases} -2.085+0.475pH\_soil & \text{if } pH\_soil \le 6.5 \\ 1 & \text{if } 6.5 < pH\_soil < 8 \\ 1.0-2.0pH\_soil & \text{if } pH\_soil \ge 8 \end{cases}$$

Mean Climatic Suitability: average climatic suitability for agriculture within the respective unit of analysis.
Source: The Climatic Suitability is the climatic component of the land quality index. I wish to thank Navin Ramankutty, the author of this dataset for providing the individual components of the land quality index.
Variation in Climatic Suitability: standard deviation in climatic suitability for agriculture
Source: See Mean Climatic Suitability.
Mean Soil Suitability: average soil suitability for agriculture within the respective unit of analysis.
Source: The Soil Suitability is the soil component of the land quality index.
Variation in Soil Suitability: standard deviation in soil suitability for agriculture
Source: See Mean Soil Suitability.

Difference in Land Quality: absolute difference in land quality between adjacent regions.
Source: See Mean Land Quality
Dispersion in Land Quality: the difference between the maximum and the minimum land quality
Source: See Mean Land Quality
Variation in Land Quality: the standard deviation of land quality within the unit of analysis.
Source: See Mean Land Quality
Migratory Distance from East Africa: geodesic migratory distance of Addis Ababa from the centroid
Source: Constructed following Ramachandran et al (2005). The waypoint locations for Europe are Cairo and Istanbul. For the Americas are Anadyr in Northeastern Russia and Prince Rupert in Canada. Finally, for territories in the Southeast Asia and the Pacific, Pnomh Penh in Cambodia is the additional waypoint.
Mean Precipitation: average monthly precipitation between 1961 and 1990 in 1000's of millimeters.
Source: For the cross-country and cross-virtual country the source is the G-Econ database. For the pair analysis the source is the Climatic Research Unit (CRU) 2.0, Mitchell et al (2004).
Distance from the Sea: distance from the nearest coastline in 1000s of km's from the respective centroid
Source: For the cross-country analysis with underlying linguistic coverage from the Ethnologue and the cross-virtual country the estimate is derived from the G-Econ database. For the cross-country analysis in Table 3b the average distance to the coastline comes from www.cid.harvard.edu/ciddata/geographydata.htm For the regional-pair analysis the coastlines of seas, oceans, and extremely large lakes dataset is used after excluding the latter. Publisher and place: Global Mapping International, Colorado Springs, Colorado, USA. Series name: Global Ministry Mapping System. Series issue: Version 3.0
Mean Temperature: average temperature in Celsius between 1961 and 1990.
Source: For the cross-country and cross-virtual country the source is the G-Econ database. For the pair analysis the source is the Climatic Research Unit (CRU) 2.0, Mitchell et al (2004). It provides temperature estimates at the same resolution as the land quality data set, i.e., at a 0.5 by 0.5 decimal degrees.
Water Area: total area within the respective unit of analysis covered by rivers or lakes in sq. km.
Source: "Inland Water Area Features" dataset from Global Mapping International which offers a comprehensive global source of all rivers and lakes.
Sorghum: Indicator that takes the value of 1 if sorghum is the most commonly cultivated crop in the partition.
Source: The raw data constructed by Leff et al (2004) are available at: http://www.sage.wisc.edu/iamdata/ and provide information at a resolution of 0.5 by 0.5 decimal degrees on the fraction of land that is allocated towards each of 18 different crops including sorghum in 1992. To generate a measure of average cultivation per crop for the unit of analysis I calculate the average across the cells that fall within the respective unit.
Sorghum in the non-Adjacent Partition of the Same Group: Indicator that takes the value of 1 if sorghum is the most commonly cultivated crop in the non-Adjacent partition of the same linguistic group.
Source: See Sorghum
Sorghum across Buffer Regions: Indicator that takes the value of 1 if sorghum is the most commonly cultivated crop in the regions located in the buffer zone of 0.5 decimal degrees radius around the partition.
Source: See Sorghum
Maize: Indicator that takes the value of 1 if maize is the most commonly cultivated crop in the partition.
Source: See Sorghum
Maize in the non-Adjacent Partition of the Same Group: Indicator that takes the value of 1 if maize is the most commonly cultivated crop in the non-Adjacent partition of the same linguistic group.
Source: See Sorghum
Maize across Buffer Regions: Indicator that takes the value of 1 if maize is the most commonly cultivated crop in the regions located in the buffer zone of 0.5 decimal degrees radius around the partition.
Source: See Sorghum
Pasture: Indicator that takes the value of 1 if the percent of land allocated to pasture in the respectively partition is larger than the fraction of land allocated to farming.
Source: The raw data by Ramankutty et al (1998) available at: http://www.sage.wisc.edu/iamdata/ provide information at a resolution of 0.5 by 0.5 decimal degrees on the fraction of land that is allocated towards pasture and the fraction of land allocated towards cultivation in 1992. To generate a measure of pasture and farming intensity I calculate the respective average across the cells that fall in the relevant unit.
Pasture across Buffer Regions: Indicator that takes the value of 1 if the percent of land allocated to pasture in the buffer zone of 0.5 dd radius around the partition > the fraction of land allocated to farming.
Source: See Pasture
Pasture in the non-Adjacent Partition of the Same Group: Indicator that takes the value of 1 if the percent of land allocated to pasture in the non-adjacent partition of the same group > % of land allocated to farming.
Source: See Pasture

Other Variables

Timing of Transition to Agriculture: Year when the first significant region within a current country underwent a transition from reliance mainly on hunting and gathering to reliance mainly on cultivated crops (and livestock).
Source: Putterman, L., Agricultural Transition Data Set, Brown University, http://www.econ.brown.edu/fac/Louis_Putterman/agricultural%20data%20page.htm
Within Country: dummy equals 1 if a virtual country or pair of contiguous regions fall within a real country.
Source: Constructed using ArcGIS.
Indigenous: percentage of the current population's composition which was indigenous in these countries as of 1500 AD, version 1.1. http://www.econ.brown.edu/fac/Louis_Putterman/world%20migration%20matrix.htm
Source: Putterman, L., 2007, World Migration Matrix, 1500 -- 2000, Brown University.
Ln(Number of Languages): ln(number of languages) located within the respective unit of analysis.
Source: WLMS database.
Percentage of Common Languages: number of common languages divided by the total number of languages spoken within a regional pair.
Source: see Ln(Number of Languages)
Ln(Population Density in 1500 AD): log population density in 1500.
Source: McEvedy and Jones (1978), "Atlas of World Population History".
Number of Countries: number of real countries in which a virtual country belongs to
Source: Constructed using ArcGIS.
Year of Independence: year a country achieved independence.
Source: Fearon J., "Ethnic and Cultural Diversity by Country", originally from the Correlated of War database.
ELF: This index represents for each country the probability that two individuals randomly drawn from the overall population will belong to different ethnolinguistic groups. The construction of this index is based on data from a Soviet ethnographic source, NarodovMira and augmented by Fearon and Laitin (2003)
The sources used for augmenting NarodovMira missing country observations were the CIA Factbook, Encyclopedia Brittanica, and the Library of Congress Country Studies.
ELF3, ELF5, ELF7, ELF9 : Similar to ELF where the numbers. 3, 5, 7, 9 reflect the level of linguistic aggregation at which the fractionalization measures are constructed.
Source: Desmet, Ortuño-Ortín and Wacziarg (2011)

References

Atlas Narodov Mira (Atlas of the People of the World). 1964. *Atlas Narodov Mira (Atlas of the People of the World).* Moscow:Glavnoe Upravlenie Geodezii i Kartograi, Bruck, S.I., and V.S. Apenchenko.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. Forthcoming. "The Political Economy of Ethnolinguistic Cleavages." *Journal of Development Economics.*

Fearon, James, and David Laitin. 2003. "Ethnicity, Insurgency and Civil War." *American Political Science Review*, 97: 75–90.

G-Econ. 2006. *Available on-line at http://gecon.yale.edu.*

Leff, B, N Ramankutty, and J.A. Foley. 2004. "Geographic Distribution of Major Crops Across the World." *Global Biogeochemical Cycles*, 18(1).

Mitchell, Carter, Jones, and Hulme. 2004. "A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100)." Tyndall Centre Working Paper 55.

Ramachandran, Sohini, Omkar Deshpande, Charles C. Roseman, Noah A. Rosenberg, Marcus W. Feldman, and L. Luca Cavalli-Sforza. 2005. "Support from the Relationship of Genetic and Geographic Distance in Human Populations for a Serial Founder Effect Originating in Africa." *Proceedings of the National Academy of Sciences*, 102: 15942–15947.

Ramankutty, Navin, Jonathan A. Foley, and John Norman. 1998. "Characterizing Patterns of Global Land Use: An Analysis of Global Croplands Data." *Global Biogeochemical Cycles*, 12: 667–685.

Ramankutty, Navin, Jonathan A. Foley, John Norman, and Kevin McSweeney. 2002. "The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change." *Global Ecology and Biogeography*, 11: 377–392.

WLMS. 2006. *World Language Mapping System, Version 3.2 Available on-line at http://www.gmi.org/wlms/.*