STAT 230A Final project
Replication of Michalopoulos: The Origins of Ethnolinguistic Diversity

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

1 Paper summary & summary statistics table

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 this examination is that differences in local land characteristics induce different levels of human capital across locations, which in turn, gives rise to localized ethnicities that are characterized by separate languages. The results of the empirical study presented are found to be consistent with this hypothesis.

The empirical results are obtained separately by three regressions:

- **Cross-country**: this takes the current political borders as the unit within which covariates such as the number of languages are counted.
- **Virtual countries**: To account for the arbitrary nature of some political boundaries with respect to ethnolinguistic groupings, the world is split into arbitrary *virtual countries* and the regression is performed again.
- **Adjacent regions**: To account for a potentially high "baseline" effect in some regions, adjacent regions are compared directly, which neutralizes region-specific fixed effects and focuses on the effect of the variables under consideration.

1.2 Data statistics and summary table

The data comes from multiple sources: the standard geographic data was sourced from the *Geographically Based Economic Data database*, the data on land quality for agriculture comes from *Ramankutty et al. (2002)*, and the data on the distribution of languages comes from the *World Language Mapping System*. 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.

The paper lacks a true summary table and shows a couple of 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 - the *cross-country model*. Figure 1 shows the distribution of land suitability for agriculture across the world at a resolution of .5-by.5 decimal degrees. The dependent variable represents the probability that a particular grid cell may be cultivated.

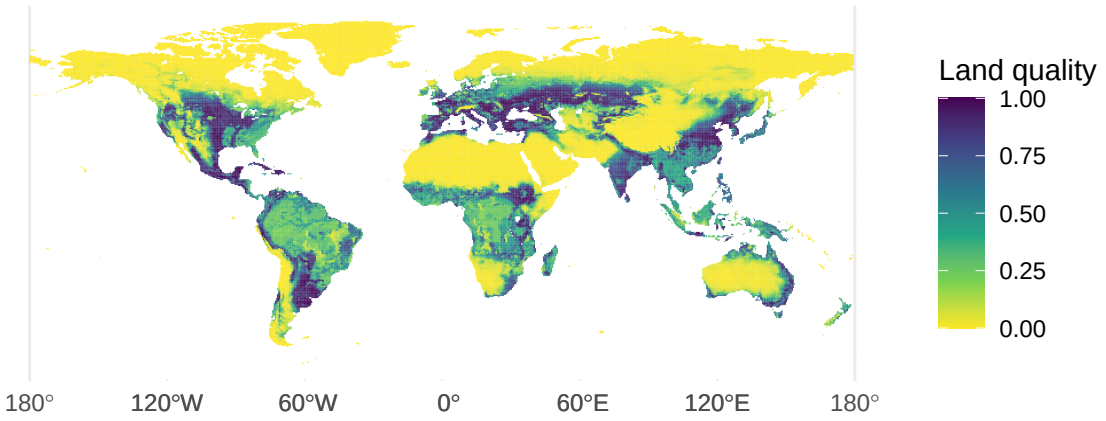

Figure 1: Land quality for agriculture across countries

Figure 2 shows the distribution of land quality within two countries selected in the paper—Greece and Nepal, obtained with a KDE using the Epanechnikov kernel.

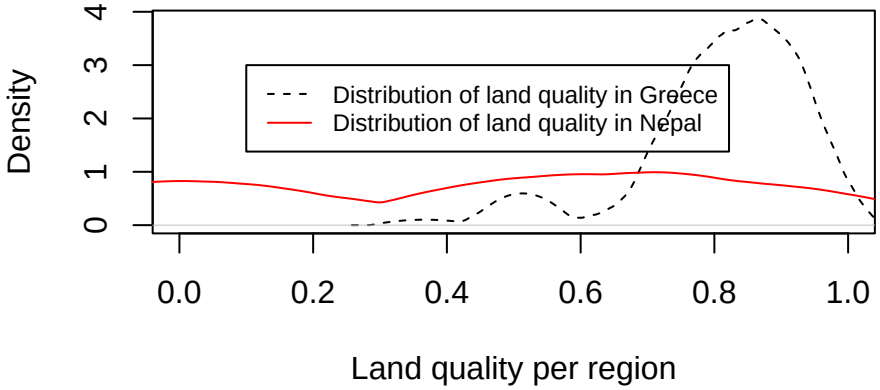

Figure 2: Kernel Density of Land Quality in Greece and Nepal

Below is our summary table of important variables for the first model. 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. Additional covariates are measures of centrality and variablity, 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

2 Model 1

2.1 Replication

```
##
## Call:
## lm(formula = lnLang ~ absLat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27119 -0.59554  0.03279  0.55611  2.06275
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.564e-17  7.051e-02   0.000        1
## absLat      -4.791e-01  7.073e-02  -6.773 2.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8806 on 154 degrees of freedom
## Multiple R-squared:  0.2295, Adjusted R-squared:  0.2245
## F-statistic: 45.87 on 1 and 154 DF,  p-value: 2.506e-10
```

```
##
## Call:
## lm(formula = lnLang ~ sdElev + sdSuitable + avgElev + avgSuitable +
##     absLat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70851 -0.53781 -0.03409  0.57628  1.73248
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.226e-17  6.280e-02   0.000  1.00000
## sdElev       3.096e-01  1.025e-01   3.020  0.00297 **
## sdSuitable   3.404e-01  7.752e-02   4.391 2.12e-05 ***
## avgElev     -2.489e-01  9.963e-02  -2.499  0.01355 *
## avgSuitable -1.791e-01  6.769e-02  -2.646  0.00901 **
## absLat      -5.467e-01  6.451e-02  -8.475 2.03e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7843 on 150 degrees of freedom
## Multiple R-squared:  0.4047, Adjusted R-squared:  0.3848
## F-statistic: 20.39 on 5 and 150 DF,  p-value: 1.693e-15
```

```
##
## Call:
## lm(formula = lnLang ~ sdElev + sdSuitable + avgElev + avgSuitable +
##     absLat + avgPrecip + avgTemp + lnArea + seaDist + migrationDist,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56760 -0.40968 -0.00431  0.36628  1.84350
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.554e-16  4.734e-02   0.000  1.00000
## sdElev         2.557e-01  9.364e-02   2.731  0.00710 **
## sdSuitable     1.775e-01  6.219e-02   2.854  0.00495 **
## avgElev       -1.111e-01  1.067e-01  -1.041  0.29953
## avgSuitable   -6.855e-02  5.736e-02  -1.195  0.23400
```

```
## absLat        -5.759e-02  1.871e-01  -0.308  0.75873
## avgPrecip      4.682e-01  8.447e-02   5.543 1.37e-07 ***
## avgTemp        2.698e-01  1.690e-01   1.597  0.11241
## lnArea         5.174e-01  5.929e-02   8.727 5.68e-15 ***
## seaDist        5.336e-02  6.639e-02   0.804  0.42282
## migrationDist -2.812e-01  6.292e-02  -4.469 1.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5912 on 145 degrees of freedom
## Multiple R-squared:  0.673,  Adjusted R-squared:  0.6504
## F-statistic: 29.84 on 10 and 145 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = lnLang ~ sdElev + sdSuitable + avgElev + avgSuitable +
##     absLat + avgPrecip + avgTemp + lnArea + seaDist + migrationDist +
##     lnPop95 + africa + europe + americas + reg_eap, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47171 -0.37683 -0.01812  0.35744  1.76098
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.675e-16  4.715e-02   0.000  1.00000
## sdElev         2.908e-01  9.777e-02   2.975  0.00345 **
## sdSuitable     2.083e-01  6.398e-02   3.256  0.00142 **
## avgElev       -1.044e-01  1.084e-01  -0.963  0.33723
## avgSuitable   -2.950e-02  6.875e-02  -0.429  0.66853
## absLat        -3.293e-02  2.187e-01  -0.151  0.88052
## avgPrecip      4.474e-01  9.380e-02   4.770 4.58e-06 ***
## avgTemp        3.848e-01  1.859e-01   2.069  0.04035 *
## lnArea         4.818e-01  6.283e-02   7.669 2.65e-12 ***
## seaDist        6.327e-02  6.920e-02   0.914  0.36210
## migrationDist -5.178e-01  1.940e-01  -2.670  0.00849 **
## lnPop95       -1.176e-01  7.139e-02  -1.647  0.10180
## africa        -8.817e-02  9.747e-02  -0.905  0.36723
## europe         5.777e-02  9.459e-02   0.611  0.54235
## americas       1.589e-01  1.690e-01   0.940  0.34868
## reg_eap        1.429e-01  8.885e-02   1.609  0.10997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5888 on 140 degrees of freedom
## Multiple R-squared:  0.6868, Adjusted R-squared:  0.6533
## F-statistic: 20.47 on 15 and 140 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = lnLang ~ sdElev + sdSuitable + avgElev + avgSuitable +
##     absLat + avgPrecip + avgTemp + lnArea + seaDist + migrationDist +
##     lnPop95 + lnPopDens1500 + entryYear + agriTran + africa +
##     europe + americas + reg_eap, data = data, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18050 -0.36356 -0.03966  0.33667  1.81222
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.816e-18  4.669e-02   0.000  1.00000
## sdElev         2.754e-01  1.014e-01   2.717  0.00754 **
## sdSuitable     2.113e-01  6.397e-02   3.303  0.00125 **
```

```
## avgElev       -8.478e-02  1.099e-01  -0.771  0.44203
## avgSuitable    6.195e-03  6.694e-02   0.093  0.92641
## absLat        -1.313e-01  2.170e-01  -0.605  0.54626
## avgPrecip      4.787e-01  9.515e-02   5.031 1.68e-06 ***
## avgTemp        4.041e-01  1.812e-01   2.230  0.02760 *
## lnArea         4.636e-01  7.378e-02   6.284 5.20e-09 ***
## seaDist        7.260e-02  6.924e-02   1.048  0.29650
## migrationDist -5.135e-01  2.124e-01  -2.418  0.01709 *
## lnPop95        2.284e-02  8.653e-02   0.264  0.79231
## lnPopDens1500 -2.349e-01  1.002e-01  -2.346  0.02058 *
## entryYear     -1.082e-01  7.100e-02  -1.524  0.12996
## agriTran       1.338e-01  1.042e-01   1.284  0.20143
## africa        -1.414e-02  1.501e-01  -0.094  0.92511
## europe         1.744e-01  1.125e-01   1.550  0.12367
## americas       6.025e-02  1.725e-01   0.349  0.72742
## reg_eap        1.011e-01  8.840e-02   1.143  0.25514
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5564 on 123 degrees of freedom
##   (14 observations deleted due to missingness)
## Multiple R-squared:   0.73,  Adjusted R-squared:  0.6904
## F-statistic: 18.47 on 18 and 123 DF,  p-value: < 2.2e-16
```

2.2 Robustness check

```
##
## Call:
## glm.nb(formula = exp(lnLang) ~ sdElev + sdSuitable + africa +
##     europe + americas + reg_eap, data = data, na.action = na.exclude,
##     init.theta = 5.342049308, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3019  -0.9245  -0.6440   0.3882   2.9764
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.28553    0.08318   3.433 0.000597 ***
## sdElev       0.18639    0.08099   2.301 0.021373 *
## sdSuitable   0.24233    0.08675   2.793 0.005216 **
## africa       0.41474    0.11686   3.549 0.000386 ***
## europe      -0.15924    0.13396  -1.189 0.234571
## americas     0.05827    0.10086   0.578 0.563485
## reg_eap      0.38348    0.08613   4.453 8.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(5.342) family taken to be 1)
##
##     Null deviance: 222.35  on 155  degrees of freedom
## Residual deviance: 158.13  on 149  degrees of freedom
## AIC: 480.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  5.34
##           Std. Err.:  2.08
##
##  2 x log-likelihood:  -464.097
```

NOTE: ~ matches on the land quality coeff, but not on elevation. What to do with the theta MLE?

```
##
## Call:
## lm(formula = lnLang ~ dispElev + dispSuitable + africa + europe +
##     americas + reg_eap, data = data, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31931 -0.43570  0.06013  0.48643  1.65734
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.32210    0.11028  -2.921 0.004035 **
## dispElev      0.23161    0.06649   3.484 0.000650 ***
## dispSuitable  0.28060    0.06761   4.150 5.57e-05 ***
## africa        0.37075    0.09420   3.936 0.000127 ***
## europe       -0.09152    0.09373  -0.976 0.330477
## americas      0.02498    0.07850   0.318 0.750781
## reg_eap       0.24755    0.07638   3.241 0.001468 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7758 on 149 degrees of freedom
## Multiple R-squared:  0.4034, Adjusted R-squared:  0.3794
## F-statistic: 16.79 on 6 and 149 DF,  p-value: 9.487e-15
```

3 Appendix: code

```
cells = read_sf(dsn = 'data_raw/Virtual_country', layer = 'virtual_cntrygrid')
countries = read_sf(dsn = 'countries', layer = 'countries')
data = read.dta13("data_raw/Tables1-3a.dta")
colnames(data) = c('countryCode', 'entryYear', 'countryName', 'avgTemp',
                   'avgPrecip', 'seaDist', 'avgElev', 'sdElev', 'absLat',
                   'dispElev', 'numLang', 'suitableCells', 'dispSuitable',
                   'climate', 'soil', 'sdClimate', 'sdSoil', 'sdSuitable',
                   'avgSuitable', 'pop95', 'area', 'lnLang', 'africa',
                   'europe', 'americas', 'lnPop95', 'migrationDist',
                   'lnArea', 'pctIndigenous', 'lnPopDens1500',
                   'agriTran', 'reg_eap')
greeceCells = countries %>% filter(COUNTRY == 'Greece') %>%
  st_intersection(y = cells)
nepalCells = countries %>% filter(COUNTRY == 'Nepal') %>%
  st_intersection(y = cells)
plot(
  density(greeceCells$suit_new, kernel = "epanechnikov"),
  xlim = c(0, 1),
  xlab = 'Land quality per region',
  ylab = 'Density',
  main = '',
  lty = 2
)
lines(density(nepalCells$suit_new, kernel = "epanechnikov"), col = 'red')
legend(
  .1,
  3,
  legend = c(
    'Distribution of land quality in Greece',
    'Distribution of land quality in Nepal'
  ),
  col = c("black", "red"),
  lty = 2:1,
  cex = .75
)
count = function(x) {
  (sum( ~ is.na(x)))
}

sumTable <- data %>% select(
  c(
    'numLang',
    'sdElev',
    'sdSuitable',
    'avgElev',
    'avgSuitable',
    'absLat',
    'avgPrecip',
    'avgTemp',
    'lnArea',
    'seaDist',
    'migrationDist',
    'lnPop95'
  )
) %>%
  summarise_each(
    funs(
      min = min,
      median = median,
      max = max,
      mean = mean,
```

```
      iqr = quantile(., 0.75) - quantile(., 0.25),
      sd = sd,
      n = sum(!is.na(.))
    )
  ) %>%
  gather(var, val) %>%
  separate(var, into = c("var", "stat"), sep = "_") %>%
  spread(var, val) %>% column_to_rownames(var = "stat") %>%
  select(
    c(
      'numLang',
      'sdElev',
      'sdSuitable',
      'avgElev',
      'avgSuitable',
      'absLat',
      'avgPrecip',
      'avgTemp',
      'lnArea',
      'seaDist',
      'migrationDist',
      'lnPop95'
    )
  ) %>%
  mutate_if(is.numeric, ~ round(., 2)) %>% slice(5, 4, 2, 3, 7, 6)
standardize = function(vec) {return ((vec - mean(vec, na.rm = TRUE)) / sd(vec, na.rm = TRUE))}

# NOTE: I'm lazy and don't want to add vars manually
modelCols = c('entryYear', 'avgTemp', 'avgPrecip', 'seaDist', 'avgElev',
              'sdElev', 'absLat', 'numLang', 'dispSuitable', 'climate',
              'soil', 'sdClimate', 'sdSoil', 'sdSuitable', 'avgSuitable',
              'pop95', 'area', 'lnLang', 'lnPop95', 'migrationDist',
              'lnArea', 'pctIndigenous', 'lnPopDens1500', 'agriTran',
              'americas', 'europe', 'africa', 'reg_eap')
for (col in modelCols) {
 data[,col] = standardize(data[,col])
}

# If we're standardizing all:
# data <- data %>% mutate(across(!countryName & !countryCode , standardize))
```