Diverse Mechanisms of Migration

Scholars across disciplines have identified several mechanisms that cause people to migrate. Some propose an "income maximizer" hypothesis and argue that individuals migrate because they are drawn to higher wages in receiving countries. Others argue that it is risk and uncertainty in the sending countries–such as low-wages and lack of market opportunities– that is driving migration patterns. They offer a "risk diversifier" hypothesis. Others still hypothesize that growing ties among individuals in receiving and sending countries fosters immigration, and advocate for analyses that focus on "network migrants." In this exercise, rather than examining them as competing hypotheses, we examine these theories together and test whether each represents the profile of a different stream of migrants from Mexico to the U.S. in recent decades. Using cluster analysis, we attempt to discover the "configurations of various attributes that characterize different migrant types." This exercise is based on the following article:

Garip, Filiz. 2012. "Discovering Diverse Mechanisms of Migration: The Mexico–US Stream 1970–2000." *Population and Development Review*, Vol. 38, No. 3, pp. 393-433.

The data come from the **Mexican Migration Project**, a survey of Mexican migrants from 124 communities located in major migrant-sending areas in 21 Mexican states. Each community was surveyed once between 1987 and 2008, during December and January, when migrants to the U.S. are mostly likely to visit their families in Mexico. In each community, individuals (or informants for absent individuals) from about 200 randomly selected households were asked to provide demographic and economic information and to state the time of their first and their most recent trip to the United States.

The data set is the file `migration.csv`. Variables in this dataset can be broken down into three categories:

INDIVIDIUAL LEVEL VARIABLES

Name	Description
`year`	Year of respondent's first trip to the U.S.
`age`	Age of respondent
`male`	1 if respondent is male, 0 if respondent is female
`educ`	Years of education: secondary school in Mexico is from years 7 to 12

HOUSEHOLD LEVEL VARIABLES

Name	Description
`log_nrooms`	Logged number of rooms across all properties owned by respondent's household
`log_landval`	Logged value of all land owned by respondent's household (U.S. dollars)
`n_business`	Number of businesses owned by respondent
`prop_hhmig`	Proportion of respondent's household who are also U.S. migrants

COMMUNITY LEVEL VARIABLES

Name	Description
prop_cmig	Proportion of respondent's community who are also U.S. migrants
log_npop	Logged size of respondent's community.
prop_self	Proportion of respondent's community who are self-employed
prop_agri	Proportion of respondent's community involved in agriculture
prop_lessminwage	Proportion of respondent's community who earn less than the U.S. minimum wage

Question 1

Examine the mean values for the individual level, household level, and community level characteristics in the dataset. Briefly interpret your answers.

Question 2

Use scatterplots to investigate the relationship between `prop_self` and `prop_agri`, as well as the relationship between `prop_self` and `log_npop`. Briefly interpret these scatter plots and what they imply about self-employed workers. Do these relationships appear to be independent? What does knowing that a migrant is self-employed tell us about them? Then calcuate the correlation for all possible interactions of the four community level variables: `prop_self`, `prop_agri`, `prop_lessminwage`, and `log_npop`. Use these correlations to help with your interpretation of the scatter plots. Does adding the `prop_lessminwage` variable add anything to your interpretation?

Question 3

We'll focus on the variables: `year`, `educ`, `log_nrooms`, `log_landval`, `n_business`, `prop_hhmig`, `prop_cmig`, `log_npop`, `prop_self`, `prop_agri`, and `prop_lessminwage`. Remove observations with missing values. Then, subset your dataset to all of your variables **except** `year`, and use the `scale()` function to standardize the variables in your subsetted dataset so that they are comparable. Compare the means and standard deviations before and after scaling. Standardizing substracts the mean of a variable from each observation and divides by the standard deviation.

Question 4

Fit the k-means clustering algorithm with *three* clusters, using the scaled variables from the data set with no missing values. Insert the code `set.seed(2016)` right before your cluster analysis so that you can compare your results from the kmeans clustering to exercise solutions later. How many observations are assigned to each cluster? Each cluster has a center. What do the centers of these clusters represent? Interpret the type of migrant described by cluster 1. To help witih interpretability, you can also calculate the mean value of the variables for each cluster, using their original scale. Repeat the cluster analysis. This time with *four* centers. How are the two results different? Is there one you prefer?

Question 5

Do these different clusters represent different temporal trends in migration from Mexico to the US? Use a time-series plot to graph the proportions of migrants in each of the four clusters from Question 4 over time (variable `year`). Briefly describe the major trends you discover.