READ ME
December 28, 2010

Division of Labor

I use Stata, Matlab and R for the analyses. As a rule of thumb, all data manipulations are done in Stata. Clustering is done in Matlab (in order to take advantage of multiple runs within k-means that gets us closer to global minima), and graphs are done in R. Listed below are the most important programs.

Folders: STATA Analysis and Data

mmp_cluster_data11.do: Listed under the Data folder. Creates the data with all migrants from the PERS folder on their first trip to the United States. Transfers data (Data/…/ Clustering/rdata1.raw) to Matlab.

mmp_cluster_data9_mig_nonmig.do: Listed under the Data folder. Creates the data with all individuals, migrants and non-migrants, from the PERS folder. Keeps all individuals on the survey year. This data set is necessary for the descriptive table that compares those who have migrated at least once to those who have never migrated in Analysis/Stata Analysis/descriptive_tables.do.

descriptive_tables.do: Listed under the Analysis/Stata Analysis folder. Creates a simple descriptive table comparing migrants to non-migrants, and performs t-tests. The output into a text file is later re-formatted in the Excel file Tables Graphs/clustering_tables.xls.

mmp_cluster_data_FULL.do: Listed under the Data folder. Creates the data with all migrants and non-migrants from the LIFE folder (therefore, only household heads). Transfers data (Data/…/Clustering/ fulldata.raw) to Matlab. This data set is required to observe the changes in coefficient estimates over time. (NOTE the regressions for this purpose are performed in Matlab.)

outcome_regressions.do: Listed under the Analysis/Stata Analysis folder. Takes the k-means solution from Matlab (saved under the Analysis folder as cluster.txt) along with the data. Runs regressions of post-migration outcomes. This is done in Stata in order to obtain regression diagnostics (such as pseudo-R2 that are not automatically computed in Matlab). Writes the results

in outcome_regressions.txt which is re-formatted in the Excel file Tables Graphs/clustering_tables.xls.

cluster_analysis.do: Listed under the Analysis/Stata Analysis folder. Outdated cluster (kmeans) analysis. We no longer use Stata for clustering since it allows only a single run. Matlab is better for obtaining global minima over several runs.

mmp_panel_for_clustering.do: Listed under Data/Panel data for regressions. Creates a longitudinal data set for **all** individuals (not just household heads). The basis is the PERS file. I use Maocan Guo's code – saved as mguo_data_setup_partI.do and *_partII.do – in the initial part of the code. Non-migrants are observed in all years, migrants are observed through the year of their first migration. The resulting data set mmp_panel_data_for_clustering.dta is used for regression analysis saved under Analysis/Stata Analysis/panel_data_regressions.do (explained below).

Folder: MATLAB Analysis

cluster_indivs14.do: Listed under the Analysis/Matlab Analysis folder. This is the most important analysis file generating the majority of the results. I perform k-means in Matlab, rather than in R or Stata, because it allows us to run several iterations and pick the optimal one. This way it is less likely to get stuck in local optima. This program also does several steps required for tables and plots done in R or ArcGIS. (a) Computes clusters and saves as 'cluster.txt'. (b) Computes the distance of each observation to cluster centroids for heatmaps, saves as 'dist.txt'. (c) Computes state-level descriptives (e.g., distribution of clusters across or within states) for maps in ArcGIS and saves as text files to be transferred first to Excel and then as an attribute table to ArcMap. (d) Prepares tables of cluster centroid characteristics, along with ttests. The results copied into a text file (manually) are then transferred to excel tables in Tables Graphs/clustering tables.xls. (e) Reads in the fulldata.raw (from life history file, created in mmp_cluster_data_FULL.do, and runs regressions at each year separately to show how regression coefficients change. The results are transferred to a text file 'coef.txt' to be plotted in R.

cluster_variables.m: This is an old file, in which I cluster the data with respect to variables rather than individuals. This is useful to see the correlation structure among variables. A similar analysis would be heatmaps of variables, which is easier to interpret.

explore_data.m: This is an old file, no longer used in the paper, that takes a first cut at clustering, tries a number of different visualization methods (PCA, etc.).

Matlab Plots: A folder containing plots saved in Matlab – silhouette plots, kdensity plots, etc. These are no longer used in the paper.

Folder: R Mclust

This is a folder with R analysis with Mclust package (Fraley and Raftery). Fits Gaussian mixture models, and determines the optimal number of clusters under a variety of models. For our purposes, this is too rigid - we cannot use binary variables, and the solution always seems to be with a high number of (>9) clusters.

Cluster1.R: Mclust clustering with raw data.

Cluster1_pca_results.R: Mclust with 3 PCA dimensions based on the full (binary and continuous) data.

Folder: R Descriptives

Almost all figures are done in R. The data comes from text files saved by Matlab.

R_Heatmaps3.R: Draws heatmaps of individuals on full and restricted (randomly selected or n-closest to cluster centroids, obtained from Matlab) data. The program takes about 10 hours on full data. Restricted data is much faster, and also allows us to display other variables (in a picket plot under the heatmap), such as repeat migration, under the heatmap. (In full data, there are too many points to generate intelligible graphs). Several different versions (with different data cuts, and variable selections) are saved as pdf files in the R Descriptives folder, and then copied to Tables Graphs.

R_Cluster_Validation2.R: Uses cluster.stats and clValid packages to compute several cluster validation measurs. Some measures take a long time

(like Hubert Gamma) therefore a subset of the data is used for validation. The results of several runs are saved manually to Cluster_Validation_Results.doc and then transferred to R_Cluster_Validation_Graphs.R for plots.

R_Cluster_Validation_Graphs.R: Takes cluster validation measures computed in R_Cluster_Validation2.R and saved in Cluster_Validation_Results.doc and saves plots as pdf.

All Graphs in R.R: Reads in text files generated in Matlab (cluster_indivs.m) and plots the following graphs: (1) trends in the distribution of migrants across clusters, (2) distribution of migrants across clusters and economic trends, (3) trends in coefficient estimates over time, (4) trends in the distribution of migrants across clusters and emergence of theories. All graphs are saved as pdf in R Descriptives folder, and then copied to Tables and Graphs folder.

Folder: ArcGIS Maps

Mexico state map is used to display the distribution of clusters within and across states. The data comes from the Matlab program (cluster_indivs.m), saved in the Analysis folder (e.g., state_clust*.txt, clust_across_state.txt) and transferred to an excel file under ArcGIS Maps/Map Descriptives.xls. All maps are in .mxd files, and excel data is transferred to an attribute table. (See the description in How to Use ArcGIS.doc) A separate map is drawn for each cluster (and year interval) and maps are combined in Adobe Illustrator. Final figures are saved as pdf and copied to Tables Graphs folder.