

California data analysis read-me

September 10, 2013

- California4.do:** Starts with the original data (all campuses) – keeps only individuals who have died (for whom we have more extensive information), and those who have registered after 2009 (pre-need conversion). The code checks for the viability of the MCAR (missing completely at random) assumption, creates variables and saves the data for Matlab. Computes descriptives by cluster ids. The code also codes occupation categories per SSM reviewers' request. The analyses do not appear in the final product.
- Clusters_new*.m:** Uses the new data created by Californi4.do. Implements k-means clustering on the data imported from Stata. Then transfers data back to Stata to compute the descriptives by cluster ids. * - original data + occupations, orig – original data from the first round, full – original data + observations with missing migprt added + occupations, temp – same as the unnamed, but with alternative coding of occupations
- Calif_cluster_validation.r:** Uses the new data created by California4.do. Implements cluster validation measures in R. Uses the data imported from Stata. Naming convention same as the Matlab file.s