Stage 1 – Creation of the Distance Matrix for All of the 3,141 U.S. Counties

Software: Stata SE, version 8.2
Program: 2000DistanceMatrix_AllCounties.do
Input: 2kresco_us.dct, 2kresco_us.txt
Output: 2000DistanceMatrix_AllCounties.txt, 2000DistanceMatrix_AllCounties_log.txt

Notes:

File *2kresco_us.txt* is the Census 2000 dataset on commuting patterns and can be downloaded from http://www.census.gov/population/cen2000/commuting/2KRESCO_US.zip.

File *2kresco_us.dct* is the dictionary file used by Stata to read *2kresco_us.txt*. It is based upon the original file layout found at http://www.census.gov/population/cen2000/commuting/coxcolayout.txt.

File *2000DistanceMatrix_AllCounties.txt* is the distance matrix that is used in the following stage.

File *2000DistanceMatrix_AllCounties_log.txt* is the log of the process.

This stage of the method requires Stata SE rather than standard Intercooled Stata due to the size of the distance matrix.

Stage 2 – Cluster Analysis of 3,141 U.S. Counties

Software: Base SAS and SAS/Stat, version 9.1.3 Service Pack 1
Program: 2000_AllCounties.sas
Input: 2000DistanceMatrix_AllCounties.txt
Output: Tree2000_AllCounties.csv, Results2000_Calibration.csv, 2000_AllCounties.log, 2000_AllCounties.lst

Notes:

File *2000DistanceMatrix_AllCounties.txt* is the distance matrix created in the previous stage.

File *Tree2000_AllCounties.csv* is the full dendrogram of the cluster analysis. A full description of the structure of this output file can be found in the SAS/Stat documentation for the TREE procedure. The documentation can be found at http://support.sas.com/documentation/onlinedoc/91pdf/index_913.html. Two variables are most important here: _NCL_ and _HEIGHT_. _NCL_ indexes the partitions by the number of clusters in the partition. _HEIGHT_ indicates the height, or average between-cluster distance of the partition.

File *Results2000_Calibration.csv* defines the composition of each of the clusters for 516 of the 3,141 possible partitions. Each row of this file is a county, while each column represents a different partition and lists a numerical identifier of the cluster that the county belongs to in that partition. This file is used in the following stage.

File *2000_AllCounties.log* is the log of the process.

File *2000_AllCounties.lst* is the text from the SAS output window. It is an abbreviated version of the dendrogram.

Stage 3 - Calibration

Software: Stata SE, version 8.2
Program: Calibration.do
Input: Results2000_Calibration.csv, Counties_ArbitronCoded.csv
Output: Calibration_log.txt

Notes:

File *Results2000_Calibration.csv* lists the counties in each cluster for 516 of the 3,141 possible partitions. It is created in the previous stage.

File *Counties_ArbitronCoded.csv* lists the counties in each of the 292 Arbitron markets in the 50 States and the District of Columbia as of Fall 2004. The source for this is http://www.arbitron.com/radio_stations/mktdefs.asp. In addition the file lists all of counties that are not in an Arbitron market. The list of counties are those in existence as of April 1, 2000; the date of the 2000 decennial Census. Consequently it does not include Broomfield County, Colorado. It does include the formerly independent city of Clifton Forge, Virginia. The field "ID" is simply a numerical identifier for the Arbitron market the county belongs in. It is blank if the county is not in an Arbitron market. The field "total" takes on a value greater than 0 when a county is part of more than one Arbitron market or only a portion of the county is inside an Arbitron market. The "total" field was developed by examining the map of Arbitron markets at http://www.arbitron.com/downloads/Arb_US_Metro_Map_04.pdf. Field "Population_2000" is used to construct the calibration weights. It is the population of the county from the 2000 decennial Census. The purpose of this file at this stage is to identify the composition of the Arbitron markets and the weights that will be used for the calibration.

File *Calibration_log.txt* is the log of the process. It contains the score of each of the 516 partitions evaluated in this stage. Examination of this file indicates that the maximum score occurs when the partition consists of 619, 620, or 621 clusters. The value of the height of these partitions can be found in file

Tree2000_AllCounties.csv. The partition size is contained in variable _NCL_ and the height is in variable HEIGHT .

Stage 4 – Creation of the Distance Matrix for 2,199 U.S. Counties not in an Arbitron Market

Software: Stata SE, version 8.2
Program: 2000DistanceMatrix_NonArbNoSplit.do
Input: 2kresco_us.dct, 2kresco_us.txt, Counties_ArbitronCoded.csv
Output: 2000DistanceMatrix_NonArbNoSplit.txt,
2000DistanceMatrix_NonArbNoSplit_log.txt

Notes:

File *2kresco_us.txt* is the Census 2000 dataset on commuting patterns and can be downloaded from http://www.census.gov/population/cen2000/commuting/2KRESCO_US.zip.

File *2kresco_us.dct* is the dictionary file used by Stata to read *2kresco_us.txt*. It is based upon the original file layout found at http://www.census.gov/population/cen2000/commuting/coxcolayout.txt.

File *Counties_ArbitronCoded.csv* lists the counties in each of the 292 Arbitron markets in the 50 States and the District of Columbia as of Fall 2004. The source for this is http://www.arbitron.com/radio_stations/mktdefs.asp. In addition the file lists all of counties that are not in an Arbitron market. The list of counties are those in existence as of April 1, 2000; the date of the 2000 decennial Census. Consequently it does not include Broomfield County, Colorado. It does include the formerly independent city of Clifton Forge, Virginia. The field "ID" is simply a numerical identifier for the Arbitron market the county belongs in. It is blank if the county is not in an Arbitron market. The field "total" takes on a value greater than 0 when a county is part of more than one Arbitron market or only a portion of the county is inside an Arbitron market. The "total" field was developed by examining the map of Arbitron markets at http://www.arbitron.com/downloads/Arb_US_Metro_Map_04.pdf. The purpose of the file at this stage is to identify the counties that are not in an Arbitron market.

File *2000DistanceMatrix_NonArbNoSplit.txt* is the distance matrix that is used in the following stage.

File *2000DistanceMatrix_NonArbNoSplit_log.txt* is the log of the process.

This stage of the method requires Stata SE rather than standard Intercooled Stata due to the size of the distance matrix.

Stage 5 – Cluster Analysis of 2,199 U.S. Counties

Software: Base SAS and SAS/Stat, version 9.1.3 Service Pack 1
Program: 2000_NonArbNoSplit.sas
Input: 2000DistanceMatrix_NonArbNoSplit.txt
Output: Tree2000_NonArbNoSplit.csv, Results_NonArbNoSplit626.csv, 2000_ NonArbNoSplit.log, 2000_ NonArbNoSplit.lst

Notes:

File *2000DistanceMatrix_NonArbNoSplit.txt* is the distance matrix created in the previous stage.

File *Tree2000_NonArbNoSplit.csv* is the dendrogram of the cluster analysis performed on 2,199 counties not in an Arbitron market. A full description of the structure of this output file can be found in the SAS/Stat documentation for the TREE procedure. The documentation can be found at http://support.sas.com/documentation/onlinedoc/91pdf/index_913.html. Two variables are most important here: _NCL_ and _HEIGHT_. _NCL_ indexes the partitions by the number of clusters in the partition. _HEIGHT_ indicates the height, or average between-cluster distance of the partition. An examination of the _HEIGHT_ variable indicates that only one partition has a height that falls within the optimal range identified in stage 3. This partition contains 626 clusters.

File *Results_NonArbNoSplit626.csv* defines the composition of the 626 clusters in the optimal partition. The first column contains the FIPS code of each county, while the second column is a numerical identifier of the cluster that the county is assigned to.

File *2000_ NonArbNoSplit.log* is the log of the process.

File *2000_ NonArbNoSplit.lst* is the text from the SAS output window. It is an abbreviated version of the dendrogram.