Flow Mapping with
Graph Partitioning and Regionalization

User Manual (Version 1.0)

Diansheng Guo

Department of Geography
University of South Carolina

Website:
www.SpatialDataMining.net

June 25, 2010

1. Overview

Flow Mapping with Graph Partitioning and Regionalization is an integrated software tool to explore flow patterns in large spatial interaction data. It involves two major steps: (1) using spatially constrained graph partitioning to find a hierarchy of natural regions defined by spatial interactions; and (2) rendering a flow map based on the discovered regions and related attributes. The first step will need the "**GraphREDCAP**" software package, which is to partition the graph and derive regions. The second step will need the "**FlowMap**" tool, which displays aggregated region-to-region flows. This manual focuses on the “**FlowMap**” tool.

The flow map allows the examination of flows between regions based on given regionalization results. The multivariate information of flows is aggregated based on the same regionalization result. A self-organizing map is then used to perform the multivariate clustering analysis. Flow lines are colored based upon the clustering results, enabling the understanding of multivariate information and flow structure at the same time.

A variety of user interactions (e.g., brushing and focusing) are supported to efficiently facilitate the exploration and accurate interpretation of spatial interaction patterns.

2. Relevant Publications

Guo, D., M. Gahegan, et al. (2005). "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach." Cartography and Geographic Information Science 32(2): 113-132.

Guo, D. (2009). "Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data", IEEE Transactions on Visualization and Computer Graphics, **15**(6), pp. 1041-1048

Guo, D. (2009). "Greedy Optimization for Contiguity-Constrained Hierarchical Clustering", The Fourth International Workshop on Spatial and Spatiotemporal Data Mining, IEEE International Conference on Data Mining (ICDM 2009), Miami, FL.

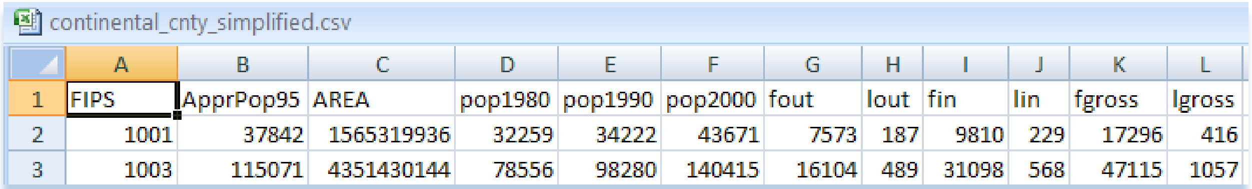
3. Data Files

This package requires following data files:

- **Shape file** (*. **shp**): an ArcGIS file storing spatial data (polygon or point type);
- **Attribute table** (*.**csv**): containing the IDs and attributes of the spatial objects;
- **Region file**(*.**rgn**, or *.**csv**).: regionalization result
- **Flow files** (*.**mvr**, or *.**csv,** one or more): a list of the flow volume and the multivariate data associated with origin-destination pairs.

1. Shape file

The spatial information needs to be stored in the ArcGIS shape file. Pint-type shapes are transformed to polygon-type and are then treated as polygons in the flow map. One CSV file is required to store the **ID** of spatial objects. Other attribute fields are allowed but not mandatory. A part of an example CSV file is shown in Figure 1.



continental_cnty_simplified.csv

	A	B	C	D	E	F	G	H	I	J	K	L
1	FIPS	ApprPop95	AREA	pop1980	pop1990	pop2000	fout	lout	fin	lin	fgross	lgross
2	1001	37842	1565319936	32259	34222	43671	7573	187	9810	229	17296	416
3	1003	115071	4351430144	78556	98280	140415	16104	489	31098	568	47115	1057

Figure 1 CSV File Structure

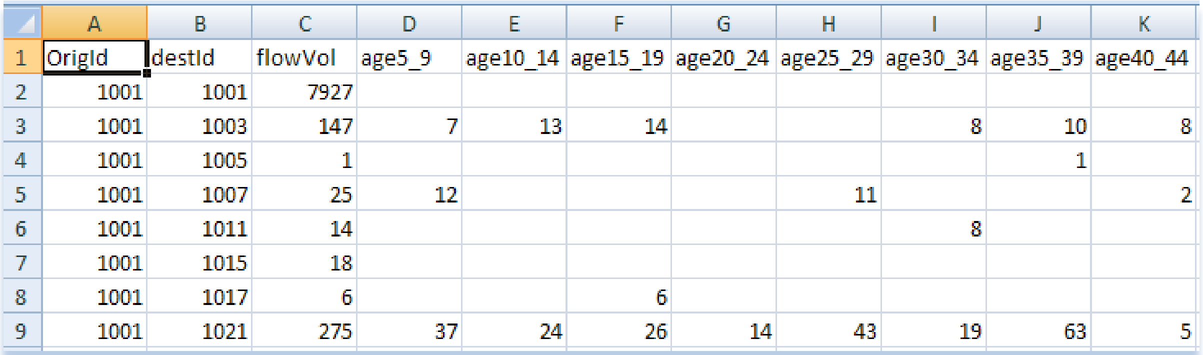
Following the procedures below, a dBASE file (*.**dbf**) in the ArcGIS shape can be converted to a **CSV** file.

1) Open the dbf file (of the shape file) in Excel. Make sure the names of attribute fields are in the first row.
2) Make sure that the **first** column of this file (FIPS in Figure 1) is the ID field of the spatial units. Make sure there are **no duplicate** IDs. The number and the order of the IDs should be **consistent** with the polygons (or points) in the shape file. The IDs can be strings/texts (i.e., state names) or integers (i.e., 5-digit zip code).
3) Save the file with the **same** name in CSV format.

User can choose other sources of CSV files as long as it meets the requirements above: field names in the first row; ID in the first column; no duplicate IDs; the number and the order of IDs are consistent with the shape file.

2. Flow File

The flow file stores **flow volumes**. It can also store **multivariate information** of flows. A part of an example flow file is displayed in Figure 2. Each row represents the flow between a unique pair of locations. For example, cell C3 in Figure 2 represents the flow size (147) from the location 1001 (cell A3) to the location 1003 (cell B3). The other cells (D3-K3) in the same row store the age composition flows. Empty cells in flow files are equivalent to zeros. In other words, if there is no flow between a pair of locations, there can be no entry for that pair in the flow table. In other words, it is possible the flow file does not contain all IDs in the CSV file. This makes the storage of sparse flow network very efficient. If there are no values (i.e., zeros) for certain multivariate fields, they can be blanks as well. When creating the flow file, it is helpful to note that:



	A	B	C	D	E	F	G	H	I	J	K
1	OrigId	destId	flowVol	age5_9	age10_14	age15_19	age20_24	age25_29	age30_34	age35_39	age40_44
2	1001	1001	7927								
3	1001	1003	147	7	13	14			8	10	8
4	1001	1005	1							1	
5	1001	1007	25	12				11			2
6	1001	1011	14						8		
7	1001	1015	18								
8	1001	1017	6			6					
9	1001	1021	275	37	24	26	14	43	19	63	5

Figure 2 Flow File Structure

1) The first row stores field names. Flow file has **at least three** columns. The first column must be the **origin ID** and the second column must be the **destination ID**. The third column is the flow volume. If there is multivariate information, it is stored in other columns.

2) If the flow file has only three columns (no multivariate information) and it is the only flow file, the flow map will be only about flow structure.

3) The IDs in the flow list must be consistent with those in the CSV file. In other words, one ID in the flow file and the same ID in the CSV represent the same spatial object.

4) If the flow file contains IDs that do not exist in the flow file, entries with these IDs as the origin or the destination are ignored and will **NOT** be processed.

5) The order of the origin-destination ID pair does not affect the analysis.

6) Multiple flow files are allowed. Each multivariate must follow the same rules as stated above. The first three columns in each flow file can be exactly the same.

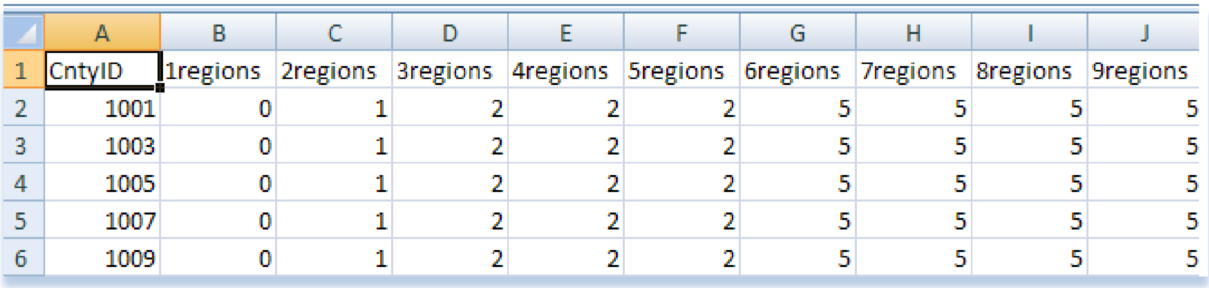
7) The multivariate data are aggregated by adding. In some cases, it may be inappropriate to add the values of units to obtain the value of regions, i.e., average individual income.

3. Region File

Region file stores the region partitions of the study area in different hierarchical levels. An example region file is provided in Figure 3. About the flow file, it shall be noted that:

1) The **first** column must be the IDs of the spatial objects.

2) Each row stores the region IDs of a spatial object in different hierarchical levels. The region IDs must be represented by **integers**. For example, in Figure 3, the region ID of the unit “1001” has a region ID “0” at the “1 region” level, and a region ID “1” at the “2 region” level, and a region ID “5” at the “6 region” level.



	A	B	C	D	E	F	G	H	I	J
1	CntyID	1regions	2regions	3regions	4regions	5regions	6regions	7regions	8regions	9regions
2	1001	0	1	2	2	2	5	5	5	5
3	1003	0	1	2	2	2	5	5	5	5
4	1005	0	1	2	2	2	5	5	5	5
5	1007	0	1	2	2	2	5	5	5	5
6	1009	0	1	2	2	2	5	5	5	5

Figure 3 Region File Structure

3) The region file must include **all** IDs existing in the CSV file. The IDs in the first column must be consistent with those in the CSV and the flow files. In other words, an ID in the region file and the same ID in the CSV file represent the same spatial objects.
4) The order of the origin-destination ID pair does not affect the analysis.
5) If the region file contains IDs that do not exist in the flow file, entries with these IDs are ignored and will not be processed.
6) The region IDs must start from 0 and end in the number one less than the total number of regions for that level. For example, at the "6 regions" level, the region IDs must be 0, 1, 2, 3, 4, or 5.

4. Launch the Software

To run this program, it is recommended to have the latest version of Java installed on your machine. You can visit http://java.com/en/download/index.jsp and test if you have the latest version. Use the following command in a command window to start "**FlowMap**":

```
java -jar -Xmx1024m flowmap.jar
```

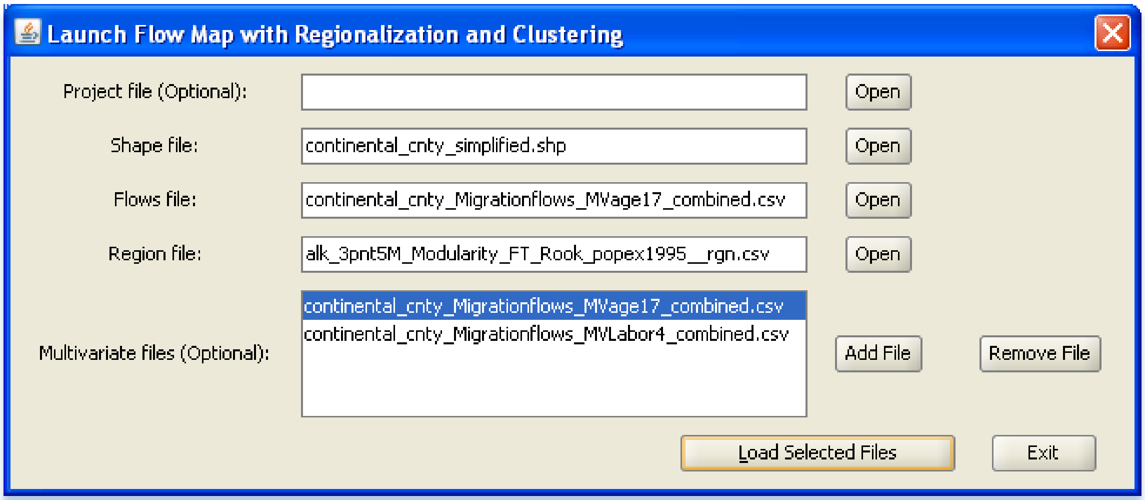
If for some reason, your downloading software actually saves the file as "**.ZIP**", you need to rename the file back to "**.JAR**" before running it as instructed above.

5. Load Data Files

After the package is launched, a data load dialog will appear as shown in Figure 4. There are two options for loading the data: (1) load data files individually, or (2) load a project file.

1. Load Data Files One-by-one

Clicking the "Open" buttons, the shape file, flow file, and region file can be chosen one by one (Figure 4). As mentioned earlier, the flow file can contain multivariate information of flows. If the flow file contains more than three columns, the file is considered containing multivariate information and its name appears in the text box next to the label "Multivariate files (optional)". Additional multivariate files can be added or removed by clicking the button "Add File" and the button "Remove File" in the lower right of the dialog window.



Project file (Optional):

Shape file: continental_cnty_simplified.shp

Flows file: continental_cnty_Migrationflows_MVage17_combined.csv

Region file: alk_3pnt5M_Modularity_FT_Rook_popex1995__rgn.csv

Multivariate files (Optional): continental_cnty_Migrationflows_MVage17_combined.csv
continental_cnty_Migrationflows_MVLabor4_combined.csv

Figure 4 Data Load Dialog: load files separately

2. Quick Start Using Project File

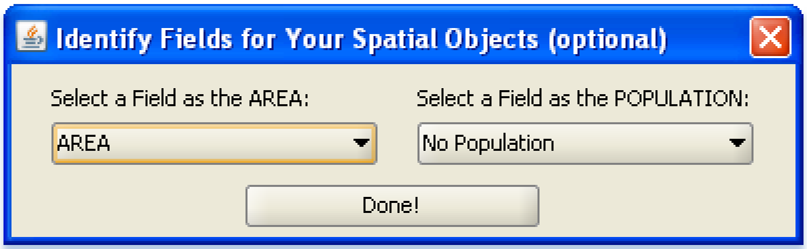
A project file stores the configuration of data files, including the project file path, file names and the field names of AREA and POPULATION (discussed in the next section) in the CSV. Once the project file is loaded, the text boxes for shape file, flow file, region file, and multivariate files are filled automatically. User is allowed to re-select data files after the project file is loaded. The project file must reside in the same folder as the data files.

Possible loading error: If there is no CSV file having the same name with the shape file in the same directory, a shape file cannot be loaded and an error message is displayed.

3. Specify the Area Field and Population Field

After the button “Load Selected Files” in the data load dialog is clicked, another smaller dialog will appear for choosing the fields of area and population of the unit locations (Figure 5). The fields to be chosen from are extracted from the CSV file. The area field and the population field are used to create population density maps at the unit and the region level. The population field is further used to calculate a population-based modularity measure of flow strength (Guo 2009). The area or population field does not have to carry the real population meaning. It can be any field as long as it can provide some useful background to the user. If no area or no population field is chosen, the population density will be uniform at the unit and all region levels. If no population field is chosen, there will be no option of *population-based modularity* as the flow measure.

The file path and file names can be saved in a new project file with a file extension ".pro" after the button "Done!" (Figure 5) is pressed. If a project filed has been loaded and has been changed, the changed project file can be saved as well. Make sure: (1) the project file is saved in the same folder as the data files; (2) the project file name ends with .pro; (3) the data files are in the same folder since the project file only stores one path.



Identify Fields for Your Spatial Objects (optional)

Select a Field as the AREA: AREA

Select a Field as the POPULATION: No Population

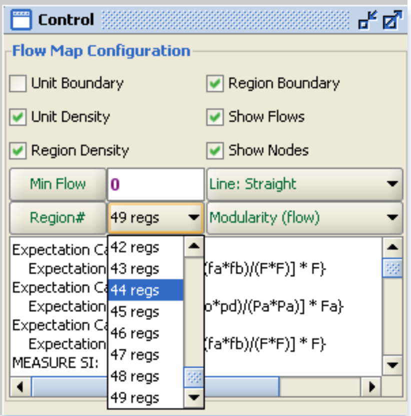
Figure 5 Field Identification Dialog

6. Flow Map

To make this manual easier to follow, this section focuses on flow map without multivariate information. The next section will discuss flow map with multivariate information. Once data loading finishes, the control panel shows up. Figure 6 shows the upper half of this panel, where the main controls of the flow map can be found.

1. Displaying Layers

- **Unit Boundary:** turn on/off the boundaries of the spatial objects in the shape file
- **Unit Density:** turn on/off the population densities of the spatial objects in the shape file
- **Region Boundary:** turn on/off the regions boundaries
- **Region Density:** turn on/off the population density of regions
- **Show Flows:** turn on/off the flow layer (region level)
- **Show Nodes** : turn on/off the flow node layer (region level)



Control

Flow Map Configuration

☐ Unit Boundary ☑ Region Boundary

☑ Unit Density ☑ Show Flows

☑ Region Density ☑ Show Nodes

Min Flow: 0

Line: Straight

Region#: 49 regs

Modularity (flow)

Figure 6 Upper Half of Control Panel

2. Setting Flow Threshold (Minimum Flow)

Flow threshold is used to filter flows. Only those between-region flows above this threshold are displayed. Flow threshold can either be set by typing the value in the Flow Threshold

Textbox (see Figure 6) or using the flow threshold plot (see Figure 7). Flow threshold plot enables setting the threshold visually and it is opened by clicking “Min Flow**”.**

- The plot shows number of links (connections between regions) on the X-Axis and migration population (the total number of migrants between those links) on the Y-Axis.
- The threshold value can be set by moving the x-bar or y-bar on the curve (Figure 7).
- When the bars are moved, the following values are updated:
  - The threshold value at the selected measure,
  - the percentage (migration population above the selected threshold / total migration population) of the migrant population above the selected threshold level,
  - migration population above the threshold (displayed on the y-bar);
  - The number of links/flows above that threshold (displayed on the x-bar).
- When the set external flow threshold button inside the plot is pressed, the links/flows that are above the selected threshold level are shown in the flow map.

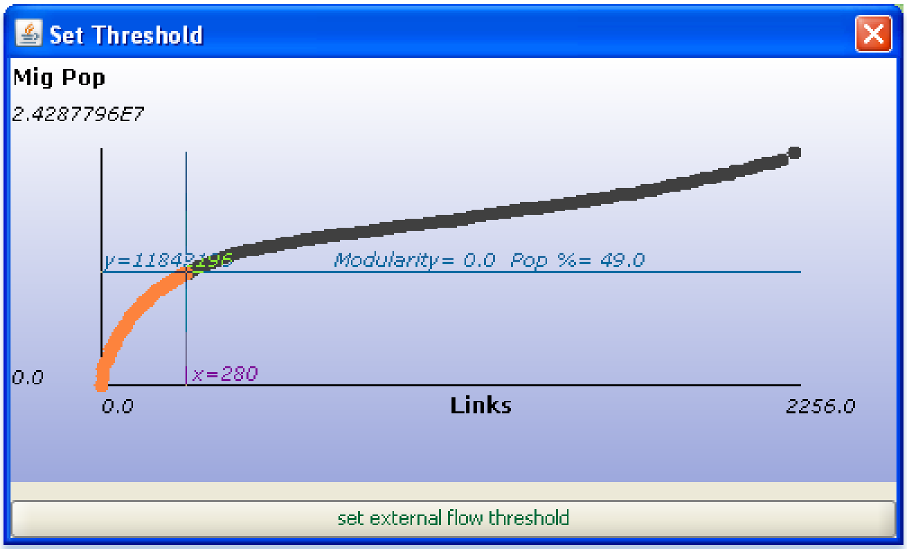


Figure 7 Flow Threshold Plot

3. Selecting Region Levels

This software package allows user to navigate up and down the region hierarchy either by selecting a region level from (1) **Region-Level drop-down list** (next to “Region#” button, see Figure 6) or (2) the region modularity plot (see Figure 8). The region modularity plot is opened by clicking “**Regions Button#**”.

- The plot draws a curve showing the region levels in the region file on the X-Axis and the values of the measure at those region levels on the Y-Axis.

- The value can be set by moving the x-bar or y-bar on the curve (see Figure 8). When the bars are moved, selected region level and the value of the measure at that level are displayed on the y-bar. In Figure 8, this measure is modularity and the modularity value which is calculated at the 39th region level is displayed on the y-bar.
- When button “set internal flow threshold” is pressed the flow map is reproduced at the selected region level.

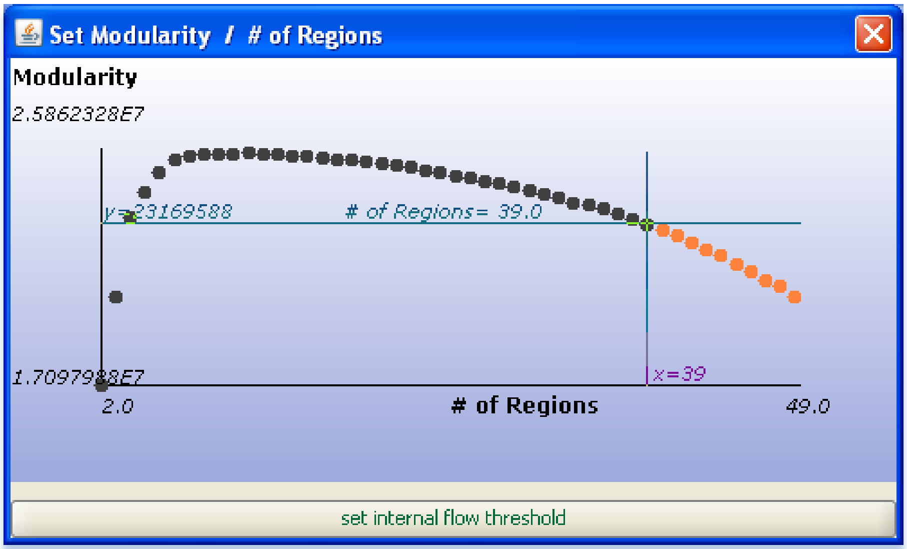


Figure 8 Region Modularity Plot

4. Selecting Flow Line Type

Flow Line type can be set as **straight** or **curve** by using **Flow Line drop-down list**. Flow lines between two locations follow the right-hand traffic rule: pointing to its destination, a flow line is on the right side of the “road”. When arrows are too small and flows are mostly one-way (so that it is difficult to tell which side of the “road” they are on), curve option becomes very convenient. This option draws Bezier curves, where a flow line is curvy at the origin and straight on the destination end.

5. Selecting Method to Display Between Region Flows

Region-to-region flows are displayed by subtracting the actual flows from the expected flows (Guo 2009). Expected flows can be calculated based on flows or populations (if a numeric field chosen as the population of unit spatial objects). If the original flow is chosen the default flow threshold is the average of all region-to-region flows. Otherwise, the default flow threshold is zero.

6. Display Information

The text area in the bottom of the upper half control panel (see Figure 6) provides information on the calculation of flow measures, the data files, and the progress of the analysis. Error messages, if any, are displayed here as well.

7. Flow Map: without Multivariate Information

The flow map displays region-to-region flows with flow lines (straight or curved) and within-region flows with nodes (see Figure 9).

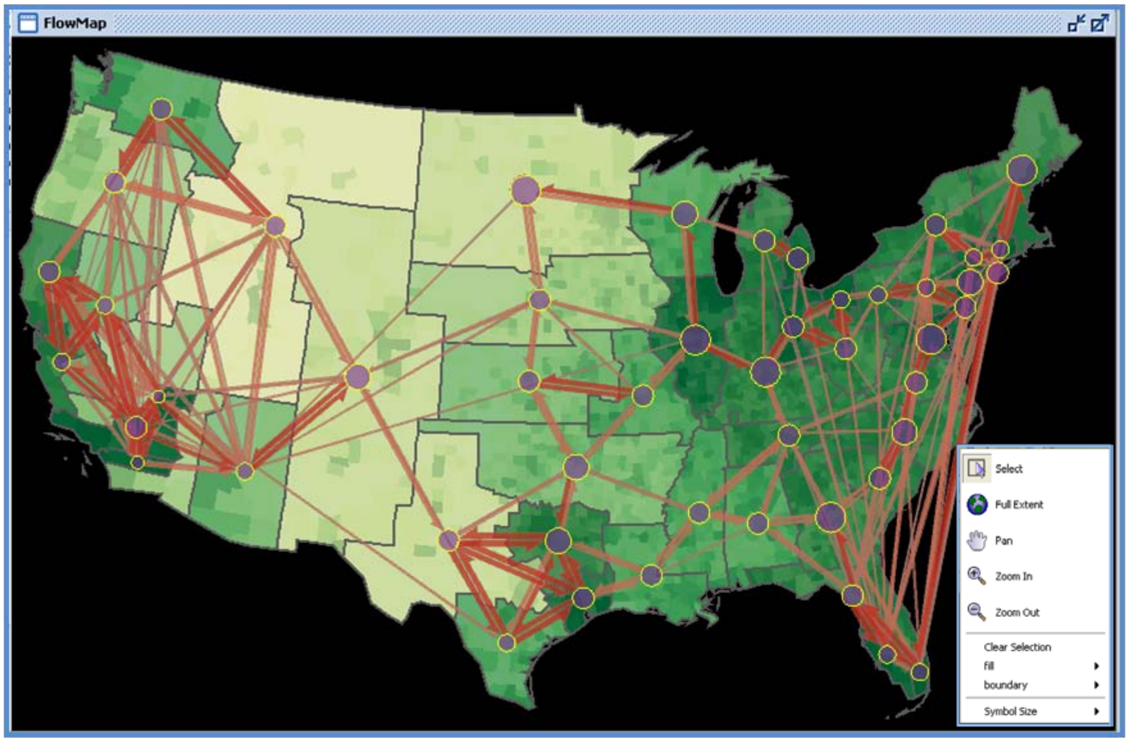


Figure 9 Flow Map without Multivariate Information

The nodes are represented by circle symbols placed at the centroid or population-weighted centroid of each region. The width of flow lines and the radius of nodes are proportional to flow measures formulated with flow strength/volumes. The flow lines share the same color but shaded base on the flow values between regions. The colors of flow nodes are uniform. The county population density and the region population density are shaded in green color to provide some background information. Figure 9 shows that strongest flows occur between nearby regions. Florida has significant long-distance in-migration from the Northeast and the Midwest. The pop-up menu appears once user right-clicks the mouse in the map panel. It can be used to clear selected features, zoom in , and change the width of flow line and the radii of nodes (which are symbol layers in the flow map).

7. Multivariate Flow Map

1. Select and Weight Variables

The first step of multivariate clustering is to select one or more variables. Multivariate flow data loaded by the user are shown in Multivariate Data sub-Panel (see Figure 10). In the example, user can choose the age variables. Once the selected variables are submitted, the normalization and weighting options and the SOM configuration appear in the same sub-panel (Figure 11). In Figure 11, the age variables are normalized by the flow volume to obtain the age composition of flows in percentage.

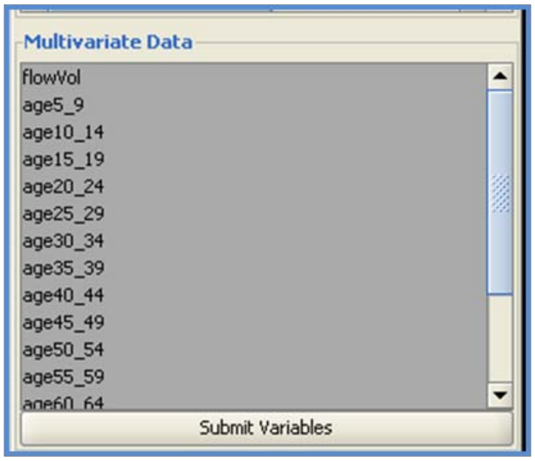
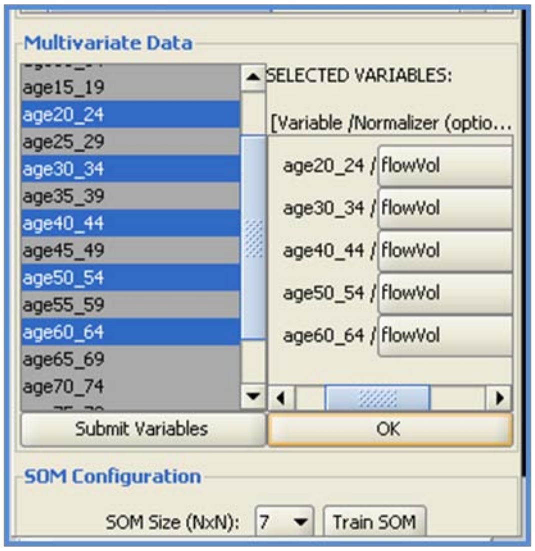


Figure 10 Multivariate Controls



SELECTED VARIABLES:

[Variable /Normalizer (optio...

age20_24 / flowVol

age30_34 / flowVol

age40_44 / flowVol

age50_54 / flowVol

age60_64 / flowVol

SOM Size (NxN): 7

Figure 11 Normalization, Weight, SOM configuration

2. Configuring the SOM Clustering

User can choose the classes of the flow clustering. The default is 49 (7*7). Depending on the data size, one may choose more or fewer. Once the button “Train SOM” is pressed, a multivariate flow map, a self-organizing map (SOM), and a parallel coordinate plot (PCP) are displayed.

3. SOM-Multivariate Flow Clustering

The **SOM** is used to identify clusters in the multivariate flow data and order clusters in a two-dimensional layout (Figure 12). Detailed explanation can be found in relevant publication (Guo, Gahegan et al. 2005).

- Each SOM node (cluster) is represented with a circle, whose size (area) represents the number of flows that it contains. The SOM uses the Euclidean distance to assess multivariate similarity between spatial objects. Therefore, nearby clusters are more similar to each other.

- Behind the nodes (circles) there is the U-matrix layer, where hexagons are shaded to show the multivariate dissimilarity between neighboring nodes, with darker tones representing greater dissimilarity.
- User can rotate or flip the 2D color scheme in case a certain corner is desired to have a certain color. These functions are enabled at the “Color Scheme” tab. The user may also change the 2D color design by going to the “Color Design” tab. The colors are then passed on to PCP and the flow map.

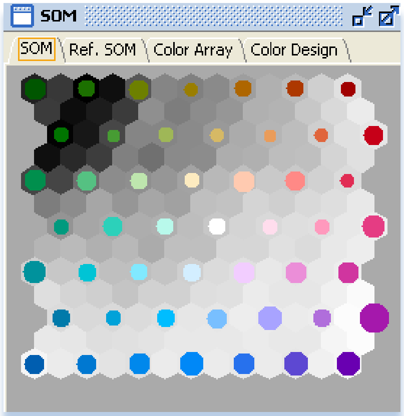


Figure 12 SOM Panel

4. PCP-Legend of Colors

A **PCP** panel (Figure 13) is used to reveal the meaning of each color assigned to each flow class by SOM. The PCP panel provides multiple options for selection mode (normal, intersect, or Union), axis scaling, view mode, and axis ordering (original or optimal).

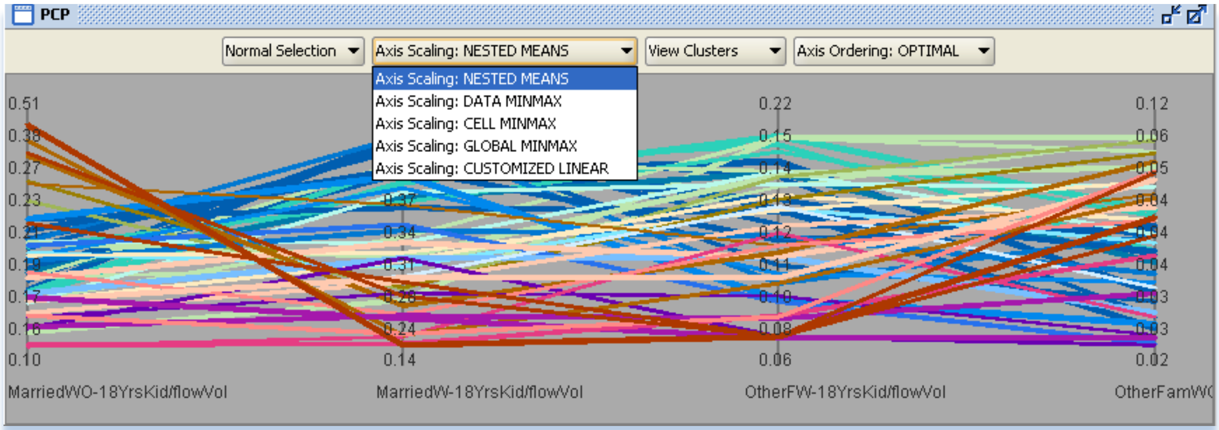

Normal Selection
Axis Scaling: NESTED MEANS
View Clusters
Axis Ordering: OPTIMAL
PCP
Axis Scaling: NESTED MEANS
Axis Scaling: DATA MINMAX
Axis Scaling: CELL MINMAX
Axis Scaling: GLOBAL MINMAX
Axis Scaling: CUSTOMIZED LINEAR
0.51
0.38
0.27
0.23
0.21
0.19
0.17
0.16
0.10
0.37
0.34
0.31
0.28
0.24
0.14
0.22
0.15
0.14
0.13
0.12
0.11
0.10
0.08
0.06
0.12
0.06
0.05
0.04
0.04
0.04
0.03
0.03
0.02
MarriedWO-18YrsKid/flowVol
MarriedW-18YrsKid/flowVol
OtherFW-18YrsKid/flowVol
OtherFamW(

Figure 13 PCP Panel

Selecting Axis Scaling Methods

- **Nested-Means:** scaling on each axis using nested means and thus adjust the spacing of intervals according to data distribution. This method can alleviate the overlapping problem in PCP for skewed data distribution. Specifically, nested-means is a *non-linear scaling* method that recursively calculates a number of mean values (and sub-means) and uses these values as break points to divide each axis into equal-length segments. Therefore, nested-means scaling always puts the mean value at the center of each axis and thus makes axes defined by different units and data ranges comparable.

- **Data Min-Max:** each axis is linearly scaled using its min and max values.
- **Cluster Min-Max** (or Cell Min-Max): each axis is linearly scaled using cluster centroid min and max values.
- **Global Min-Max:** this option is only useful when all the variables are comparable to each other, for example percentage values. Axes will be scaled linearly using the global min and max values (for all variables).
- **Customized Linear:** this option is only useful when all the variables are comparable to each other, for example percentage values. The user will define the min and max (same for all variables) to linearly scale each axis. In future versions, the user may be able to define the min/max differently for each axis.

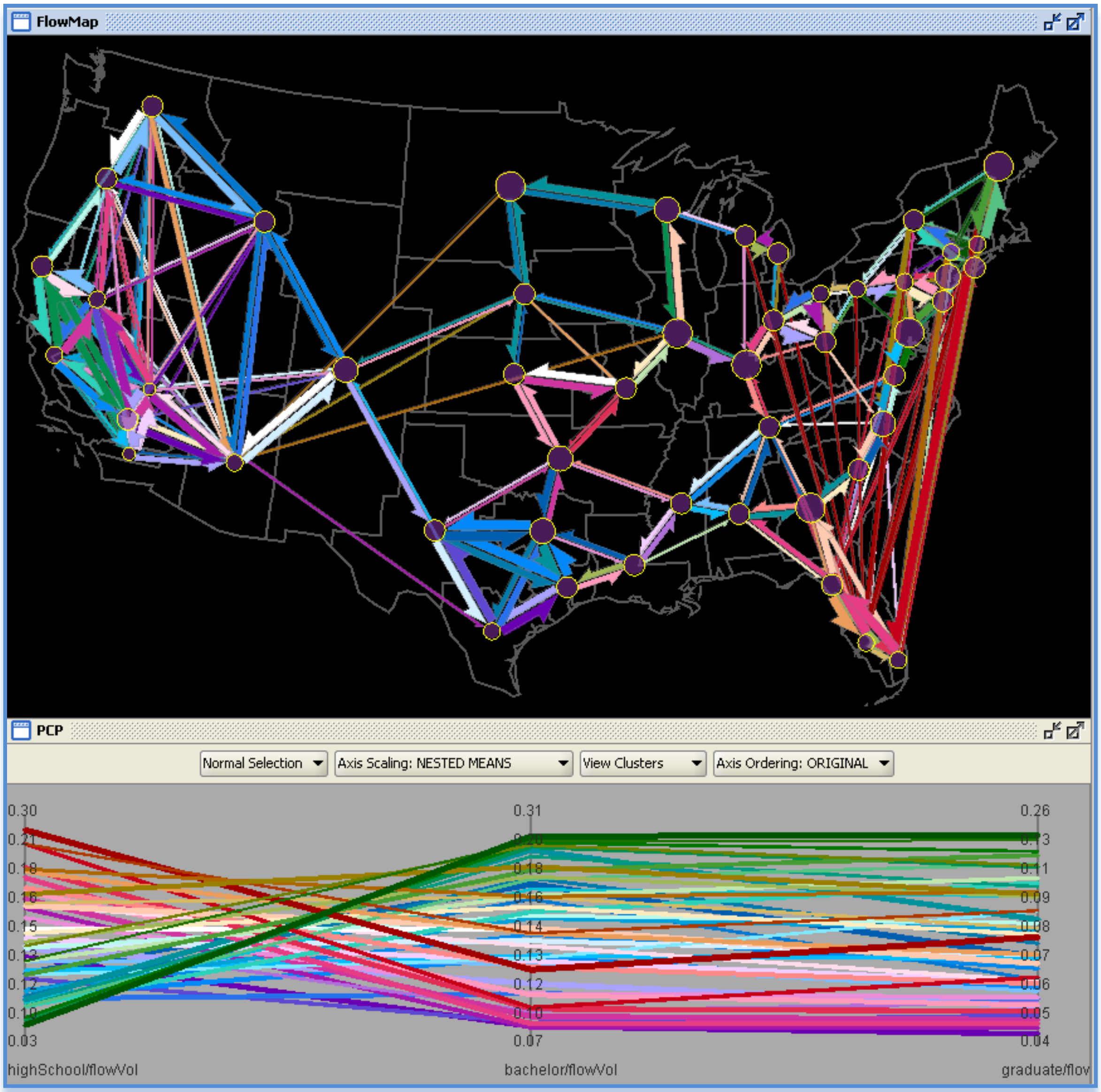
Selecting Detail Levels

- **Cluster Level:** the PCP shows each cluster as a single string, which has the same color as the cluster does in the SOM. The thickness of each string is proportional to the cluster size.
- **Data Item Level**: each string in the PCP will represent an individual region-to-region flow, in the same color of its cluster. With the clusters (and colors) derived by the SOM, and their meaning revealed by the PCP (as the legend), it is straightforward to pass on the colors to the flows in the flow map where it is obvious that there are multivariate spatial patterns of flows.

5. Integrated and Coordinated Panels

The graphic panels (map, SOM, and PCP) in the interface are fully coordinated. Selections and flow colors are consistent among them. This feature enables creative and flexible exploration of spatial flow patterns. Please refer to the SOMVIS Manual for the interpretation of colors in the map when multiple variables are chosen and for the interactive features supported in the map, PCP, and SOM to explore the result.

Figure 14 shows a flow map and the PCP where the flow lines are colored according to the multivariate clustering results. The map shows the education composition of migration lows. To better present the flow structures, the unit boundary, the unit density and the region density are turned off. It is obvious that flows dominated by migrants with high-school education (red color) tend to move long distance while migration flows dominated by migrants with higher education (green color) tend to do short-distance moving.



Normal Selection

Axis Scaling: NESTED MEANS

View Clusters

Axis Ordering: ORIGINAL

Figure 14 Flow Map Panel

Acknowledgements

This work is supported in part by the National Science Foundation under Grant No. 0748813.