REDCAP: A Regionalization Toolkit (Version 2.0.0)

User Manual

Diansheng Guo

Department of Geography
University of South Carolina

www.SpatialDataMining.org

Sept. 01, 2011

1. Overview

REDCAP (**RE**gionalization with **D**ynamically **C**onstrained **A**gglomerative clustering and **P**artitioning) is a toolkit for disease mapping using regionalization method, which is to construct spatially contiguous regions (clusters) based on multivariate similarity (homogeneity). REDCAP provides an integrated environment for the entire process of data loading, algorithm configuration, result visualization, interactive exploration, and result saving.

The software provides four hierarchical regionalization methods, which extend traditional hierarchical clustering methods with spatial contiguity constraints, including the average linkage (ALK), complete linkage (CLK), single linkage SLK, and the WARD's method.

2. Related Publication:

Guo, D. and H. Wang (2011). "Automatic Region Building for Spatial Analysis", Transactions in GIS, vol. 15, issue s1, pp. 29-45.

Guo, D. (2008). "Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)". International Journal of Geographical Information Science. 22(7), pp. 801-823.

Guo, D., J. Chen, A. M. MacEachren, and K. Liao (2006), "A Visualization System for Spatio-Temporal and Multivariate Patterns (VIS-STAMP)", IEEE Transactions on Visualization and Computer Graphics, 12(6), pp. 1461-1474.

Guo, D., M. Gahegan, A.M. MacEachren, and B. Zhou (2005). "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach". Cartography and Geographic Information Science. 32(2), pp. 113-132.

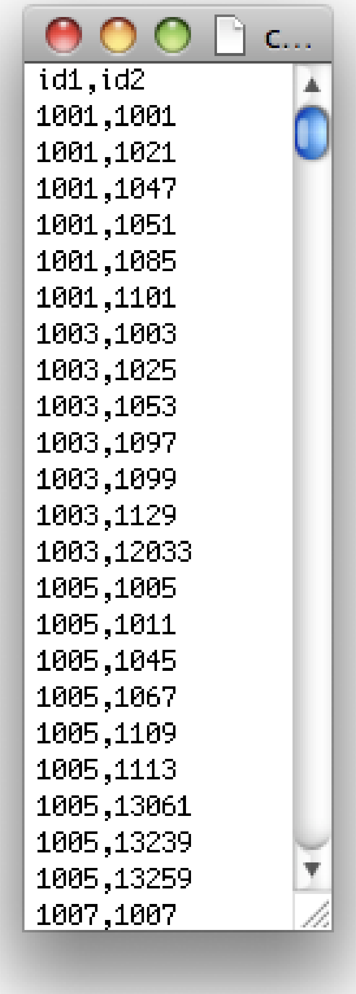
3. Launch REDCAP (from a command window)

Download **redcap.ZIP** file, extract and save it to a local folder of your choice. Open a command window, navigate to the folder chosen above, type in the following command:

```
java -jar -Xmx1024m redcap.jar
```

The option "**-Xmx1024m**" is to allocate 1G RAM memory for java, which is needed to process relatively large data sets. Depending on the data set size, this option can be modified to allow more or less memory. Note: DO NOT double-click redcap.jar to start unless your data set is small (e.g., < 500 items).

Output information is printed in the command window. If somehow REDCAP does not work for your data, you can report to us the error message in the command window so that we know what goes wrong.



```
id1,id2
1001,1001
1001,1021
1001,1047
1001,1051
1001,1085
1001,1101
1003,1003
1003,1025
1003,1053
1003,1097
1003,1099
1003,1129
1003,12033
1005,1005
1005,1011
1005,1045
1005,1067
1005,1109
1005,1113
1005,13061
1005,13239
1005,13259
1007,1007
```

4. Create Contiguity File

To derive spatially contiguous regions, a **contiguity matrix** is needed, which specifies which features are neighbors in space. Such a matrix can be compiled manually or automatically. REDCAP provides functions that can automatically construct a contiguity matrix from a given shape file, with the following steps:

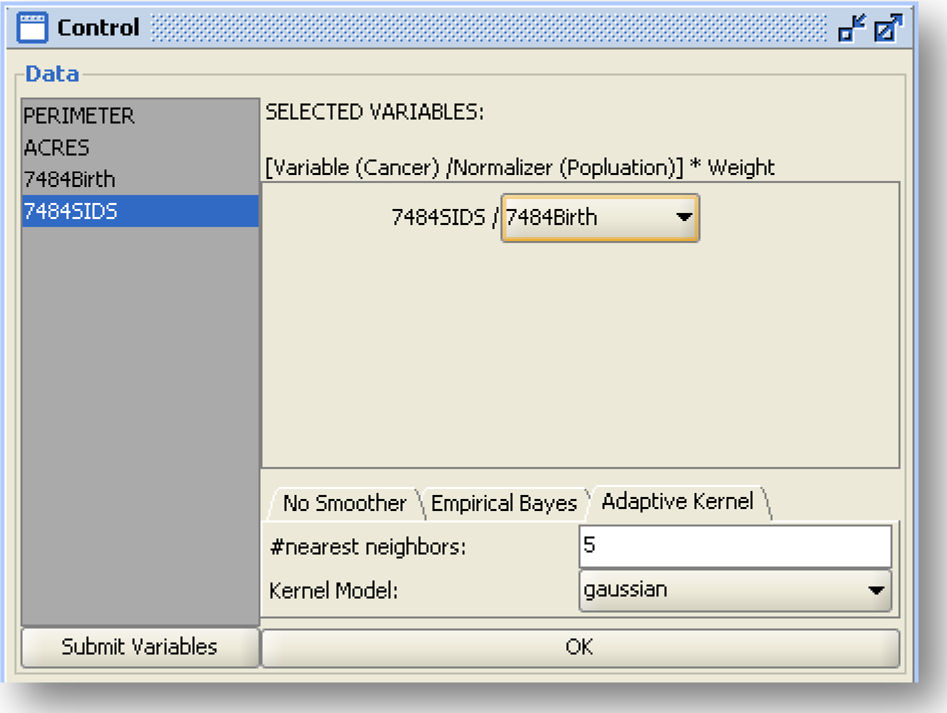
- Contiguity → Create ROOK contiguity …
  (rook is recommended over queen contiguity)
- Choose the shape file (*.shp) of the data set
- Save the contiguity file (*.ctg)

The contiguity file is in a CSV file format but with an extension of "*.ctg". The contiguity file can be opened in any text editor for editing. Each row in the contiguity file only has at two fields: the IDs of two shapes that are neighbors (see the snapshot to the right). Shape IDs are from the first column in the attribute table (*.csv).

The contiguity matrix must be a connected graph. In other words, all objects must be connected and there should not be any island. The contiguity has no direction, i.e., if A is a neighbor of B, then B is a neighbor of A. The automatic algorithm will check the connectivity and add more links if necessary to force islands connected to other components.

If you create a contiguity matrix manually, it is highly recommended that you use the "Contiguity → Check …" function to check whether the graph is connected or not. If it is not, the program will print out (in the command window) the objects in each component and then you should manually add more entries to the contiguity file to make the graph connected.

5. Load Data and Configure Variables



PERIMETER
ACRES
7484Birth
7484SIDS

SELECTED VARIABLES:

[Variable (Cancer) /Normalizer (Popluation)] * Weight

7484SIDS / 7484Birth

#nearest neighbors: 5

Kernel Model: gaussian

File → Load Data …

    Choose the shape file.
    Click Open.
    Choose the contiguity file created in step 4.
    Click Open.

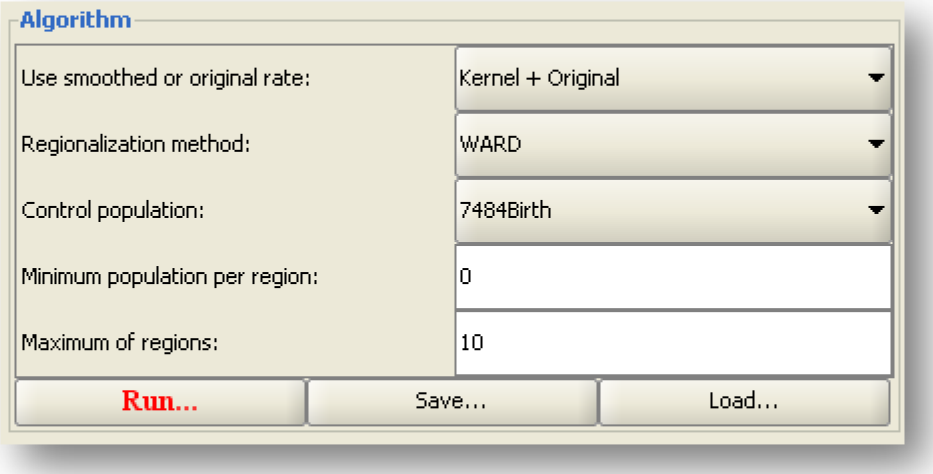
Select variable(s) in the list → Submit Variables

Set weights for variables (if there are more than one variable)

- Each variable should be normalized by another variable.
- Each variable (after the normalization if any) will be transformed to a Z-score (i.e., normalized to unit variance and zero mean). Then its weight will be multiplied.
- Choose a smoother.
    - For Empirical Bayes smoother, minimum population could be used to control the "neighbors" which are used to smooth data. The default value of minimum population is 0, which means only first-order neighbors will be used to smooth data.
    - For Adaptive kernel smoother, two parameters could be set. One is the number of nearest neighbors used to smooth data. Another is model type.

- Click "OK". (Tab color of selected smoother would be set as RED.)

6. Configure Regionalization Algorithm



Algorithm

Use smoothed or original rate: Kernel + Original

Regionalization method: WARD

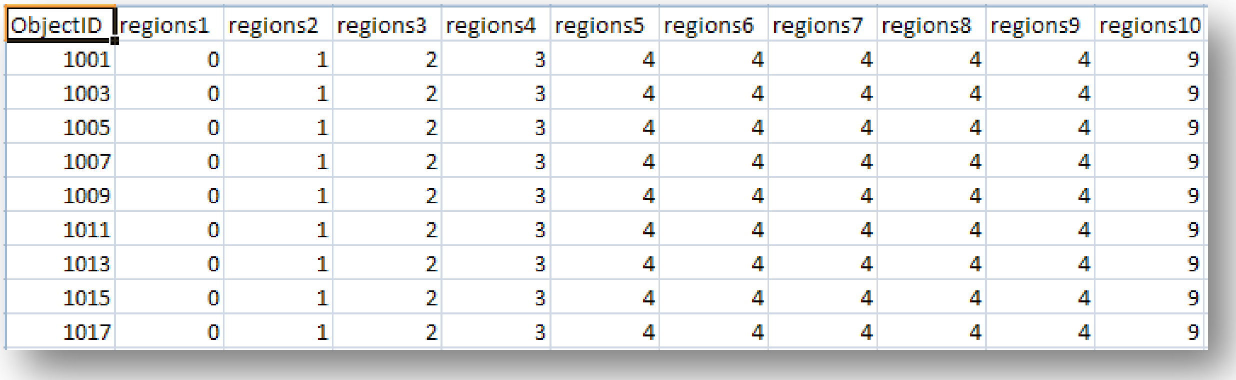
Control population: 7484Birth

Minimum population per region: 0

Maximum of regions: 10

- Choose the rate you want use
    - "Smoother" + Original: use smoothed rate to construct the hierarchical tree and original rate to partition the tree;
    - "Smoother" + "Smoother": use smoothed rate to construct and partition the tree;
    - Original + Original: use original rate to construct and partition the tree;
- Choose a regionalization method (ALK is recommended)
- In addition to the contiguity constraint, another constraint can be configured:
    - You may choose a control variable (e.g., the number of features or the population of each region)
    - Set a control population threshold (e.g., # of features in a region > 3, or region population >10,000,000).

    - If no additional constraint is needed, use “# features” and “0” for the above two fields)
- Set the maximum number of regions (the program will generate at most that many regions)
    - E.g., 10, which means the method will produce a hierarchy of 10 regions.
- Click the “Run…” button and wait until it finishes.
- Save the regionalization result
    - The saved result (which is a csv file) has the same number of rows (following the same order) as that of the input data (see Section 2). The first column, “ObjectID” are from the first column in the attribute table (*.csv).
    - Each column (regions1, regions2, etc.) has the region id (ranging from 0 to k-1, where k is the number of regions at that level) for each feature at that hierarchical level. For example, “regions1” column has only one value (zero) since there is only one region (the entire map), “regions2” column has two values: 0 and 1, and so on.

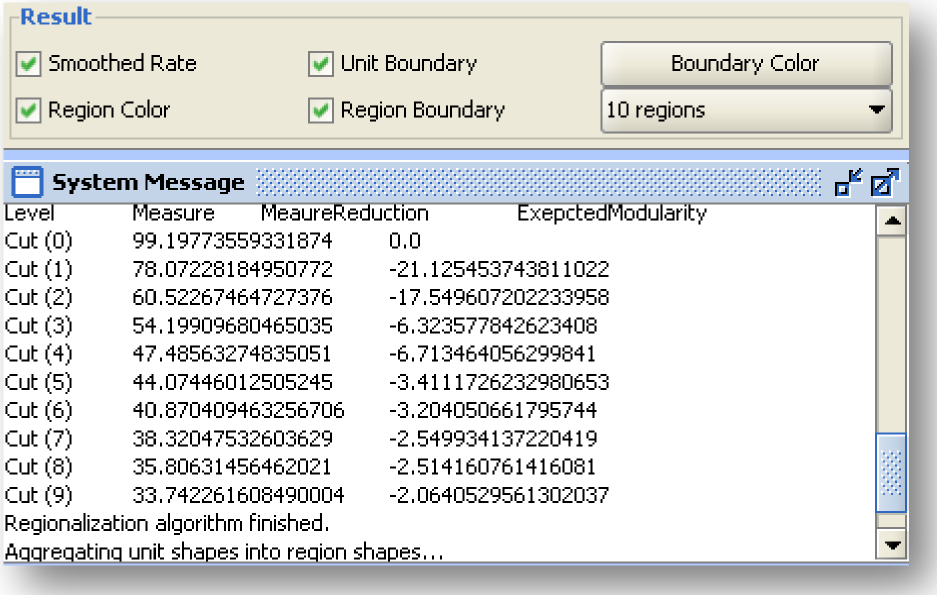


ObjectID	regions1	regions2	regions3	regions4	regions5	regions6	regions7	regions8	regions9	regions10
1001	0	1	2	3	4	4	4	4	4	9
1003	0	1	2	3	4	4	4	4	4	9
1005	0	1	2	3	4	4	4	4	4	9
1007	0	1	2	3	4	4	4	4	4	9
1009	0	1	2	3	4	4	4	4	4	9
1011	0	1	2	3	4	4	4	4	4	9
1013	0	1	2	3	4	4	4	4	4	9
1015	0	1	2	3	4	4	4	4	4	9
1017	0	1	2	3	4	4	4	4	4	9

7. Explore Regions

In the System Message window (see below), the homogeneity (i.e., SSD measure) is given for each hierarchical level. SSD for Cut (1) is 78.07. The smaller SSD is, the more homogeneous each region is on average. One may copy and paste the SSD values to Excel and draw a line plot. If more than one methods are tried, one can compare their performance by overlaying their SSD curves in the same plot.

$$SSD = \frac{1}{d}\sum_{j\,=\,1}^{d}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

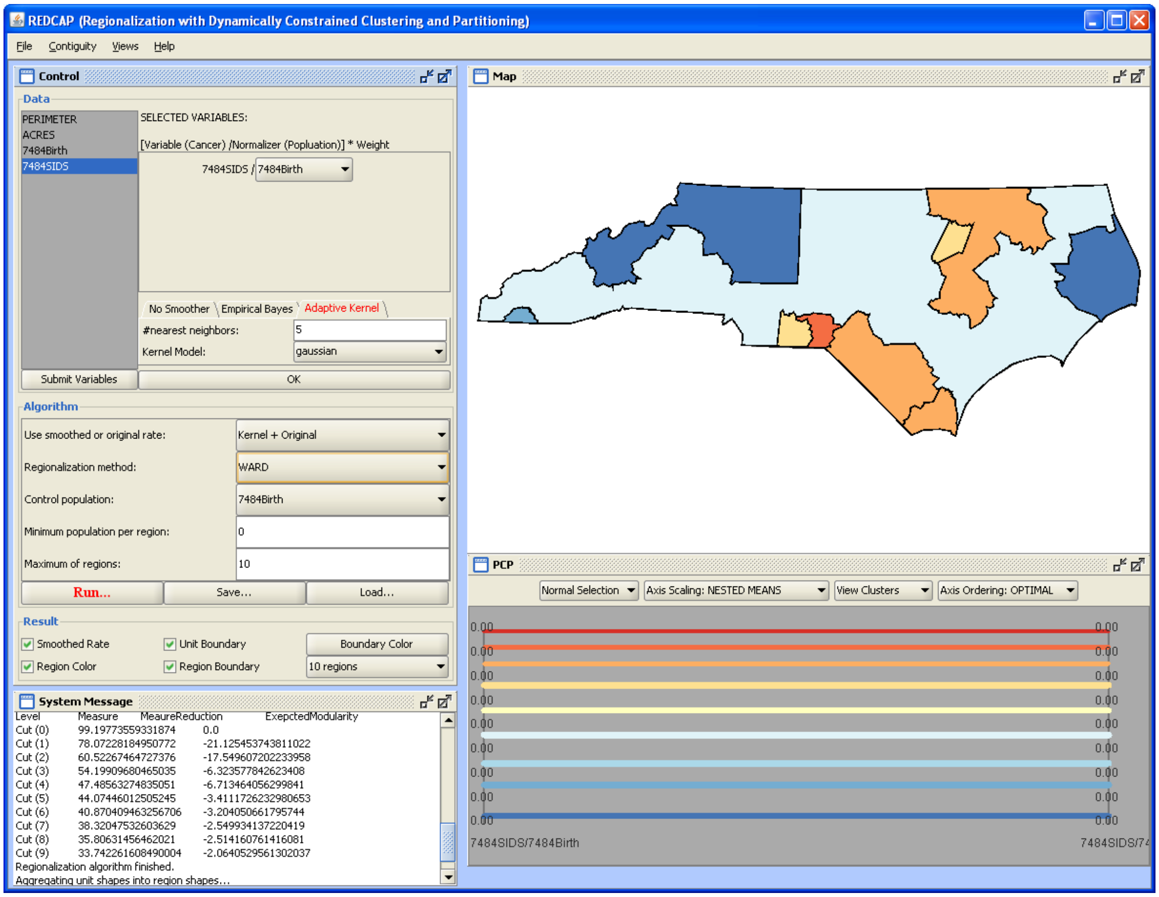


☑ Smoothed Rate ☑ Unit Boundary
☑ Region Color ☑ Region Boundary
10 regions

```
Level          Measure        MeaureReduction              ExepctedModularity
Cut (0)        99.19773559331874       0.0
Cut (1)        78.07228184950772       -21.125453743811022
Cut (2)        60.52267464727376       -17.549607202233958
Cut (3)        54.19909680465035       -6.323577842623408
Cut (4)        47.48563274835051       -6.713464056299841
Cut (5)        44.07446012505245       -3.4111726232980653
Cut (6)        40.870409463256706      -3.204050661795744
Cut (7)        38.32047532603629       -2.549934137220419
Cut (8)        35.80631456462021       -2.514160761416081
Cut (9)        33.742261608490004      -2.0640529561302037
Regionalization algorithm finished.
Aggregating unit shapes into region shapes...
```

The boundaries of each unit (e.g., counties) and regions and the filled color of each region can be turned on/off if needed. One can change the region boundary color. One can also change the number of regions to be shown (i.e., choose the hierarchical level). One can also choose the smoothed rate or original rate on the map.

Following diagram shows the overview of REDCAP interface. 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.



REDCAP (Regionalization with Dynamically Constrained Clustering and Partitioning)

SELECTED VARIABLES:

[Variable (Cancer) /Normalizer (Popluation)] * Weight

7484SIDS / 7484Birth

No Smoother | Empirical Bayes | Adaptive Kernel

#nearest neighbors: 5

Kernel Model: gaussian

Algorithm

Use smoothed or original rate: Kernel + Original

Regionalization method: WARD

Control population: 7484Birth

Minimum population per region: 0

Maximum of regions: 10

Result

☑ Smoothed Rate ☑ Unit Boundary

☑ Region Color ☑ Region Boundary

10 regions

System Message

```
Level       Measure            MeaureReduction         ExepctedModularity
Cut (0)     99.19773559331874     0.0
Cut (1)     78.07228184950772     -21.125453743811022
Cut (2)     60.52267464727376     -17.549607202233958
Cut (3)     54.19909680465035     -6.323577842623408
Cut (4)     47.48563274835051     -6.713464056299841
Cut (5)     44.07446012505245     -3.4111726232980653
Cut (6)     40.870409463256706    -3.204050661795744
Cut (7)     38.32047532603629     -2.549934137220419
Cut (8)     35.80631456462021     -2.514160761416081
Cut (9)     33.742261608490004    -2.0640529561302037
Regionalization algorithm finished.
Aggregating unit shapes into region shapes...
```

Appendix

Algorithm Configuration

Use smoothed or original rate: select different rate combination to build and partition tree.

Regionalization Method: select a method from the four options. For details about the methods, see (Guo 2008).

Control Population: select the control variable and specify the minimal value of the control variable per region.

Maximum of Regions: the number of regions to be derived. Since the regionalization is a hierarchical process, if you specify 10 regions, automatically you also get 2, 3, .., 9 regions.

Run: start the regionalization process. For 3111 counties, it may take a minute to finish.

Save: save the regionalization result to a CSV file (whose name ends with "_rgn.csv"). This file can be joined back to the shape file in ArcGIS and create new shapes for regions (and aggregate attribute values).

Result Exploration

Smoothed Rate: turn on/off the smoothed rate for each unit.

Unit Boundary: turn on / off the unit boundary (e.g., county boundaries) to better view region boundaries.

Region Color: turn on / off the region color.

Region Boundary: turn on / off the region boundaries

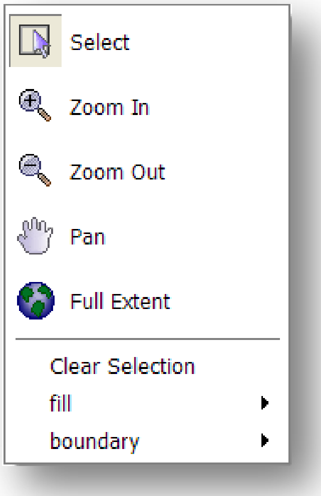
Boundary Color: change the color of region boundaries

Number of regions to show: this drop down list allows the selection of different number of regions to be displayed in the Map. Once this list is selected, one can use the UP and DOWN arrows to "animate" the hierarchy.

Map Panel

It presents the base map and the regionalization result. One can select data items in the map. Right click on the map, and in the popup menu, you can do:

Map zoom in: drag mouse to draw a rectangle, or click the desired center point
Map zoom out: map will be shrinked and centered on that point.
Map full extent
Map panning: mouse drag in the map
Map selection: drag mouse to draw a rectangle --> all features that intersect the rect will be selected.
Map selection (add): with SHIFT key down, make another selection, which will be added to the current selection.
Map selection (subtract): with SHIFT key down, selecting an already-selected feature will de-select that feature.
You can also use the mouse wheel to zoom in/out.



SOM Panel

The Self-Organizing Map is only used for mapping multivariate data (i.e., multiple variables are used). For details please refer to the reference (Guo, et al. 2005, 2006).

PCP Panel (An interactive "map legend")

Selection mode: chooses the selection mode (normal, intersection, union, or indication)

Axis scaling: choose the scaling scheme of axes (nested means, data minmax, cell minmax, global minmax, or customized linear)

View: view clusters or data items.

Axis ordering: choose the ordering of axes (optimal or original). Only useful for multivariate data