Grouping and Regionalization
(Advanced Clustering)

Supplemental Lecture | GEOG 510
GIS & Spatial Analysis in Public Health

Varun Goel

Outline

- Grouping
  - Why we group observations
  - Methods
    - Statistical (hierarchical, partitional)
    - Spatial
    - Hybrid
  - Choosing the number of groups

Clustering?

- What is clustering?
  - Clustering
    - Identifying whether events/values are clustered in space
  - Cluster detection
    - Identifying clusters of events/values in space (deviations from expected)
  - Grouping
    - Assigning observations into groups or clusters

Grouping

- Process of assigning observations into groups
  - Input: individual observations
  - Output: group membership of individual observations
  - Purpose: Organize data into sensible groupings
    - Based on measured or perceived similarity among observations

Regionalization

- Grouping process using spatial observations (e.g., small areas)
  - Assign observations to a region
  - Aggregate observation data to the "group" level
  - Creates larger geographic regions as the observation unit

Grouping Goals

- Groups should be internally cohesive and externally isolated
  - Homogeneity within groups, heterogeneity between groups
    - Observations that are similar should be in the same group and groups should be different from one another

Grouping Goals

- Output groups should have relevance
  - Method/technique should match the desired goal of the grouping analysis
    - Analyst should have a firm understanding of both the input data and output groups
      - Know what the input data represents… and know what the method produces!
      - You should be able to answer the question, "why are these observations grouped together?"

Reasons for Grouping

- Rate stabilization
  - For health data in small areas
    - Highly resolved data are great, but can lead to instability in rate calculations
- Data reduction / consolidation
  - Organization and processing
- Data exploration
  - Which observations are similar?
    - Why are these similar?

Grouping

- General approach
  - Develop conceptual model of "similarity" among observations
    - Likely, related to the output desired
  - Determine attributes or characteristics that express conceptual model of similarity
    - Choose variables that represent the desired characteristics

Grouping

- General approach (cont.)
  - Select grouping method or technique
    - Should match the conceptual model of similarity among observations
  - Choose number of groups
    - Can be a difficult process
      - Sometimes, decision is made prior to running the grouping analysis
  - Evaluate output groups
    - Do they match the conceptual model?

Grouping Methods

- Categories
  - Statistical
    - Similarity evaluated only by the "attributes" of the observations
      - e.g., group is a set of observations with similar sociodemographic characteristics
  - Spatial
    - Similarity evaluated only by spatial or topological relationships
      - e.g., group is a set of observations that are located near to one another

Grouping Methods

- Categories
  - Rule-based
    - Similarity based on a set of decision rules
      - e.g., group is a set of observations with more than 25% of some characteristic
  - Hybrid
    - Similarity based on a mixture of both spatial and statistical characteristics
      - e.g., group is a set of observations with similar sociodemographic characteristics and are located near to each other

Statistical Grouping

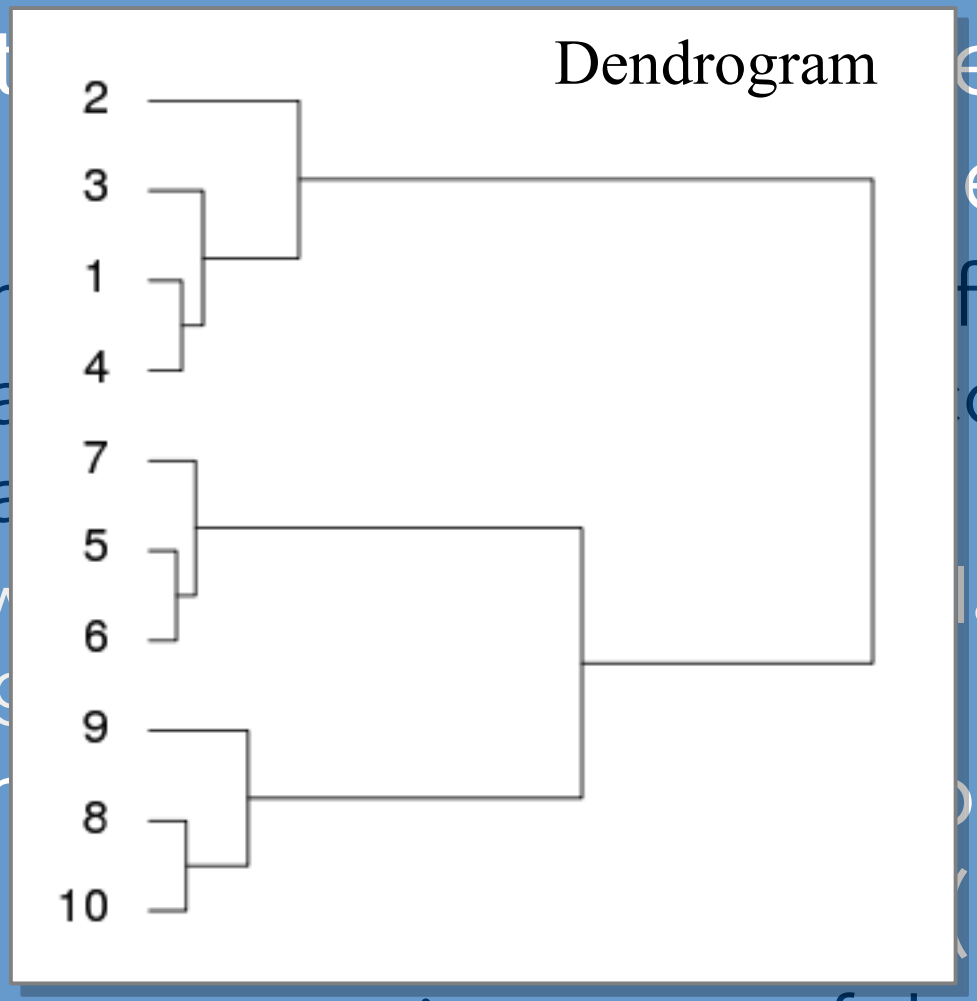
- Statistical grouping only considers the attributes of the observations
  - Does not explicitly address "space"
    - However, location as an attribute
- Hierarchical
  - Step-by-step method (iterative)
- Partitional
  - Heuristic driven
    - Attempts to find "optimal" solution for desired number of groups

Hierarchical Grouping

- Observations are grouped iteratively until one large group is formed
  - e.g., start with 20 observations, first step creates a 19 group solution, second step creates a 18 group solution, ...
    - Stops when all observations are placed into a single group
- Output includes a solution for all potential number of groups (*k = 2:n-1*)
  - Thus, user can retrieve any of the solutions

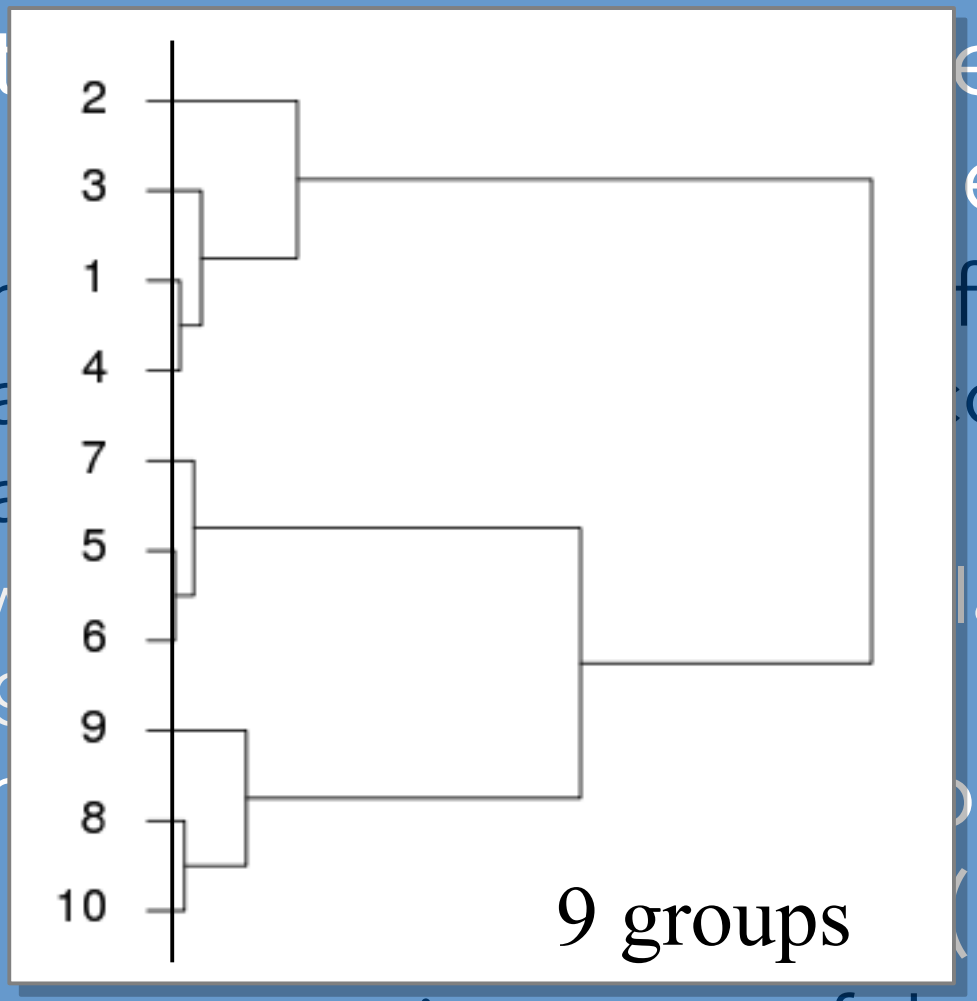
Hierarchical Grouping

- Observat… …eratively until one … …ed
  - e.g., star… …first step creates a… …ond step creates a…
    - Stops w… …laced into a single g…
- Output ir… …r all potential … *(k = 2:n-1)*
  - Thus, user can retrieve any of the solutions



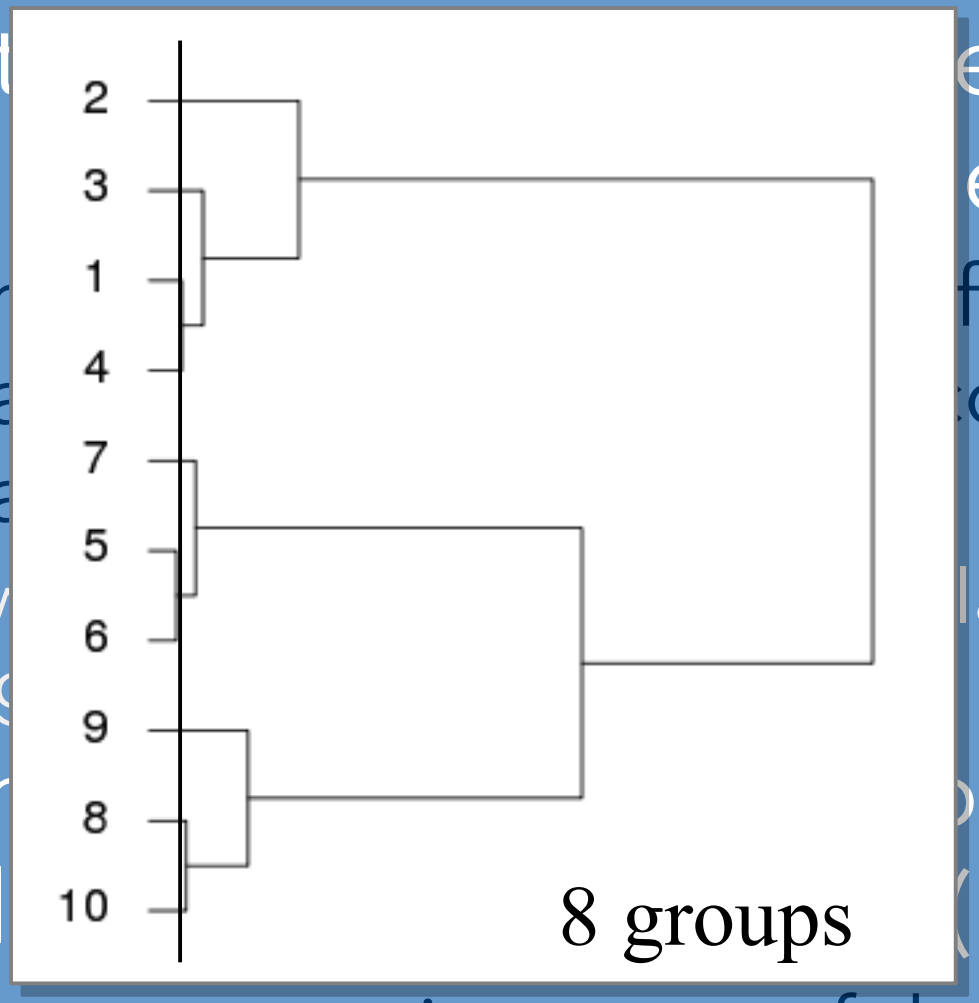
Hierarchical Grouping

- Observat… …eratively until one … …ed
  - e.g., star… …first step creates a… …cond step creates a…
    - Stops w… …laced into a single g…
- Output ir… …r all potential… …(*k = 2:n-1*)
  - Thus, user can retrieve any of the solutions



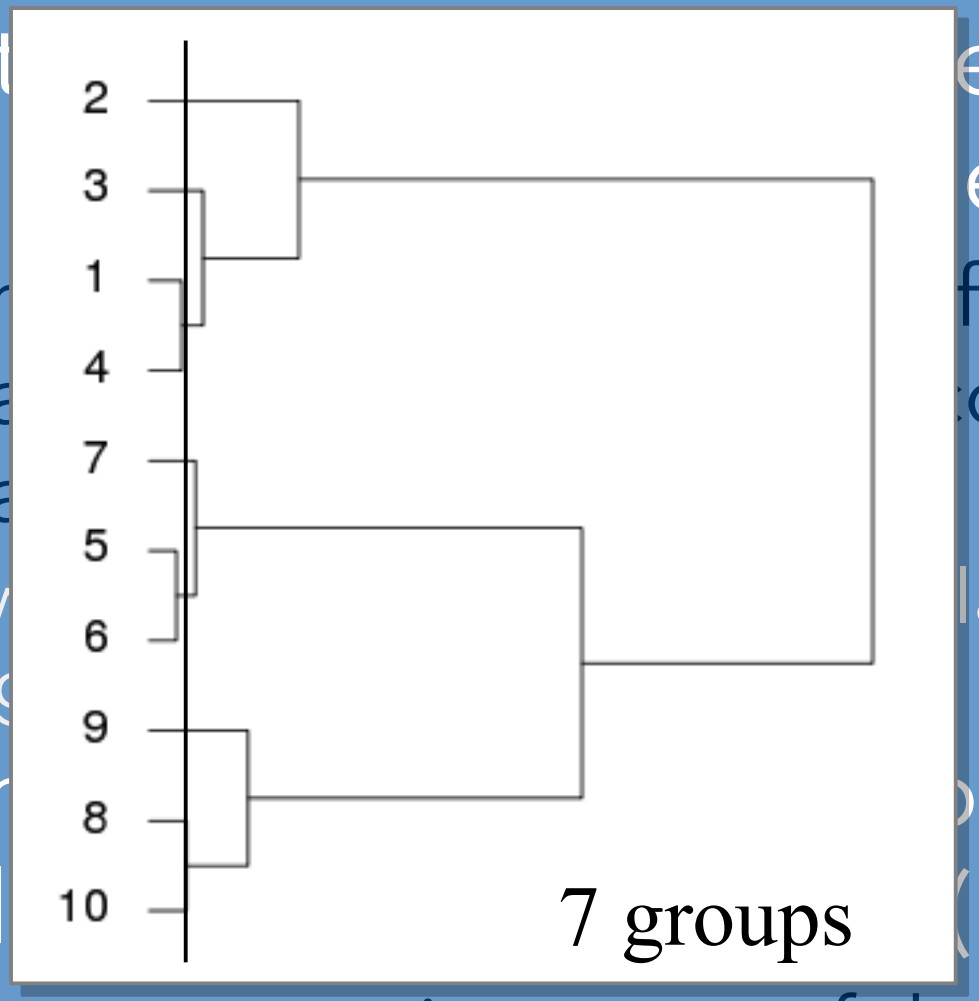
Hierarchical Grouping

- Observat eratively until one ed
  - e.g., star first step creates a ond step creates a
    - Stops v laced into a single g
- Output ir r all potential (*k = 2:n-1*)
  - Thus, user can retrieve any of the solutions



Hierarchical Grouping

- Observat eratively until one ed
  - e.g., star first step creates a ond step creates a
    - Stops v laced into a single g
- Output ir r all potential (*k = 2:n-1*)
  - Thus, user can retrieve any of the solutions

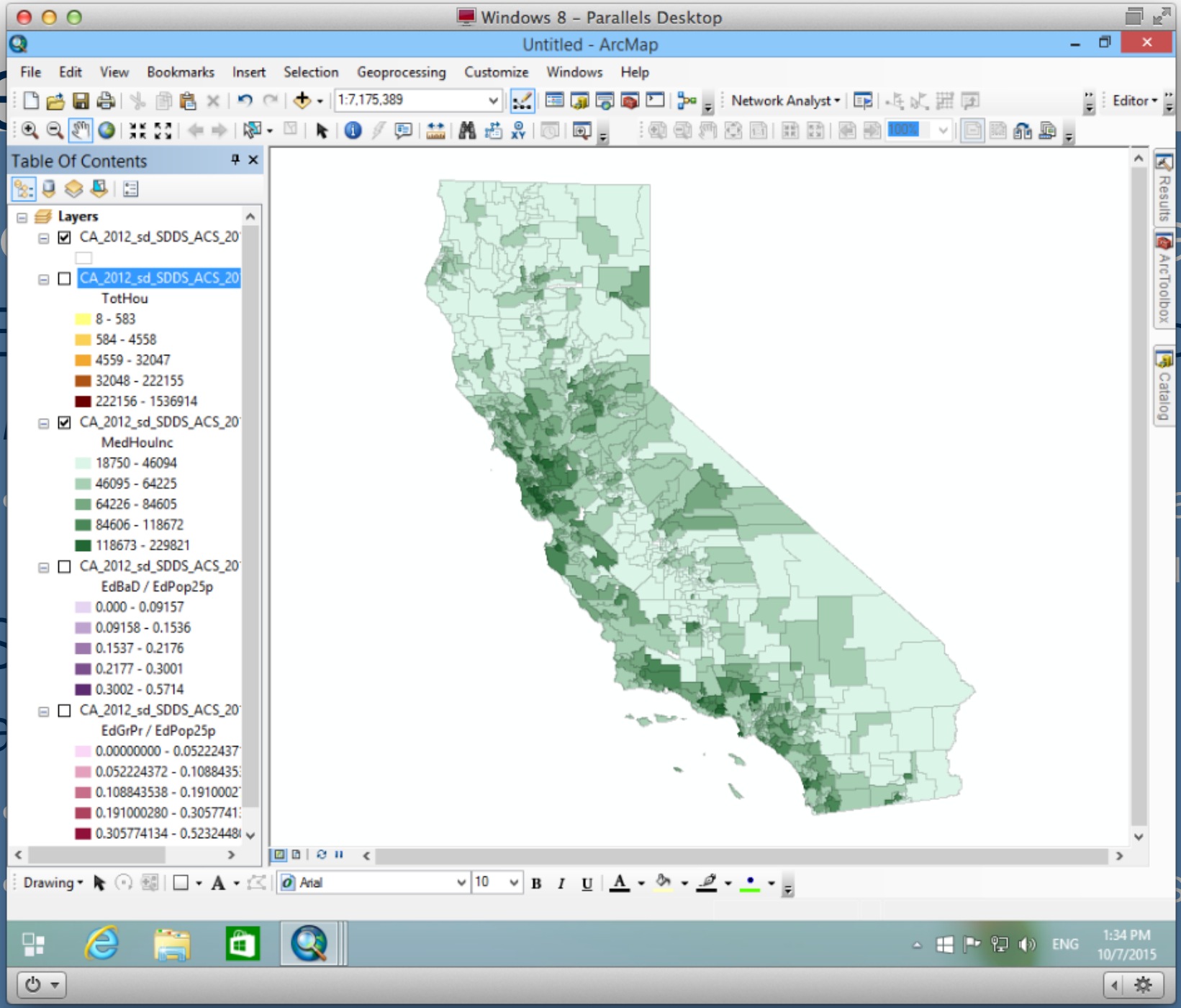


Hierarchical Grouping

- Works in *m*-dimensional data space
  - Each observation may contain multiple (*m*) attributes
    - Thinking about attribute tables in vector data, a row is an observation and a field is an attribute
  - Similarity is evaluated across all *m* attributes that are selected
    - This can be one… two… all the way to *m*
    - Important to understand "why" the attributes are chosen (link back to model of similarity)

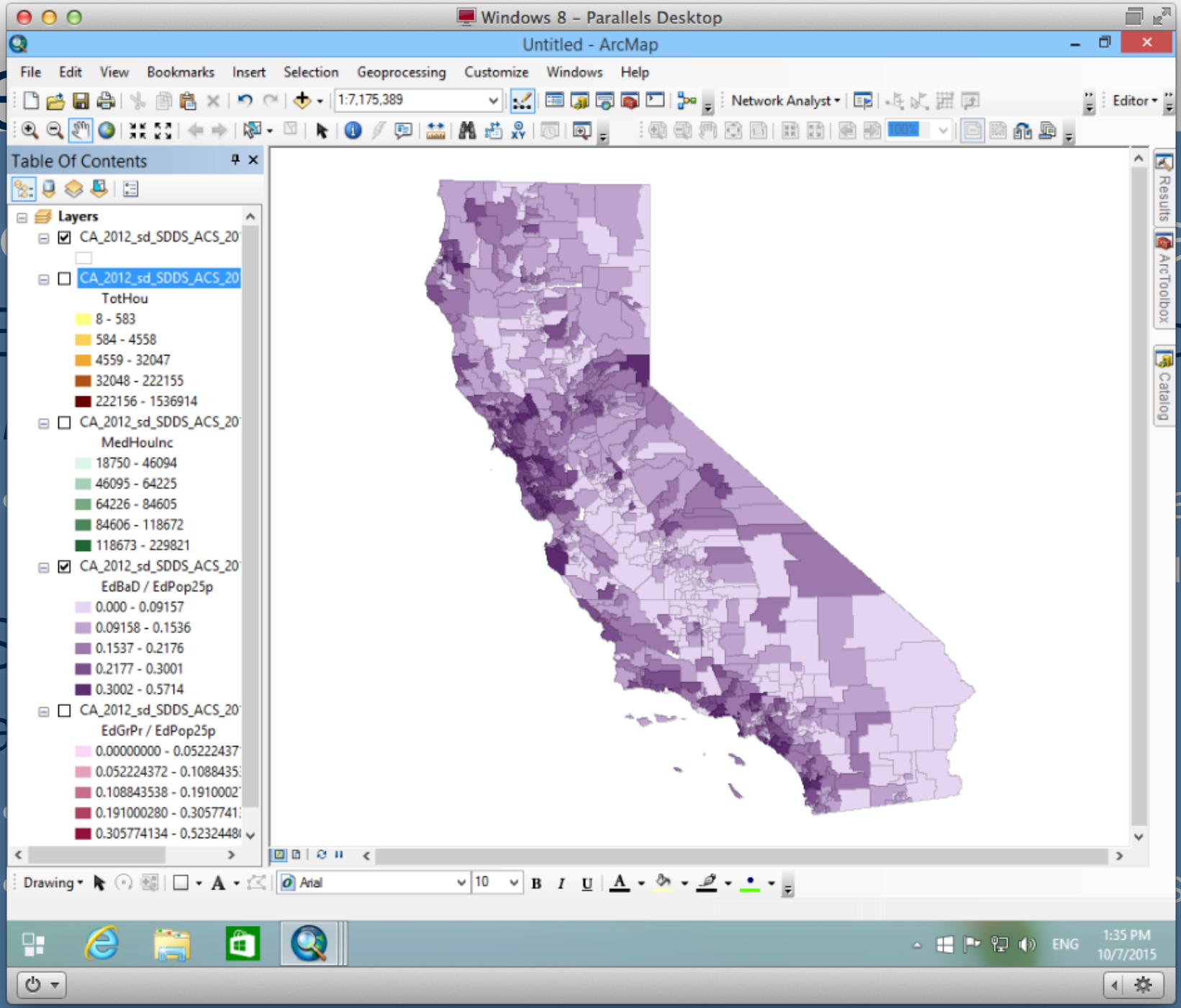
Hie

- W ace
  - E ble
  (A
  - S a, a
  a ite
  - s are



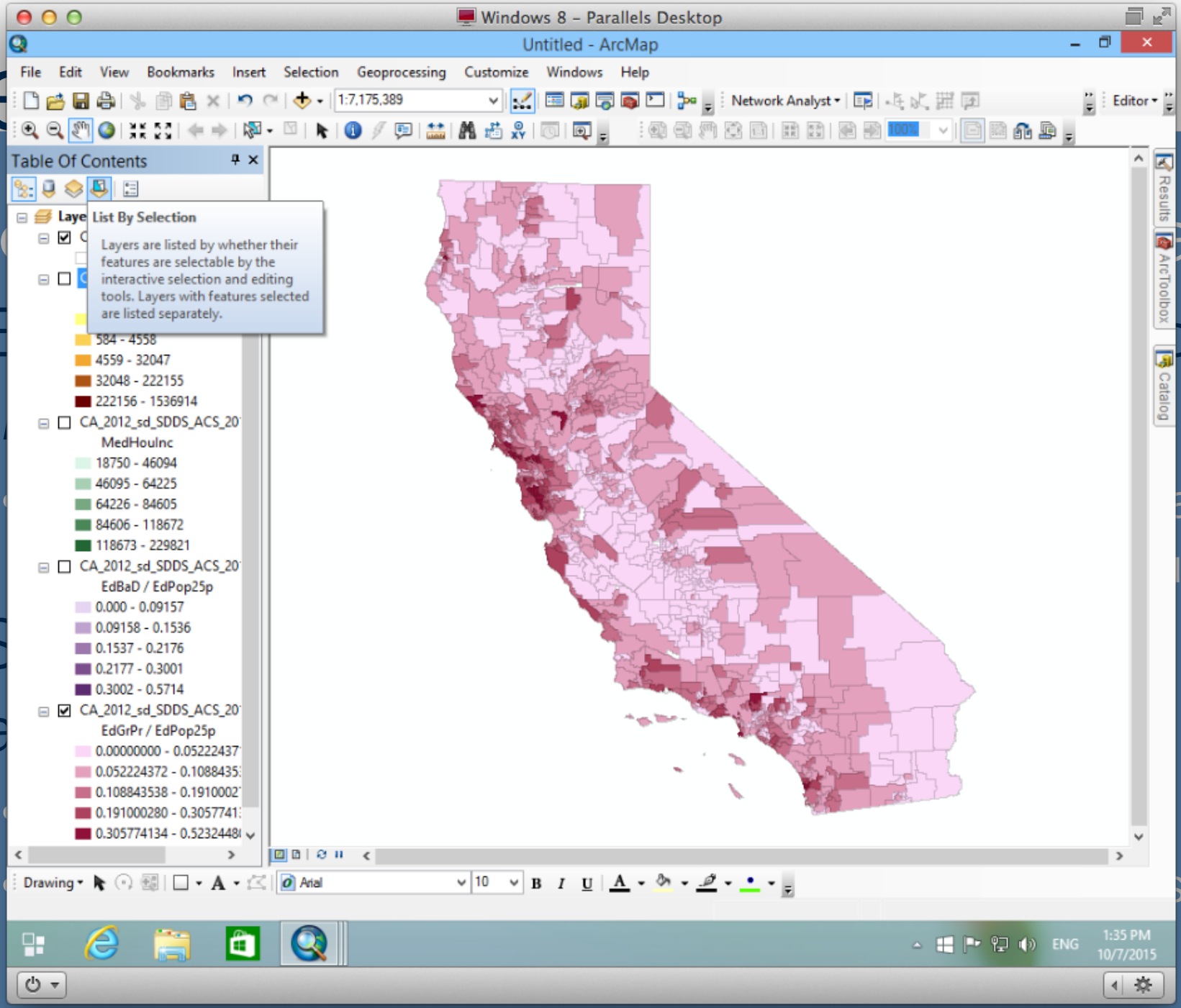
Hie

- W ace
  - E ble
  (A
  - S a, a
  a ite
  - s are



Hie

- W ace
  - E ble
  (A
  - S a, a
  a ite
  - s are

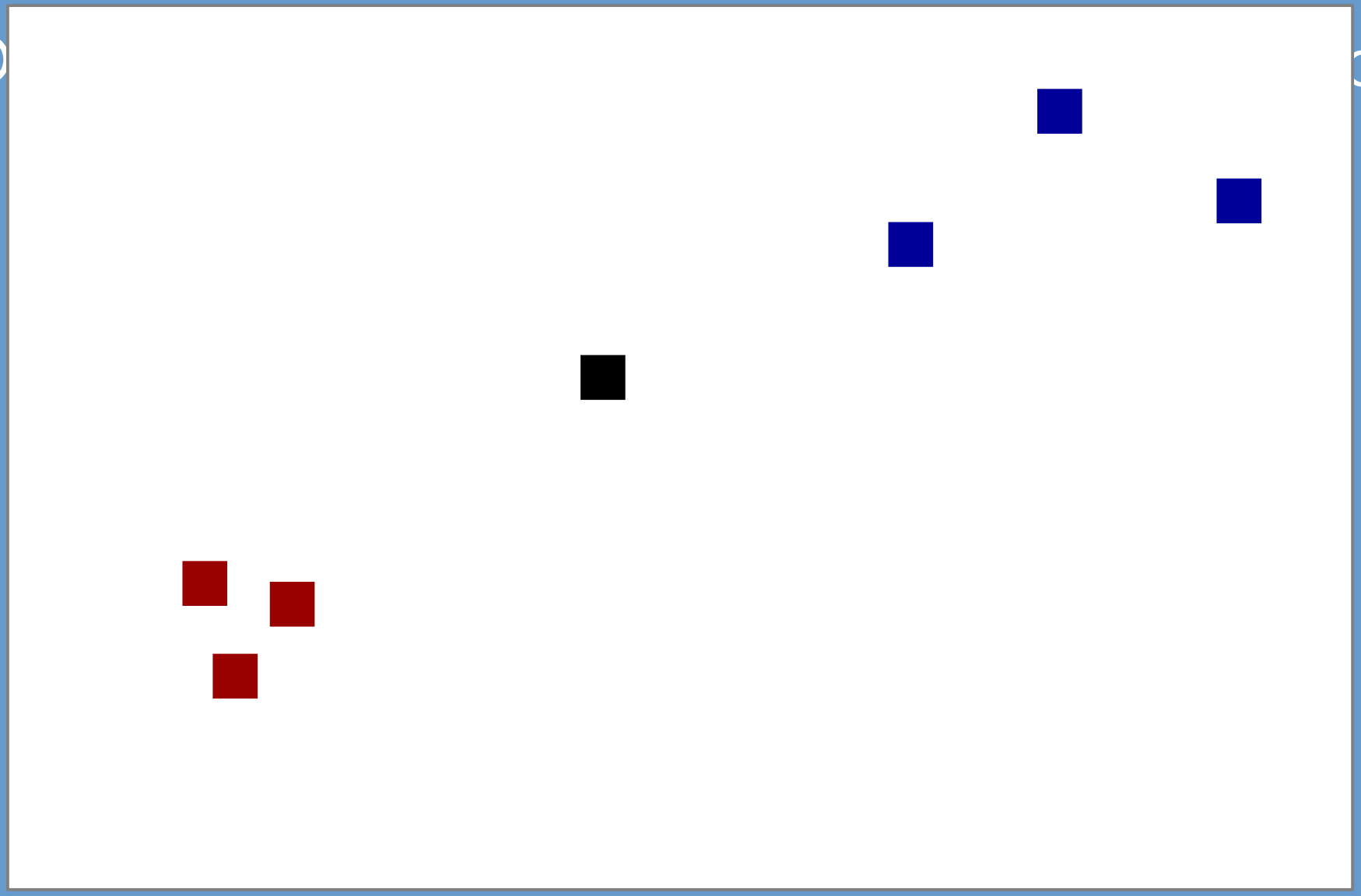


Hierarchical Grouping

- Distance measure (link) must be defined
  - Determines how "similarity" among observations is evaluated
    - Single link: nearest elements
    - Complete link: furthest elements
    - Ward's: within group variance
  - Most similar observations/groups are grouped first
    - Then, reevaluation (iteration)

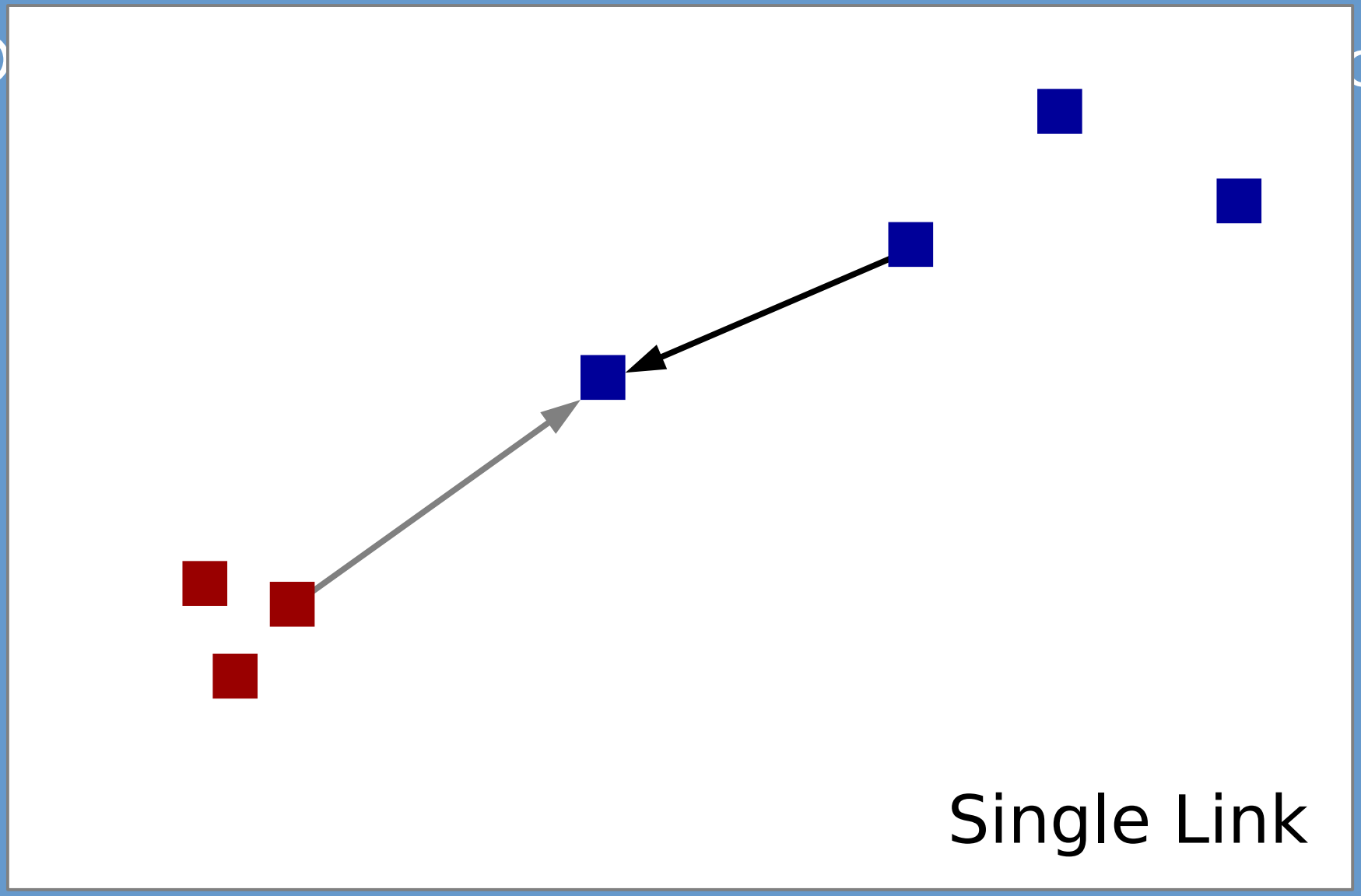
Hierarchical Grouping

- D d



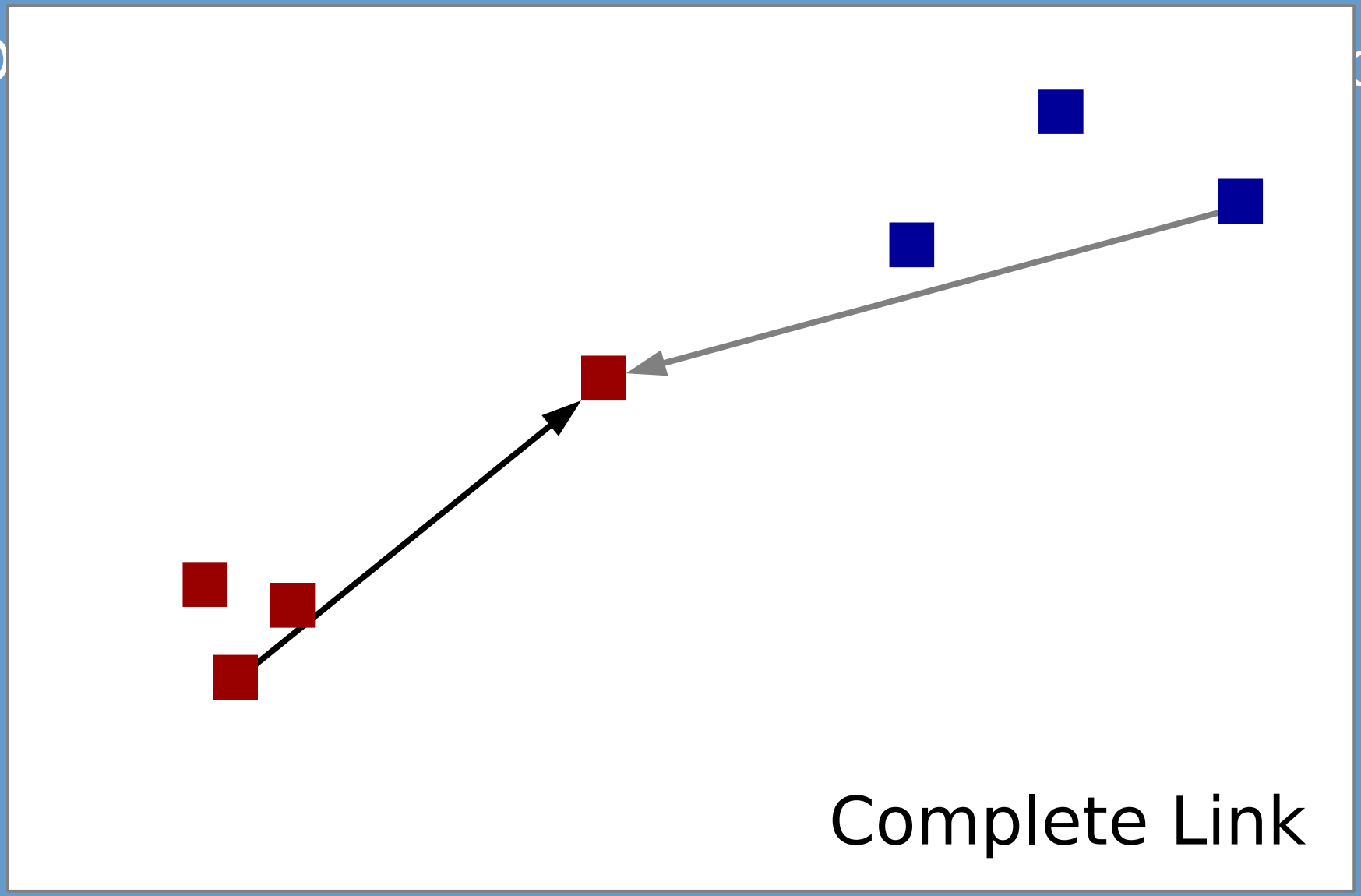
Hierarchical Grouping

- D d



Hierarchical Grouping

- D d

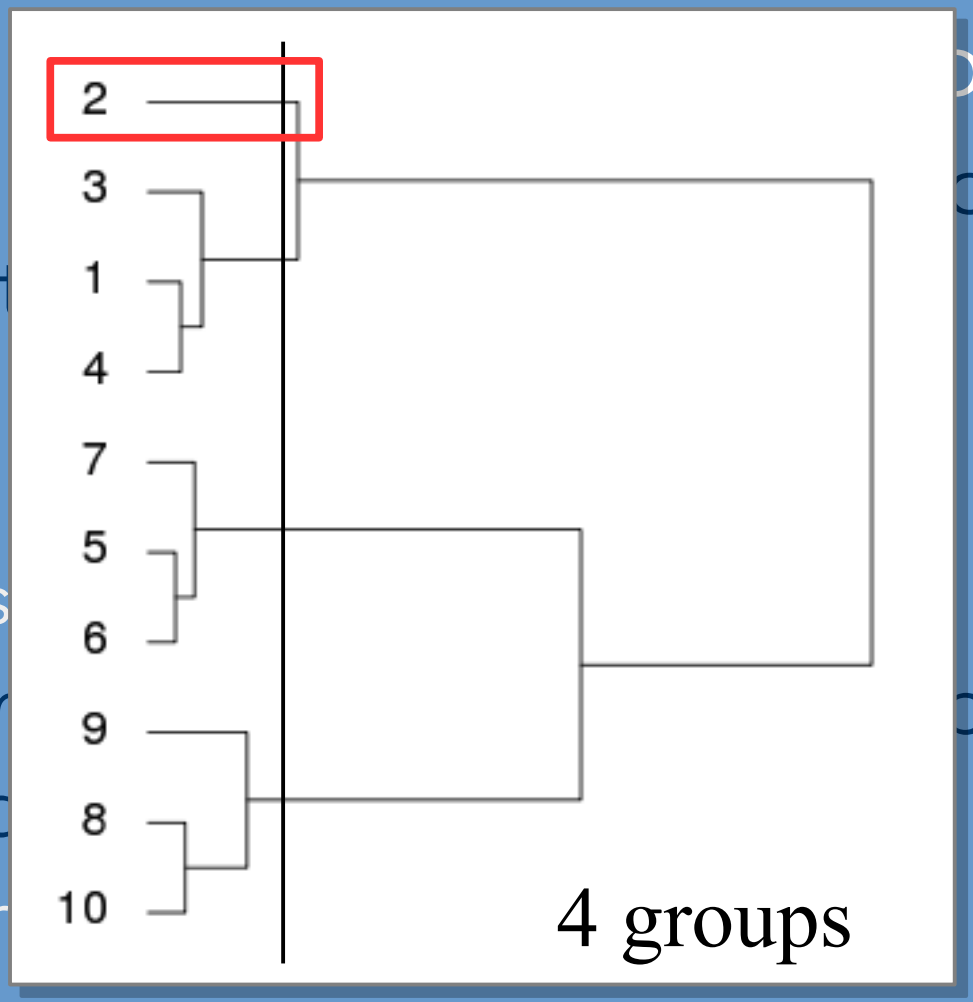


Hierarchical Grouping

- Distance measure (link) must be defined
  - Determines how "similarity" among observations is evaluated
    - Single link: nearest elements
    - Complete link: furthest elements
    - Ward's: within group variance
  - Most similar observations/groups are grouped first
    - Then, reevaluation (iteration)

Hierarchical Grouping

- Distance
  - Determi
    observat
    - Single
    - Compl
    - Ward's
  - Most sim
    groupec
    - Then, r

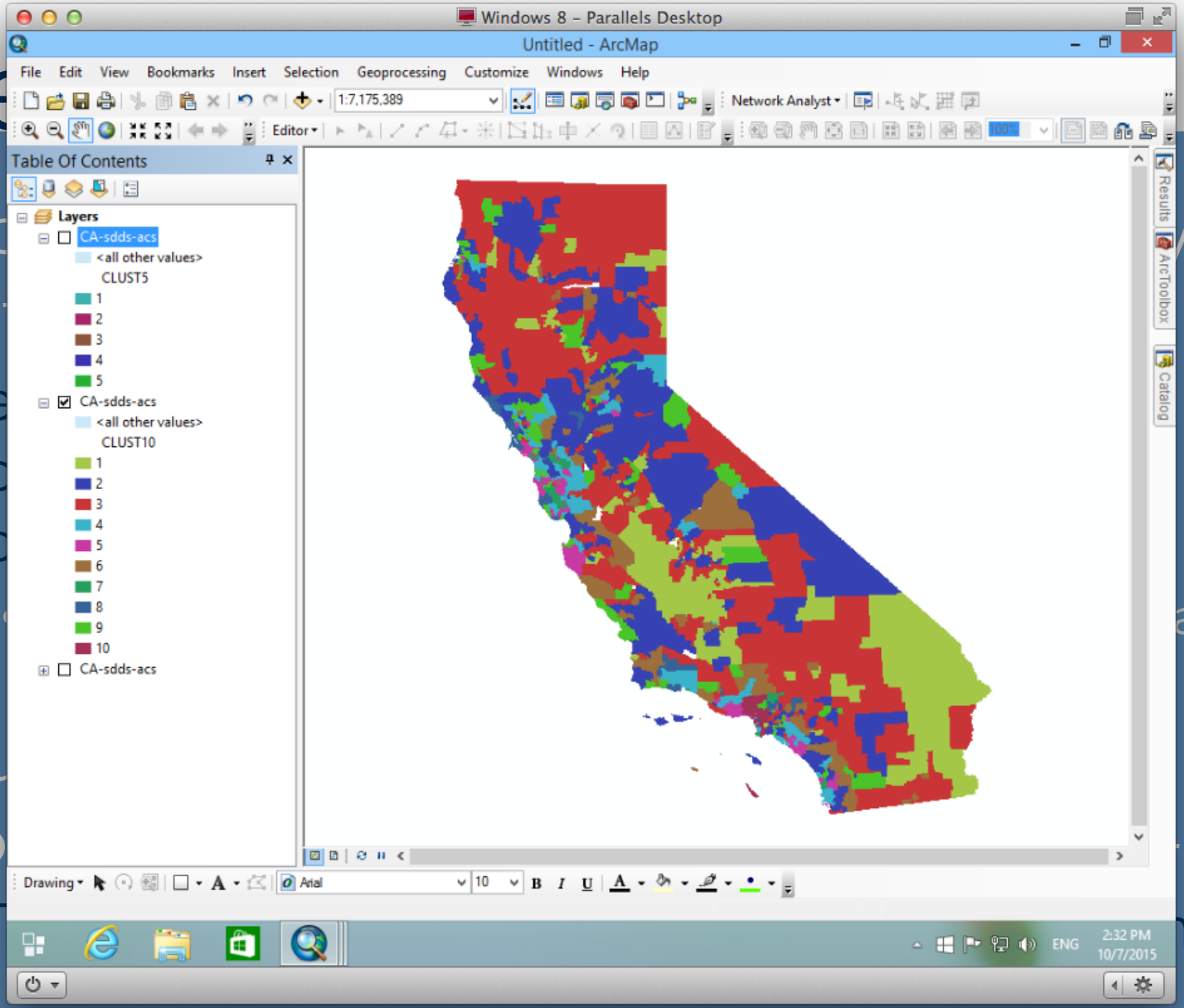


Hierarchical Grouping

- Distance measure (link) must be defined
  - Determines how "similarity" among observations is evaluated
    - Single link: nearest elements
    - Complete link: furthest elements
    - Ward's: within group variance
  - Most similar observations/groups are grouped first
    - Then, reevaluation (iteration)

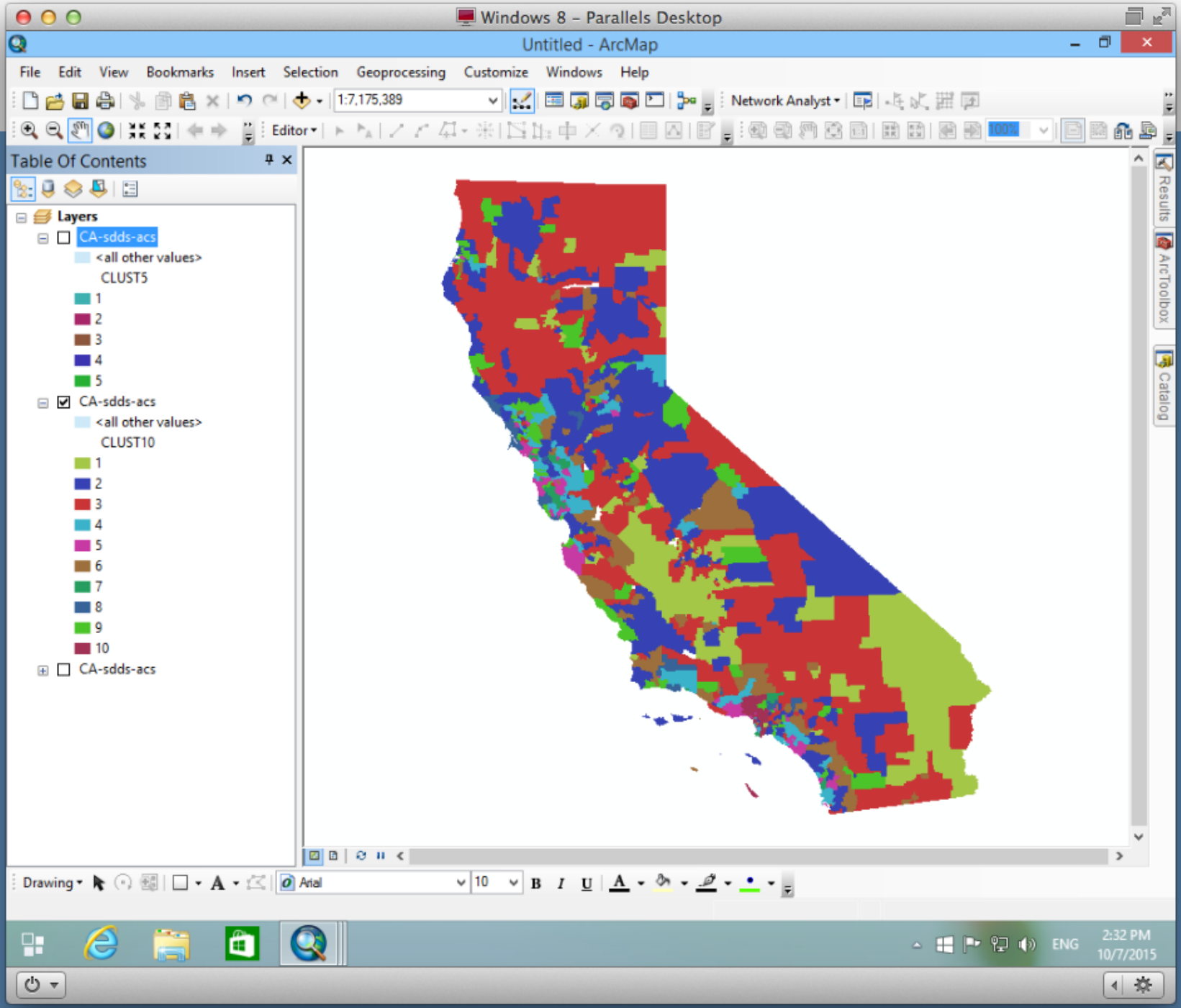
Hie

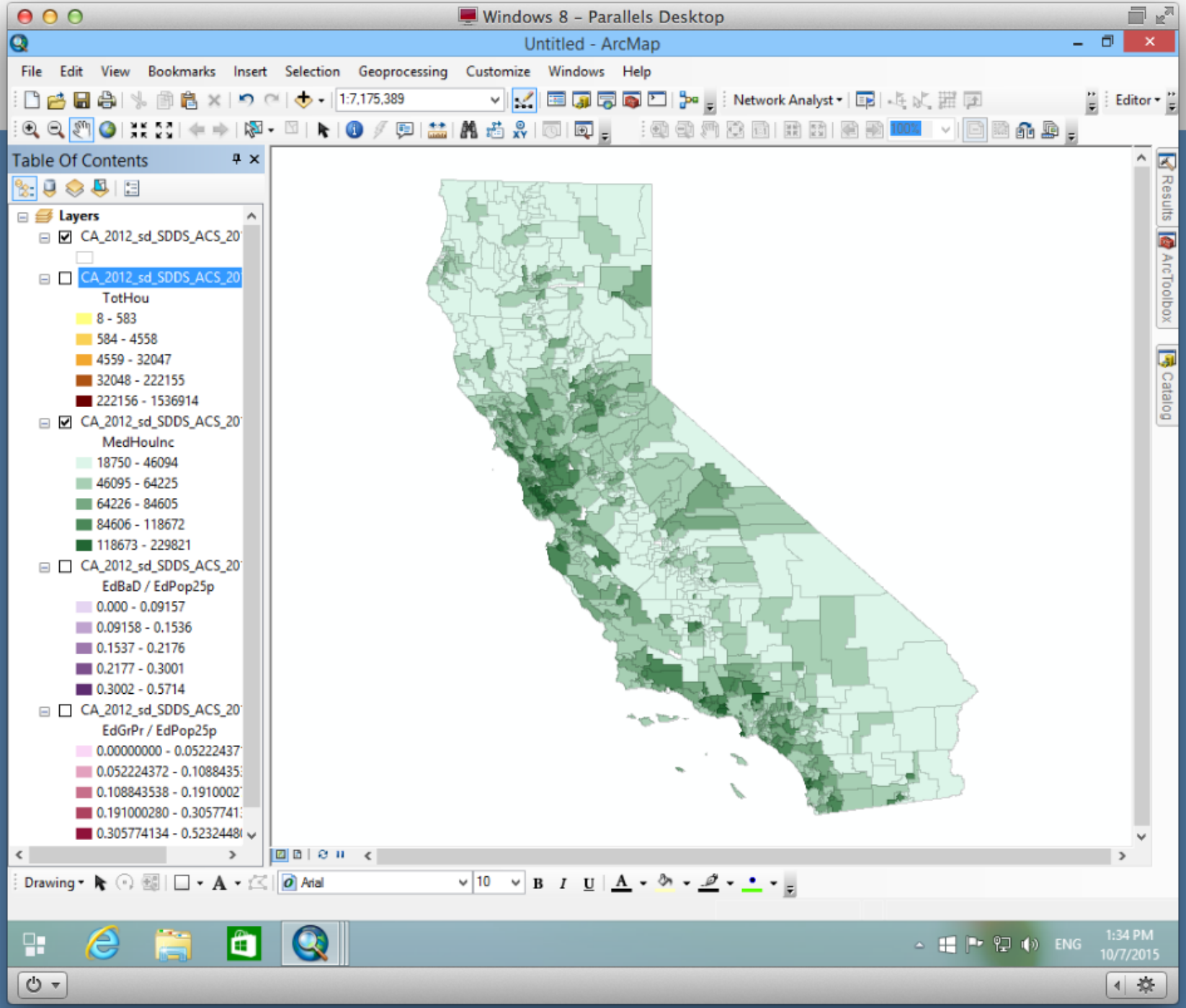
- Ob
  un
  - e
    c
    c
- Ou
  po
  - T



Spatial Autocorrelation

- Tobler's first law of Geography
  - *"Everything is related to everything else, but near things are more related than distant things"*
- Statistical grouping
  - Works in *m*-dimensional data space
    - With spatial data, location/space may be implicit within the attribute data
    - If observations are spatially autocorrelated


Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:7,175,389
Network Analyst
Editor
Table Of Contents
Layers
CA-sdds-acs
<all other values>
CLUST5
1
2
3
4
5
CA-sdds-acs
<all other values>
CLUST10
1
2
3
4
5
6
7
8
9
10
CA-sdds-acs
Results
ArcToolbox
Catalog
Drawing
Arial
10
ENG
2:32 PM
10/7/2015


Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:7,175,389
Network Analyst
Editor
Table Of Contents
Layers
CA_2012_sd_SDDS_ACS_20
TotHou
8 - 583
584 - 4558
4559 - 32047
32048 - 222155
222156 - 1536914
MedHouInc
18750 - 46094
46095 - 64225
64226 - 84605
84606 - 118672
118673 - 229821
EdBaD / EdPop25p
0.000 - 0.09157
0.09158 - 0.1536
0.1537 - 0.2176
0.2177 - 0.3001
0.3002 - 0.5714
EdGrPr / EdPop25p
Results
ArcToolbox
Catalog
Drawing
Arial
10
ENG
1:34 PM
10/7/2015

Hierarchical Grouping

- Works in *m*-dimensional data space
- Grouping is based on similarity across multiple attributes
  - Sensitive to the "scale" of the attributes
  - Important to properly weigh the attributes prior to grouping

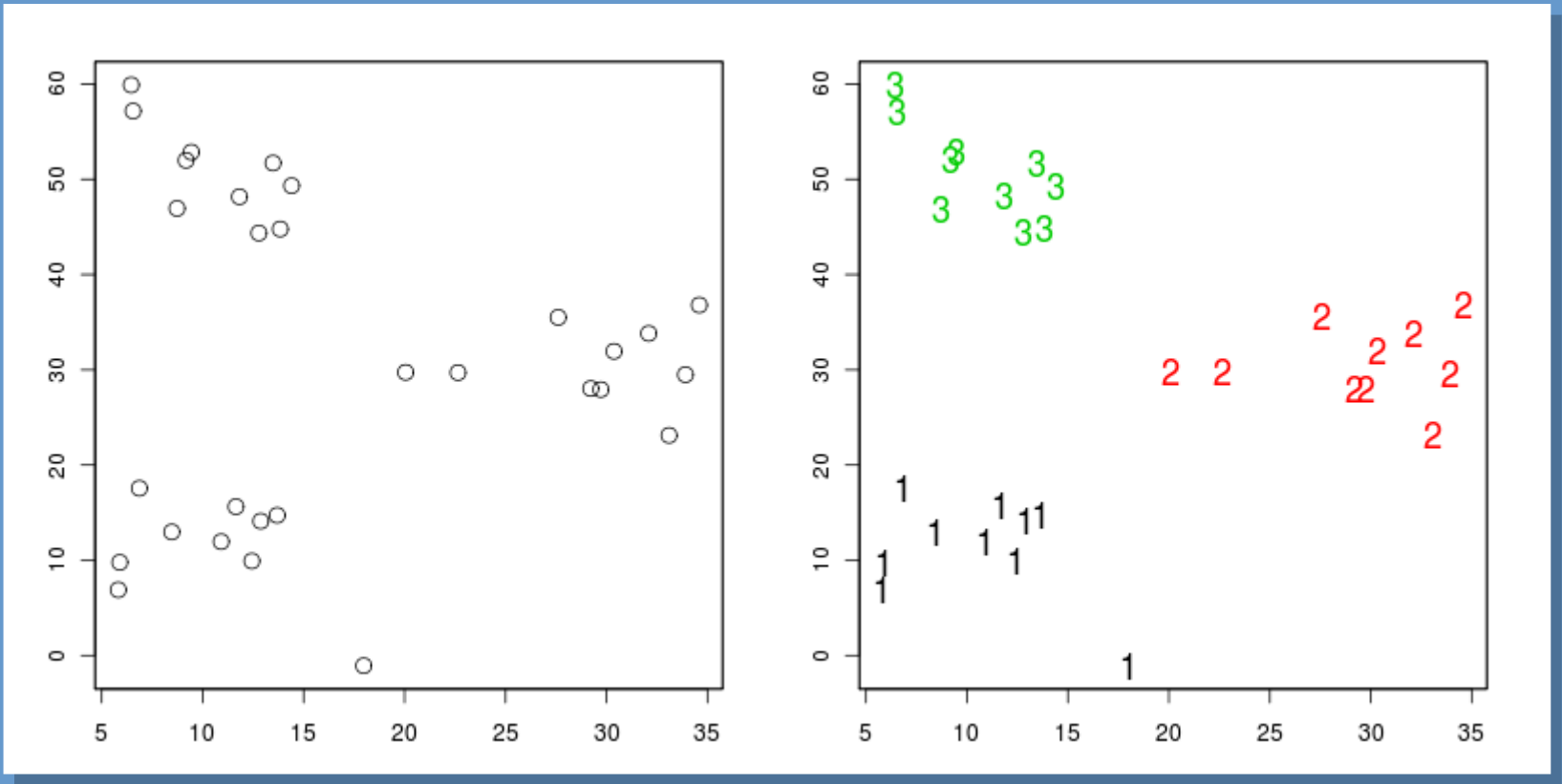
Partitional Grouping

- Observations are grouped into a user-defined number of groups (*k*)
  - Number of groups is chosen prior to implementation of the method
- Heuristic driven
  - Basically, searches for the best (optimal) solution for the user-defined *k*
    - An NP-hard problem
      - All possible solutions cannot be evaluated

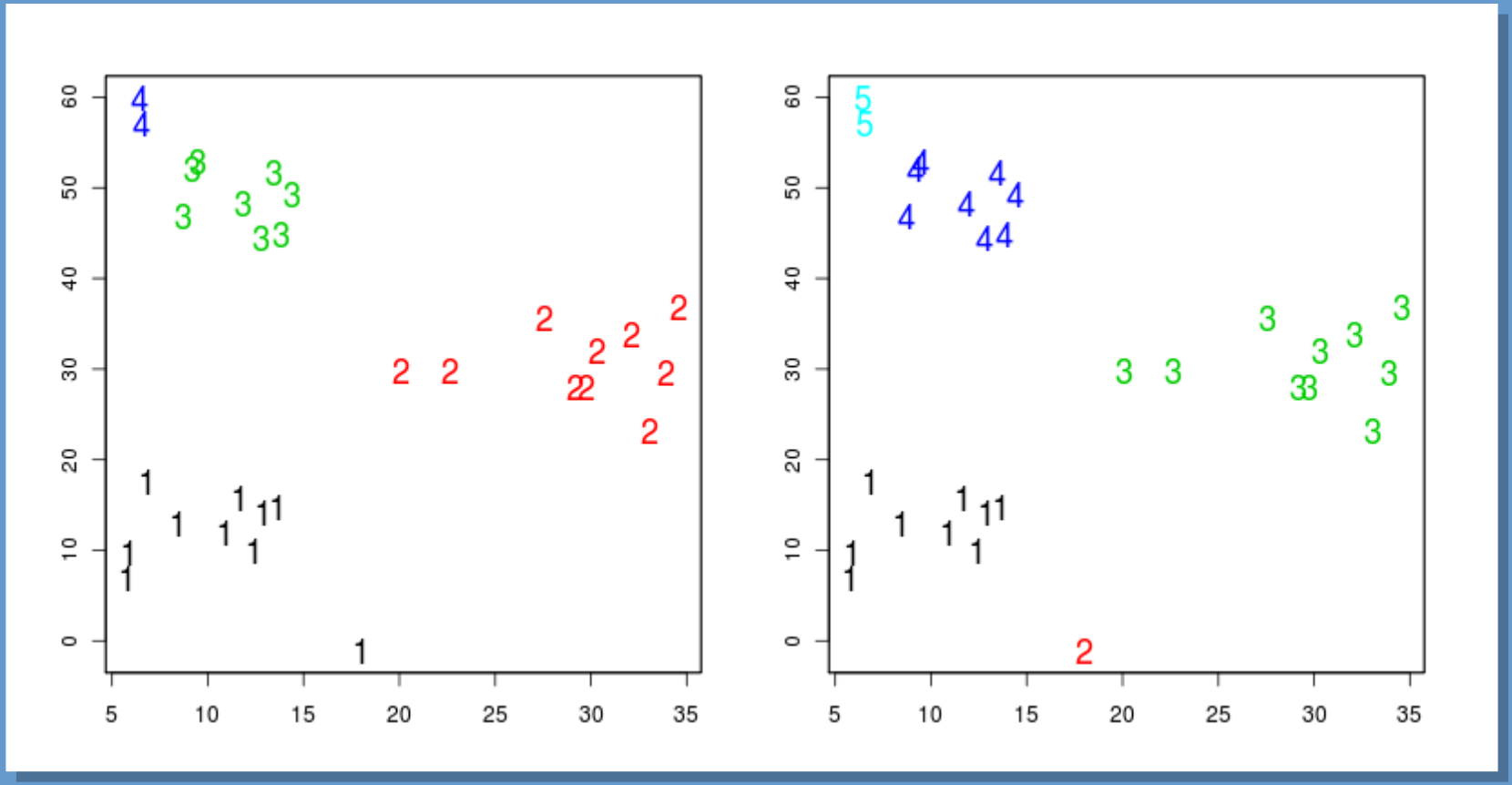
Partitional Grouping

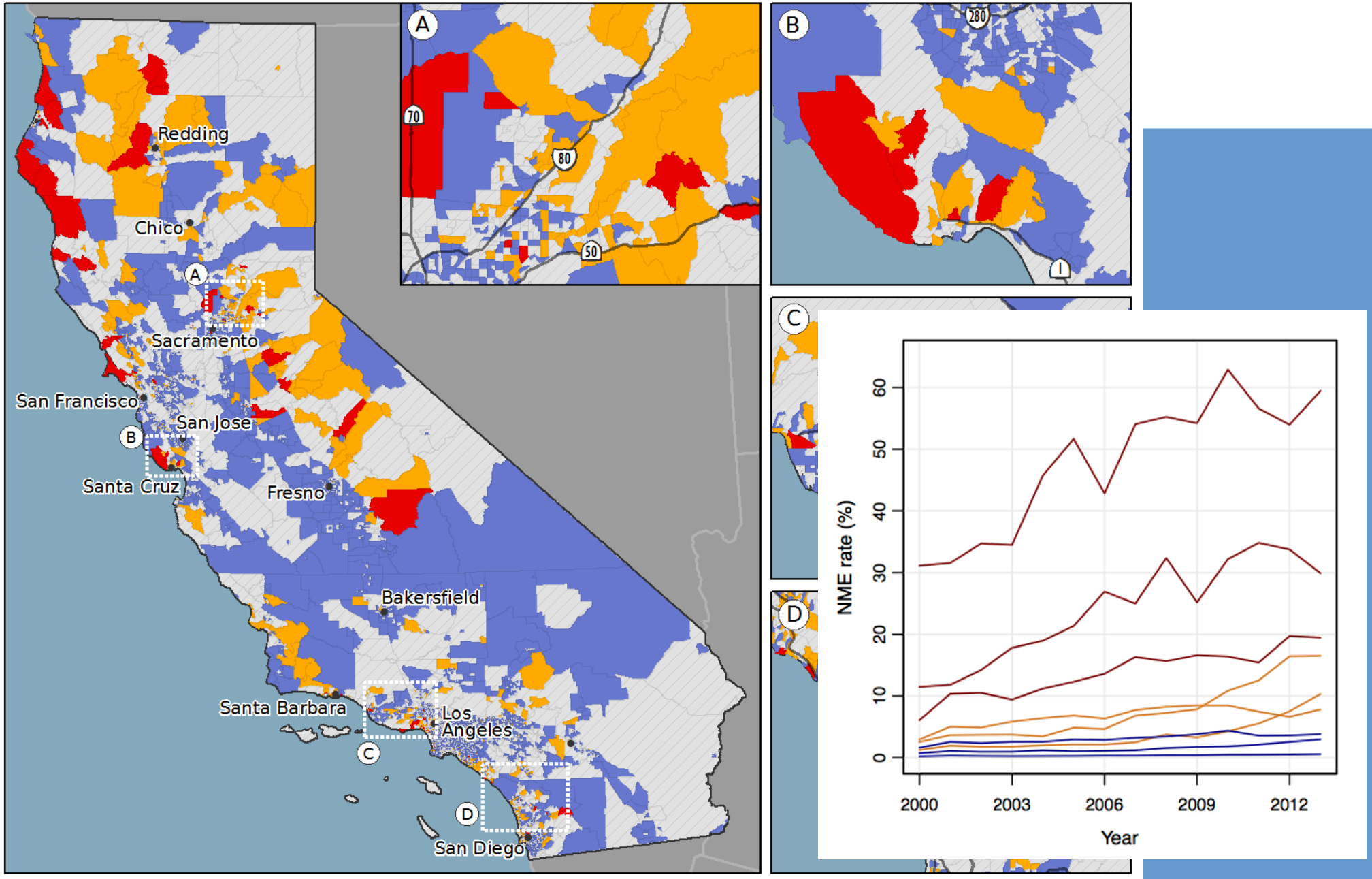
- K-means is most popular and widely used partitional grouping method
  - Works in *m*-dimensional data space
  - Heuristic finds set of groups that minimize within group variance and maximize between group variance
    - Similar logic to finding a "best fit" regression line

K-means



K-means




A
B
C
D
Redding
Chico
Sacramento
San Francisco
San Jose
Santa Cruz
Fresno
Bakersfield
Santa Barbara
Los Angeles
San Diego
70
80
50
280
1
NME rate (%)
0
10
20
30
40
50
60
2000
2003
2006
2009
2012
Year
Group Membership, NME %
Groups 1, 2, 4: Near 0% throughout or Starts low with moderate increase
Groups 3, 5, 6: Starts low with large increase
Groups 7, 8, 9: Starts high with large increase
Less than 14 years of data

Spatial Grouping

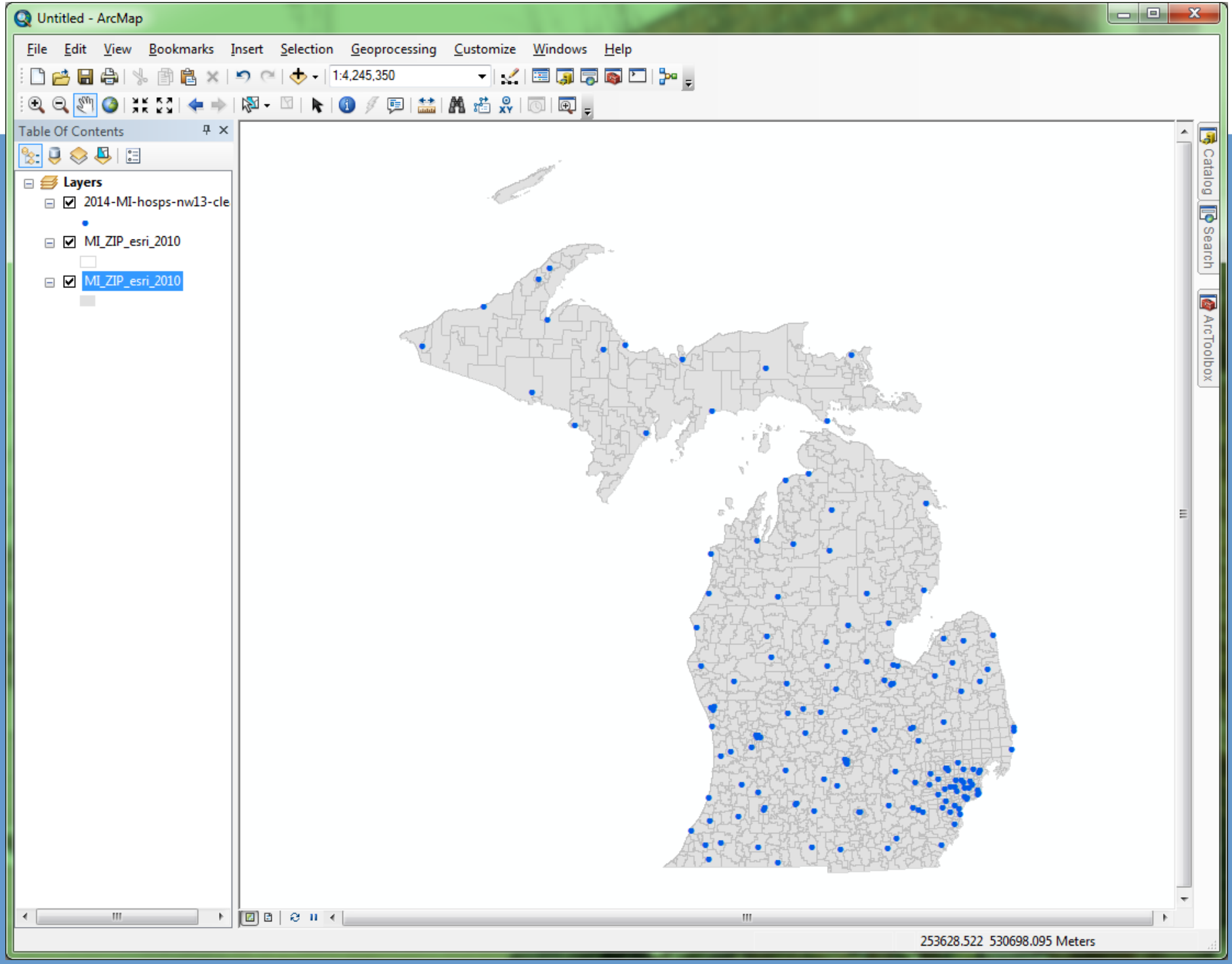
- Spatial grouping only considers the location of the observations
  - Does not address attribute similarity
- Requires a measure of similarity among locations
  - How do we define nearness?
    - e.g., distance, connectivity (topology)
- In a GIS, may require multiple steps

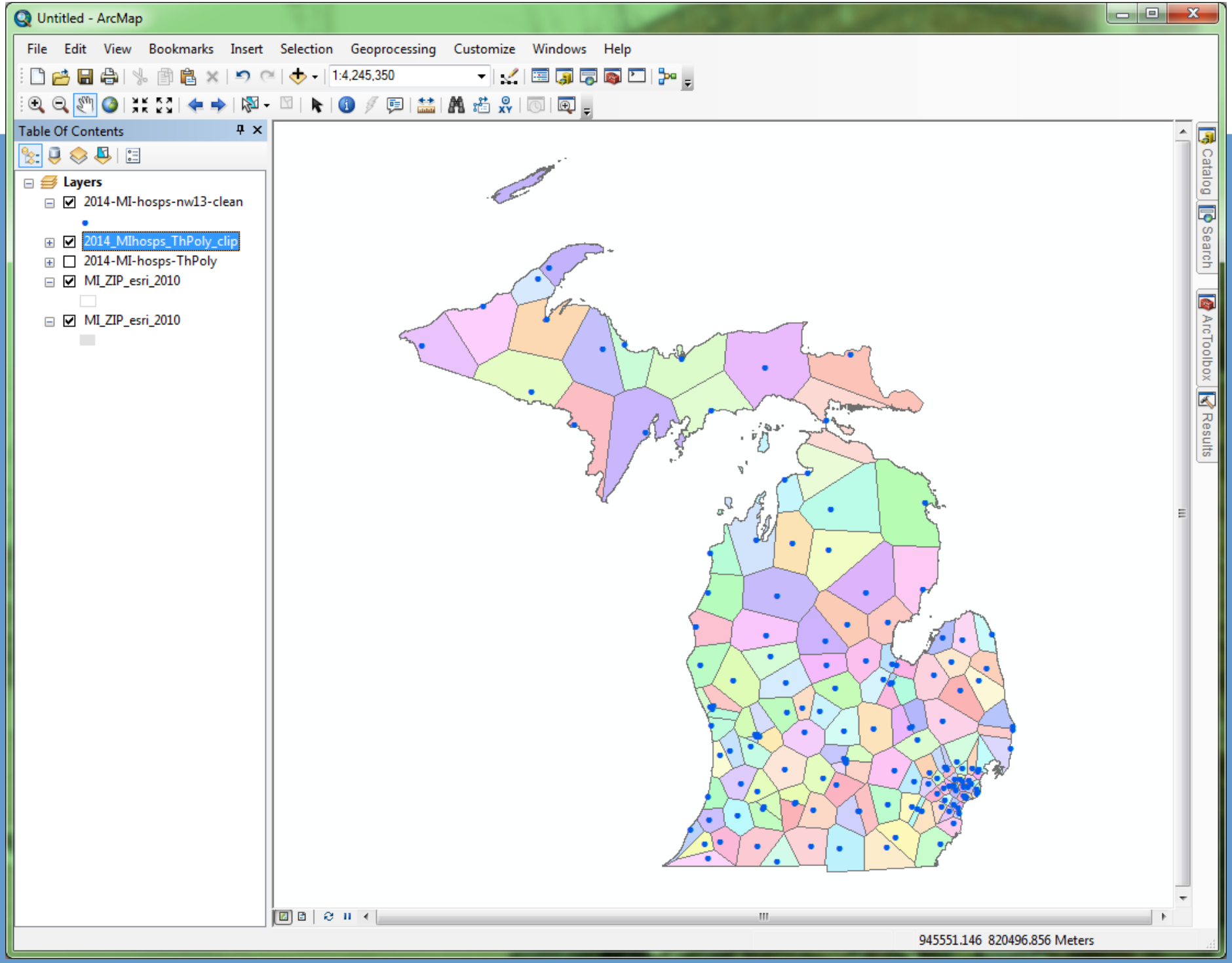
Thiessen Polygons

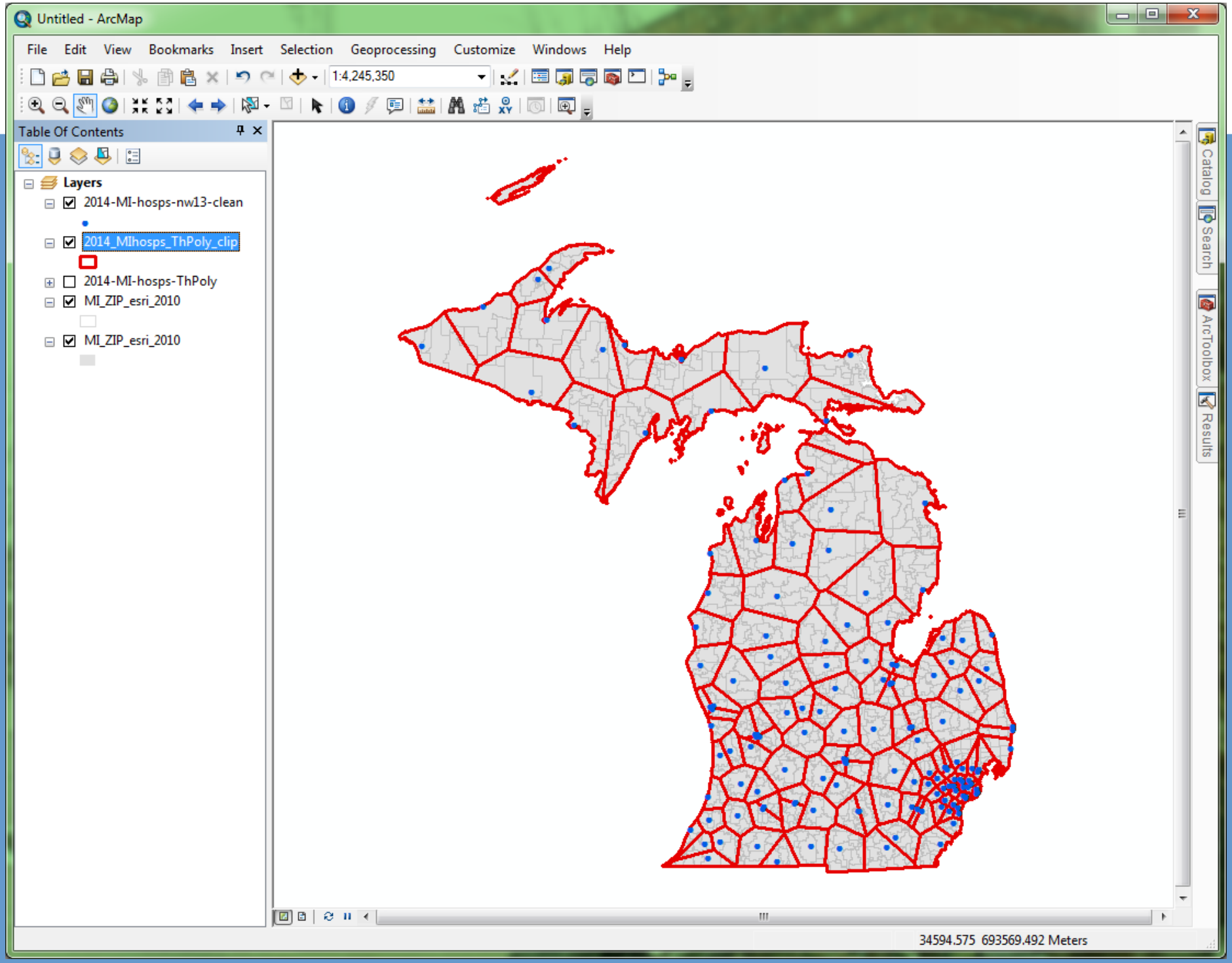
- Vector-based proximal regions
  - Based on Euclidean distance
  - Nearest region to a point location
    - Based on relative measure of distance
- Start with a set of points
  - Creates a new polygon for each input point
    - Observations within each polygon are assigned to the point to which they are nearest
    - Nearest point defines "group" membership

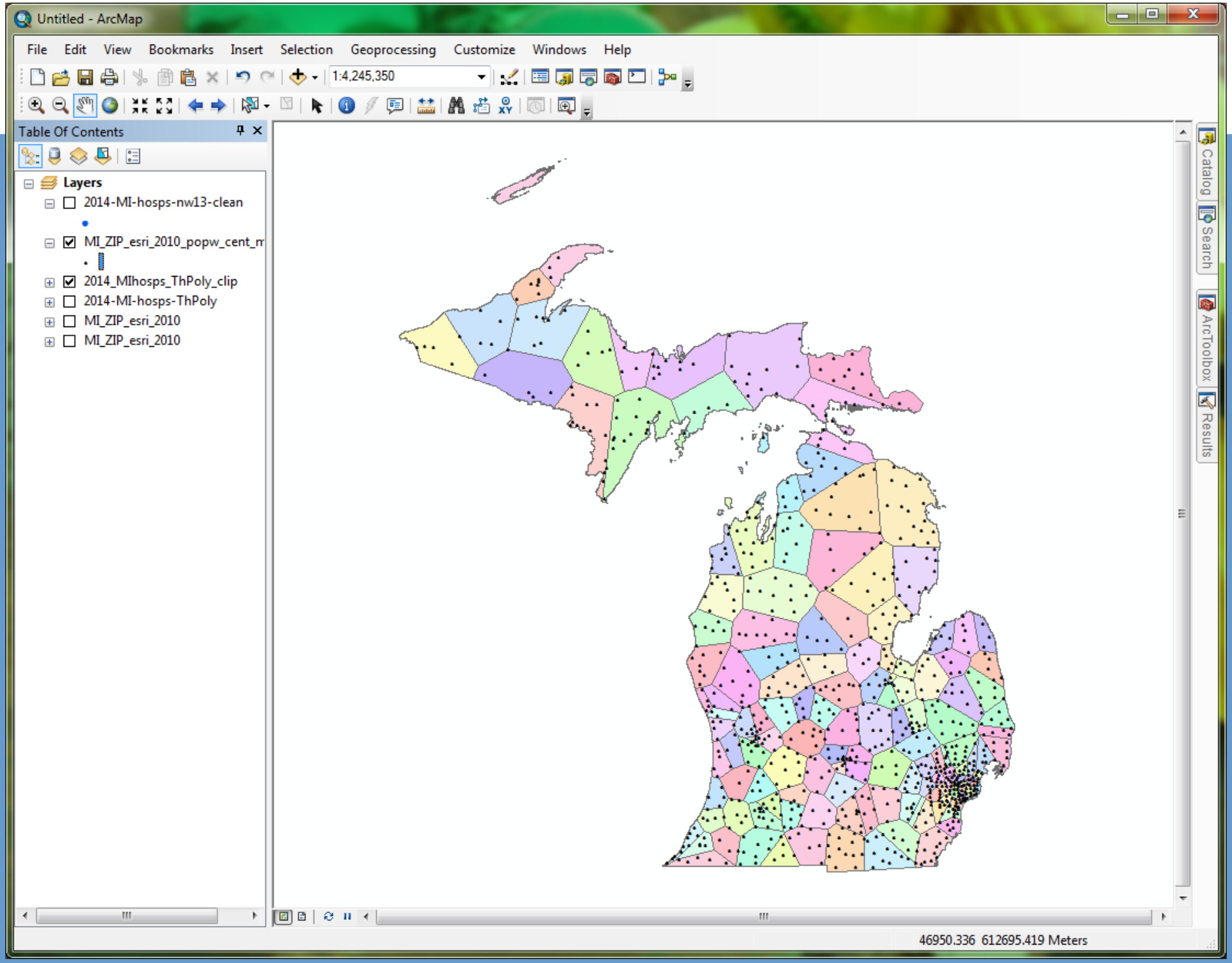
Thiessen Polygons

- Requires two steps
  - Create Thiessen polygons
  - Spatially overlay observation data
    - If observations are points, easy
    - If observations are lines or polygons, requires some decisions about assignment


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-hosps-nw13-cle
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
253628.522 530698.095 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-hosps-nw13-clean
2014_MIhosps_ThPoly_clip
2014-MI-hosps-ThPoly
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
Results
945551.146 820496.856 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-hosps-nw13-clean
2014_MIhosps_ThPoly_clip
2014-MI-hosps-ThPoly
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
Results
34594.575 693569.492 Meters

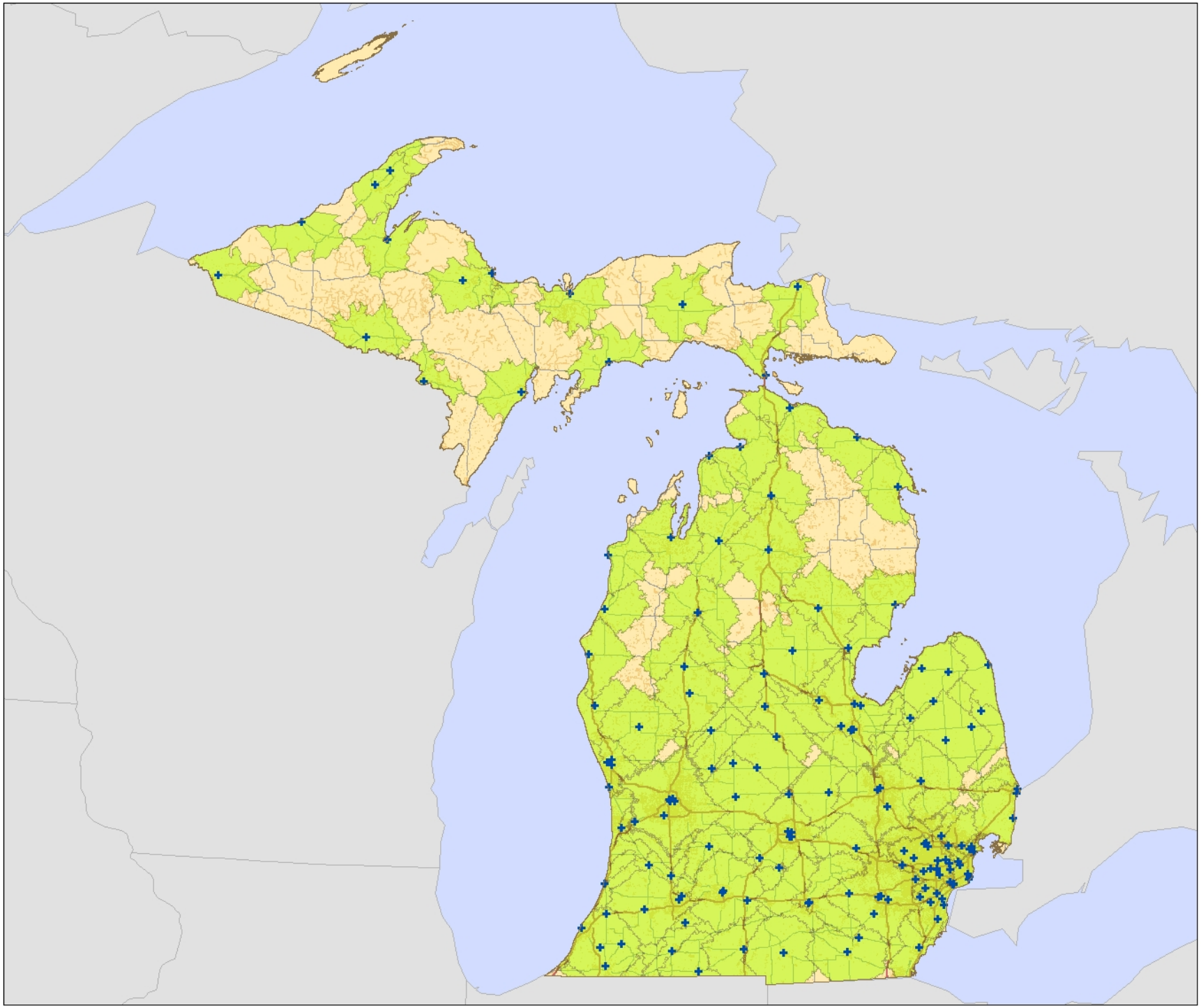

Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-hosps-nw13-clean
MI_ZIP_esri_2010_popw_cent_n
2014_MIhosps_ThPoly_clip
2014-MI-hosps-ThPoly
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
Results
46950.336 612695.419 Meters

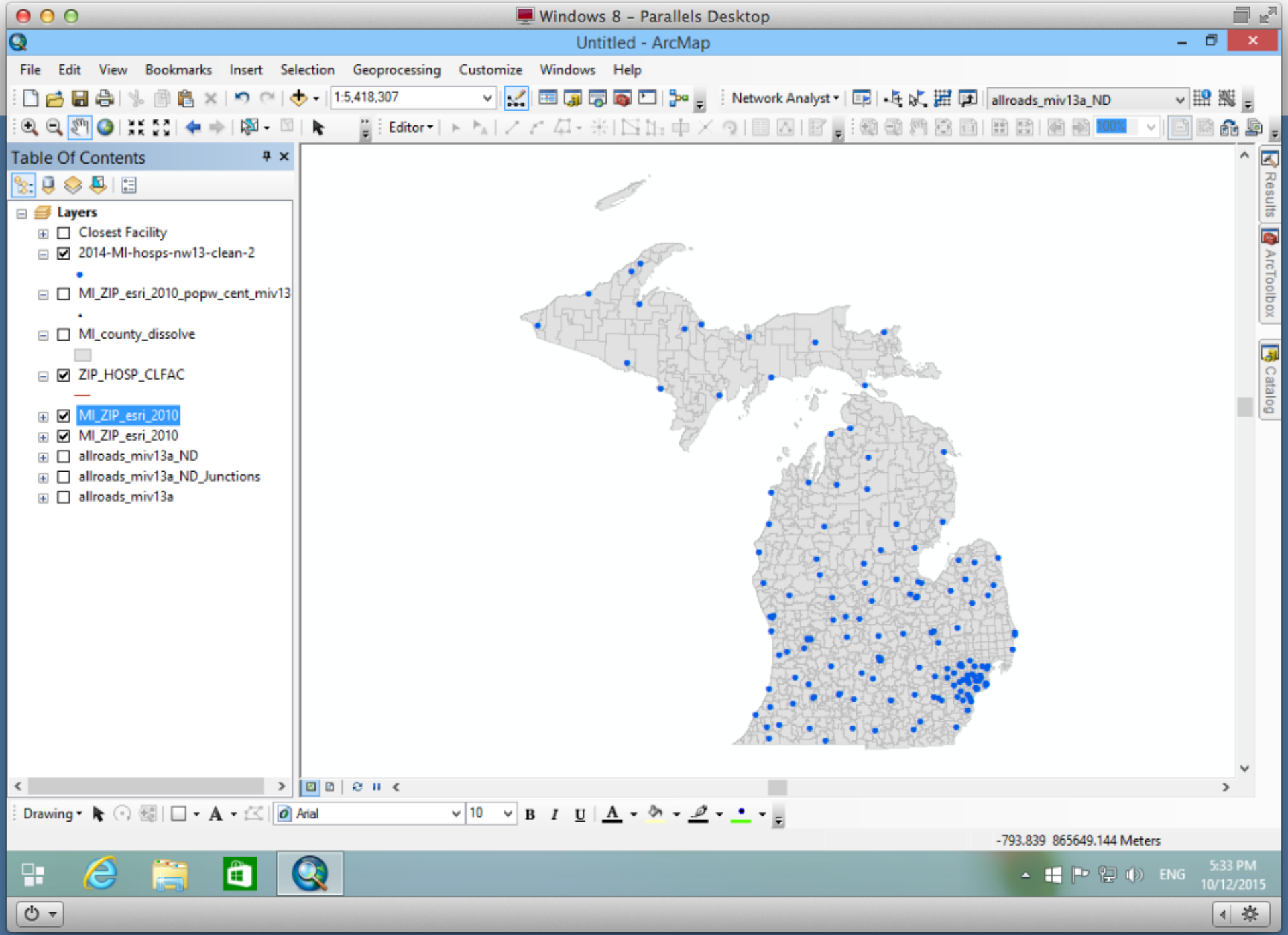
Network Service Areas

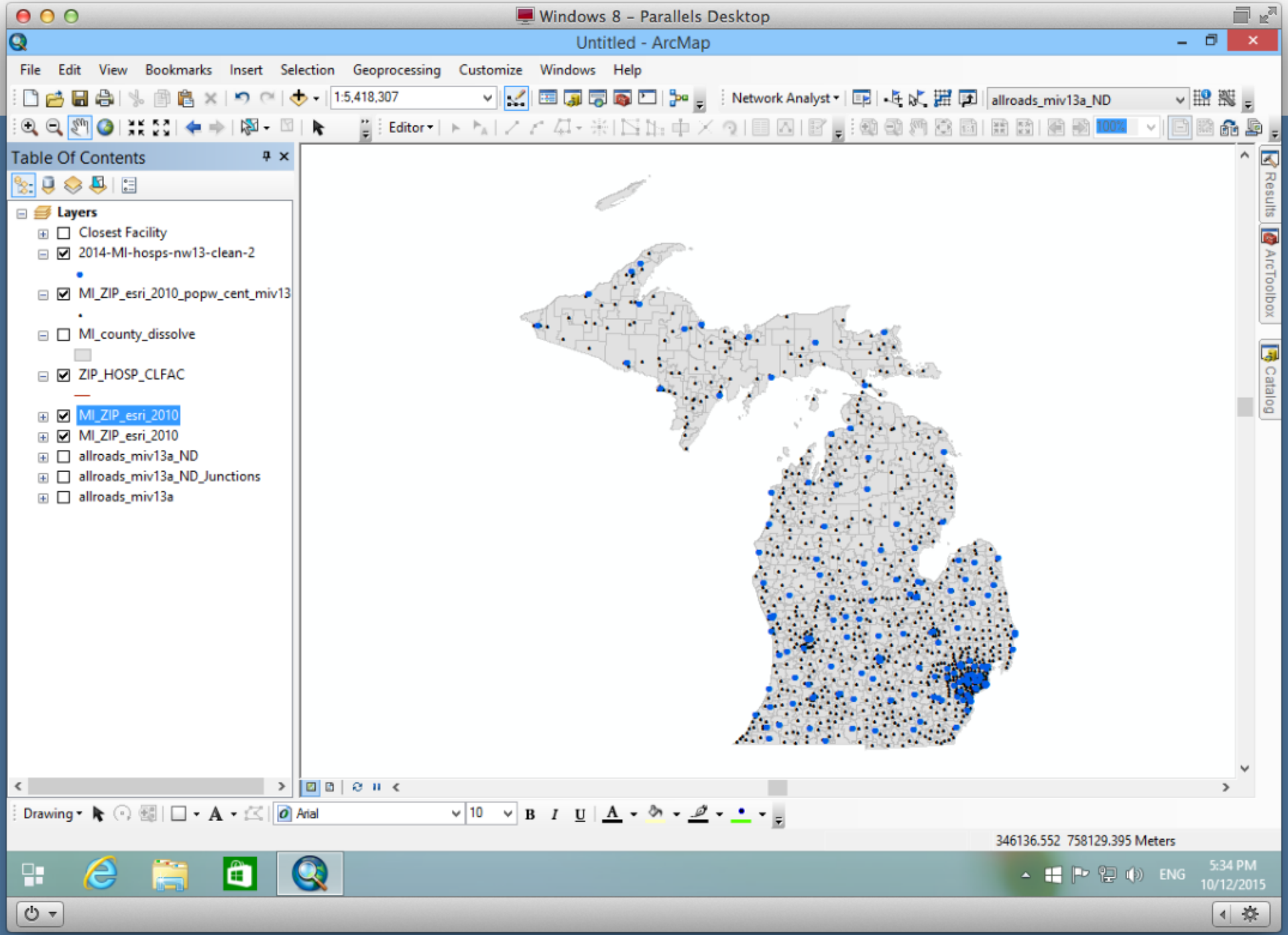
- Similar to Thiessen polygons
  - However, use "network" distance rather than Euclidean distance
    - e.g., travel distance or travel time along a road network
- Network
  - System of connected lines and points
    - In a GIS, geometry is similar to vector data
    - Connectivity is the key concept

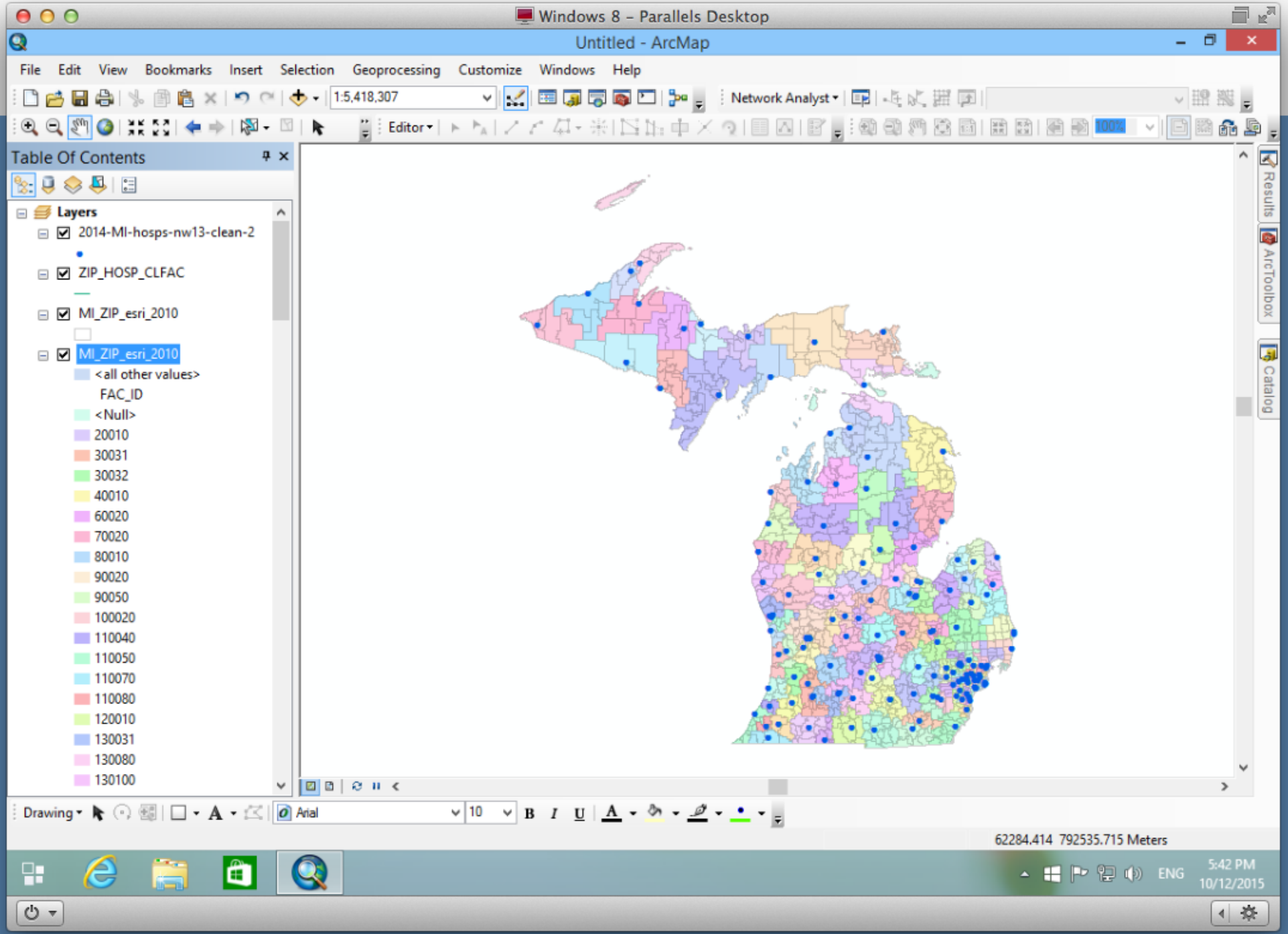
Network Service Areas

- Given a network, the goal is to find all locations serviced by a single location on the network, within a certain distance or time
  - Output is set of polygon features
- How it works
  - Trace all routes to a distance/time
  - Mark "endpoint" along all edges
  - Create polygon



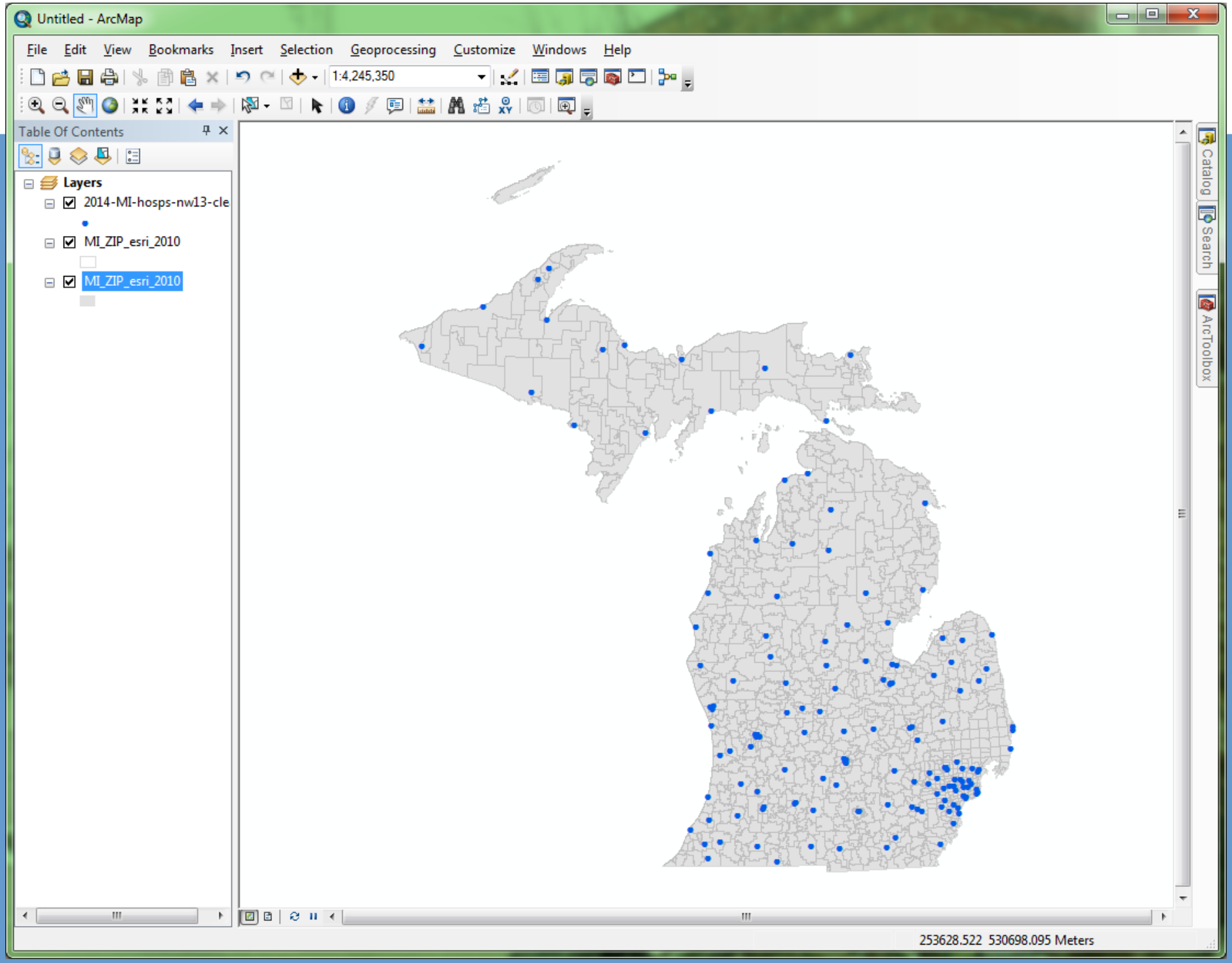

Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:5,418,307
Network Analyst
allroads_miv13a_ND
Editor
Table Of Contents
Layers
Closest Facility
2014-MI-hosps-nw13-clean-2
MI_ZIP_esri_2010_popw_cent_miv13
MI_county_dissolve
ZIP_HOSP_CLFAC
MI_ZIP_esri_2010
MI_ZIP_esri_2010
allroads_miv13a_ND
allroads_miv13a_ND_Junctions
allroads_miv13a
Results
ArcToolbox
Catalog
Drawing
Arial
10
-793.839 865649.144 Meters
ENG
5:33 PM
10/12/2015

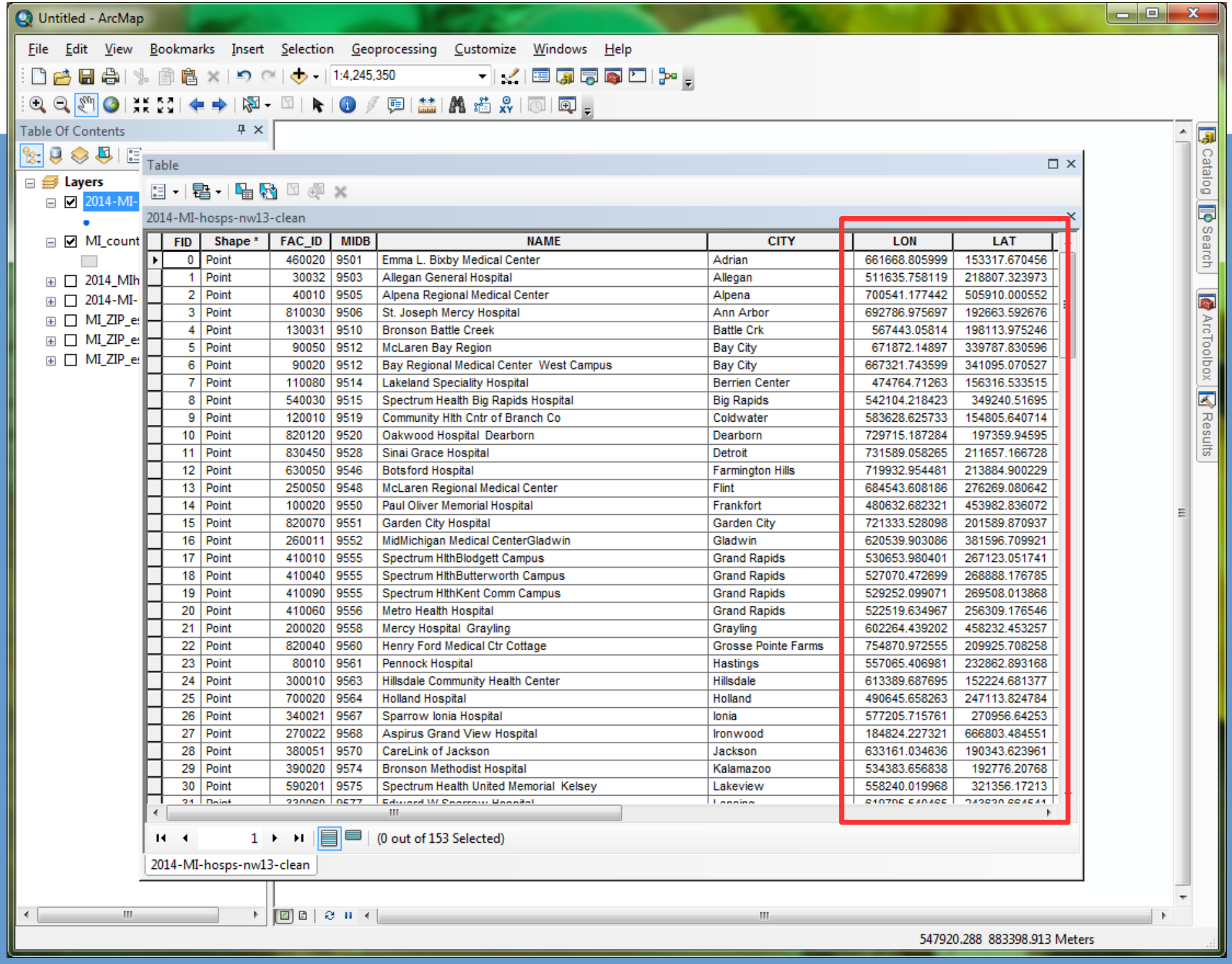

Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:5,418,307
Network Analyst
allroads_miv13a_ND
Editor
Table Of Contents
Layers
Closest Facility
2014-MI-hosps-nw13-clean-2
MI_ZIP_esri_2010_popw_cent_miv13
MI_county_dissolve
ZIP_HOSP_CLFAC
MI_ZIP_esri_2010
MI_ZIP_esri_2010
allroads_miv13a_ND
allroads_miv13a_ND_Junctions
allroads_miv13a
Results
ArcToolbox
Catalog
Drawing
Arial
10
346136.552 758129.395 Meters
ENG
5:34 PM
10/12/2015

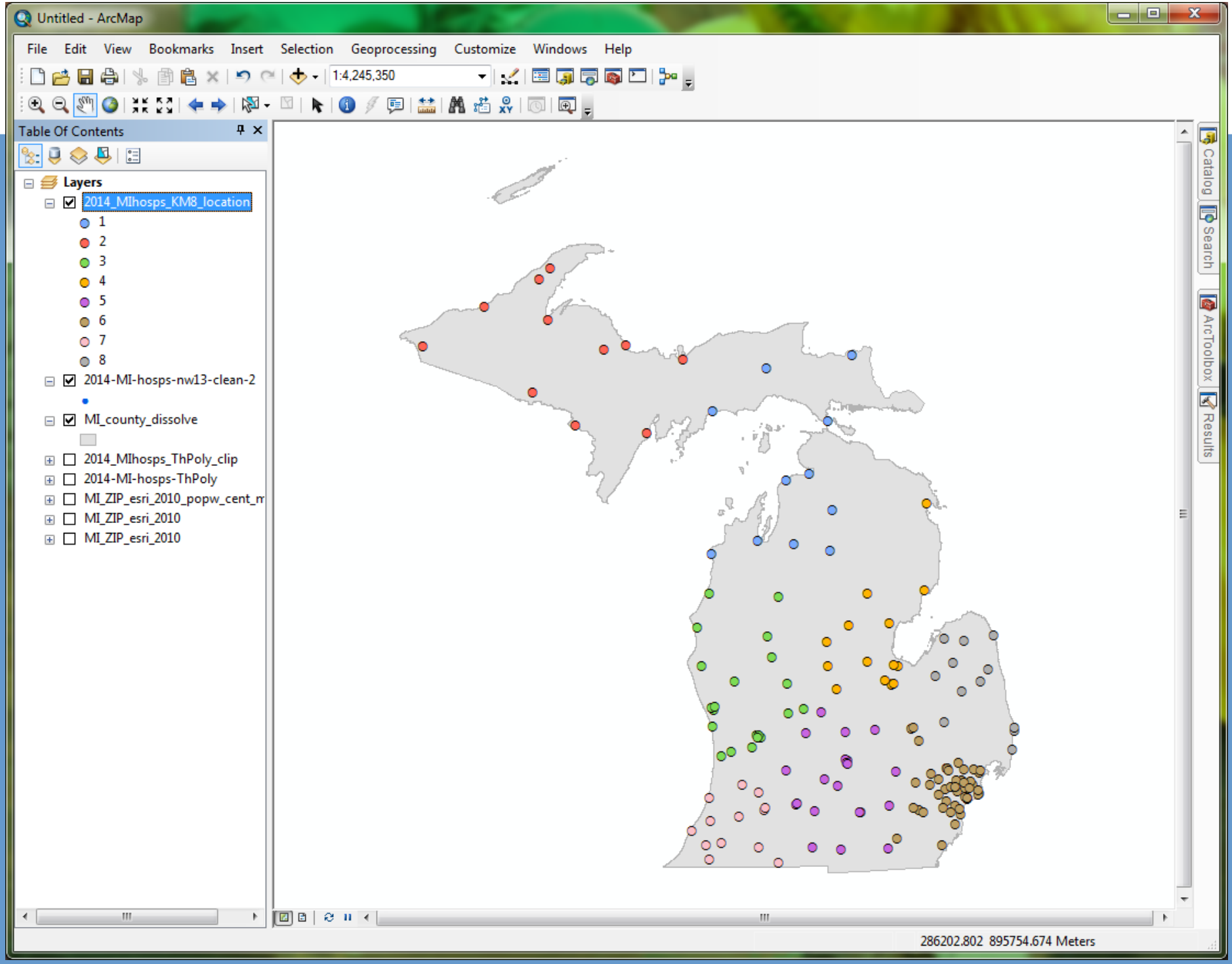

Windows 8 – Parallels Desktop
Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:5,418,307
Network Analyst
Editor
Table Of Contents
Layers
2014-MI-hosps-nw13-clean-2
ZIP_HOSP_CLFAC
MI_ZIP_esri_2010
MI_ZIP_esri_2010
<all other values>
FAC_ID
<Null>
20010
30031
30032
40010
60020
70020
80010
90020
90050
100020
110040
110050
110070
110080
120010
130031
130080
130100
Results
ArcToolbox
Catalog
Drawing
Arial
10
62284.414 792535.715 Meters
ENG
5:42 PM
10/12/2015

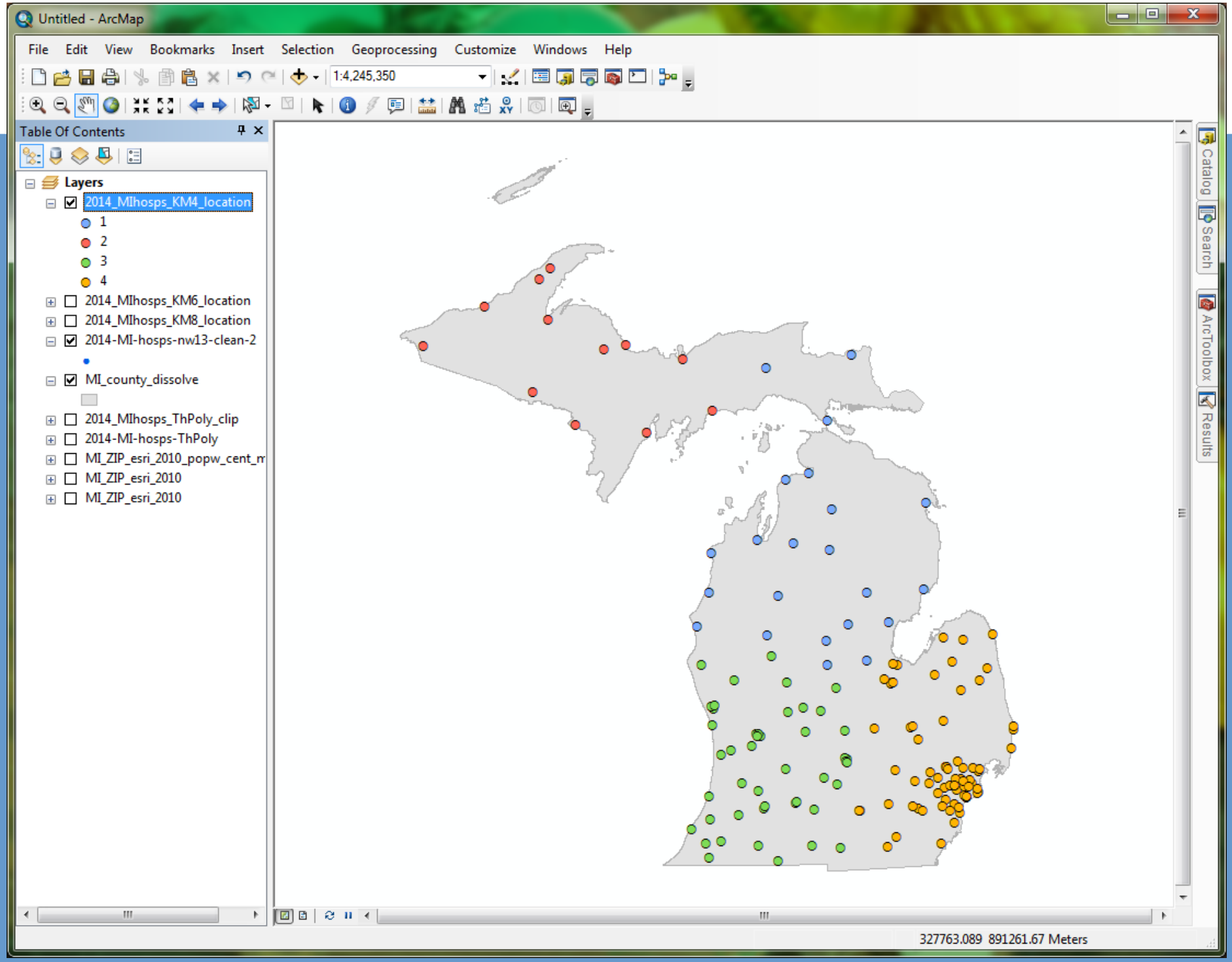
Space as an Attribute

- K-means works in *m*-dimensional data space
  - Simply use location information as attribute data!
    - Location of an object is set of attributes that define the distance from an origin point
    - For spatial data in two dimensions, we simply use distance along the X and Y axis from an origin
      - X,Y location can be used as attribute data


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-hosps-nw13-cle
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
253628.522 530698.095 Meters

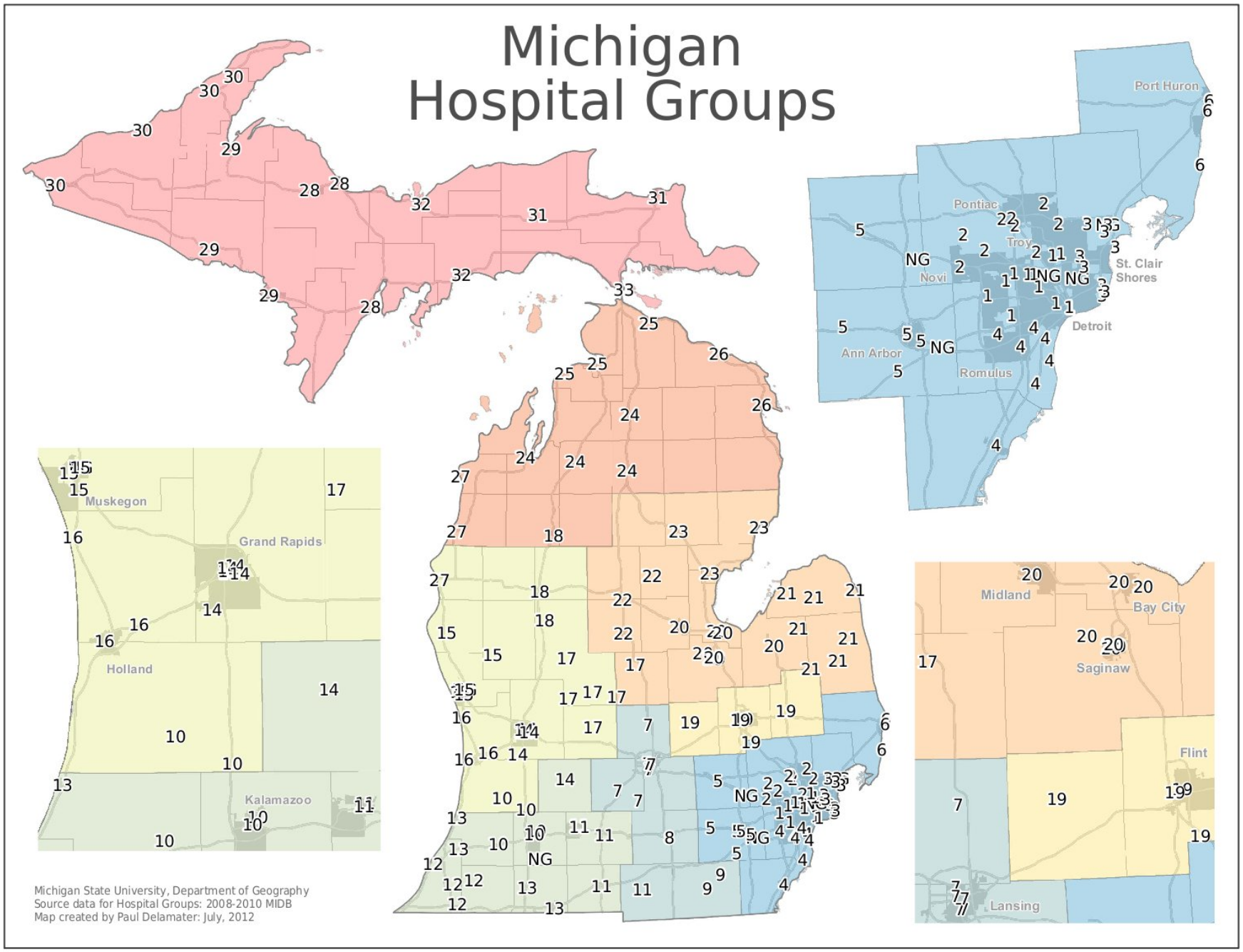

Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014-MI-
MI_count
2014_MIh
2014-MI-
MI_ZIP_e
MI_ZIP_e
MI_ZIP_e
Table
2014-MI-hosps-nw13-clean
FID Shape * FAC_ID MIDB NAME CITY LON LAT
0 Point 460020 9501 Emma L. Bixby Medical Center Adrian 661668.805999 153317.670456
1 Point 30032 9503 Allegan General Hospital Allegan 511635.758119 218807.323973
2 Point 40010 9505 Alpena Regional Medical Center Alpena 700541.177442 505910.000552
3 Point 810030 9506 St. Joseph Mercy Hospital Ann Arbor 692786.975697 192663.592676
4 Point 130031 9510 Bronson Battle Creek Battle Crk 567443.05814 198113.975246
5 Point 90050 9512 McLaren Bay Region Bay City 671872.14897 339787.830596
6 Point 90020 9512 Bay Regional Medical Center West Campus Bay City 667321.743599 341095.070527
7 Point 110080 9514 Lakeland Speciality Hospital Berrien Center 474764.71263 156316.533515
8 Point 540030 9515 Spectrum Health Big Rapids Hospital Big Rapids 542104.218423 349240.51695
9 Point 120010 9519 Community Hlth Cntr of Branch Co Coldwater 583628.625733 154805.640714
10 Point 820120 9520 Oakwood Hospital Dearborn Dearborn 729715.187284 197359.94595
11 Point 830450 9528 Sinai Grace Hospital Detroit 731589.058265 211657.166728
12 Point 630050 9546 Botsford Hospital Farmington Hills 719932.954481 213884.900229
13 Point 250050 9548 McLaren Regional Medical Center Flint 684543.608186 276269.080642
14 Point 100020 9550 Paul Oliver Memorial Hospital Frankfort 480632.682321 453982.836072
15 Point 820070 9551 Garden City Hospital Garden City 721333.528098 201589.870937
16 Point 260011 9552 MidMichigan Medical CenterGladwin Gladwin 620539.903086 381596.709921
17 Point 410010 9555 Spectrum HlthBlodgett Campus Grand Rapids 530653.980401 267123.051741
18 Point 410040 9555 Spectrum HlthButterworth Campus Grand Rapids 527070.472699 268888.176785
19 Point 410090 9555 Spectrum HlthKent Comm Campus Grand Rapids 529252.099071 269508.013868
20 Point 410060 9556 Metro Health Hospital Grand Rapids 522519.634967 256309.176546
21 Point 200020 9558 Mercy Hospital Grayling Grayling 602264.439202 458232.453257
22 Point 820040 9560 Henry Ford Medical Ctr Cottage Grosse Pointe Farms 754870.972555 209925.708258
23 Point 80010 9561 Pennock Hospital Hastings 557065.406981 232862.893168
24 Point 300010 9563 Hillsdale Community Health Center Hillsdale 613389.687695 152224.681377
25 Point 700020 9564 Holland Hospital Holland 490645.658263 247113.824784
26 Point 340021 9567 Sparrow Ionia Hospital Ionia 577205.715761 270956.64253
27 Point 270022 9568 Aspirus Grand View Hospital Ironwood 184824.227321 666803.484551
28 Point 380051 9570 CareLink of Jackson Jackson 633161.034636 190343.623961
29 Point 390020 9574 Bronson Methodist Hospital Kalamazoo 534383.656838 192776.20768
30 Point 590201 9575 Spectrum Health United Memorial Kelsey Lakeview 558240.019968 321356.17213
1 (0 out of 153 Selected)
2014-MI-hosps-nw13-clean
Catalog
Search
ArcToolbox
Results
547920.288 883398.913 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014_MIhosps_KM8_location
1
2
3
4
5
6
7
8
2014-MI-hosps-nw13-clean-2
MI_county_dissolve
2014_MIhosps_ThPoly_clip
2014-MI-hosps-ThPoly
MI_ZIP_esri_2010_popw_cent_m
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
Results
286202.802 895754.674 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:4,245,350
Table Of Contents
Layers
2014_MIhosps_KM4_location
1
2
3
4
2014_MIhosps_KM6_location
2014_MIhosps_KM8_location
2014-MI-hosps-nw13-clean-2
MI_county_dissolve
2014_MIhosps_ThPoly_clip
2014-MI-hosps-ThPoly
MI_ZIP_esri_2010_popw_cent_n
MI_ZIP_esri_2010
MI_ZIP_esri_2010
Catalog
Search
ArcToolbox
Results
327763.089 891261.67 Meters

Space as an Attribute

- Problem: Euclidean geometry may not capture true separation
  - Solution: calculated road distance from each hospital to all other hospitals in the state
    - "Location" of each hospital in relation to all other hospitals in network-like coordinates
    - A network distance O-D matrix

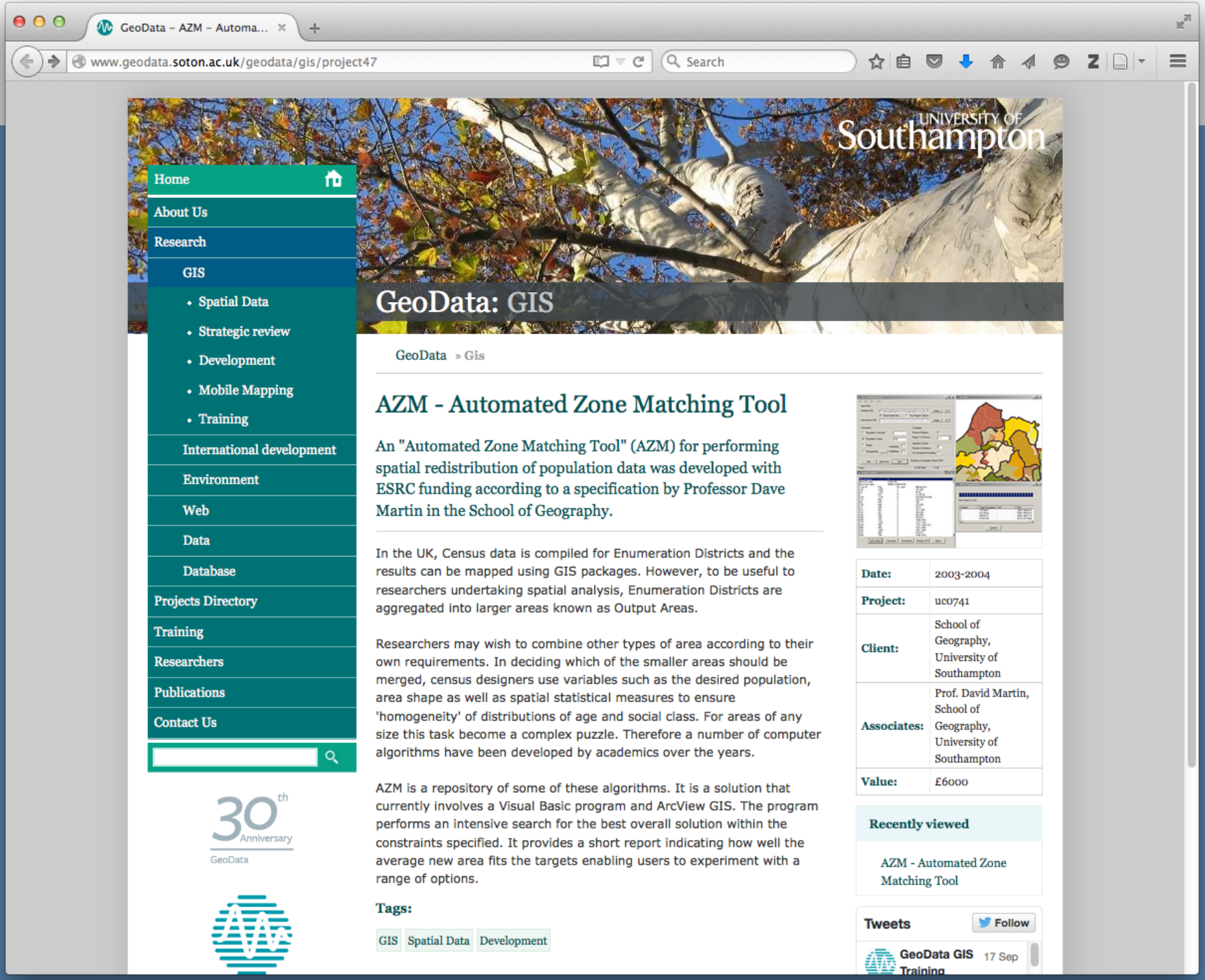

Michigan
Hospital Groups
Port Huron
Pontiac
Troy
Novi
St. Clair
Shores
Detroit
Ann Arbor
Romulus
Muskegon
Grand Rapids
Holland
Kalamazoo
Midland
Bay City
Saginaw
Flint
Lansing
Michigan State University, Department of Geography
Source data for Hospital Groups: 2008-2010 MIDB
Map created by Paul Delamater: July, 2012

Hybrid Grouping

- Hybrid grouping considers both the location and attributes of the observations
  - Specially designed algorithms
    - Simultaneously handle both spatial and statistical similarity
  - Representing space explicitly in statistical grouping approaches
    - Adding "space" as an attribute

Automated Zone Matching (AZM)

- Areal units are aggregated together to create groups
  - User-defined aggregation criteria
    - Cases or population thresholds
    - Shape properties (spatial compactness)
    - Internal homogeneity (similarity of observations)
  - Exploratory method
  - Begins with all areal units as single groups and begins aggregating
    - Performs hundreds or thousands of zone "swaps" to meet aggregation criteria


GeoData – AZM – Automa...
www.geodata.soton.ac.uk/geodata/gis/project47
Search
UNIVERSITY OF Southampton
GeoData: GIS
Home
About Us
Research
GIS
Spatial Data
Strategic review
Development
Mobile Mapping
Training
International development
Environment
Web
Data
Database
Projects Directory
Training
Researchers
Publications
Contact Us
30th Anniversary
GeoData
GeoData » Gis
AZM - Automated Zone Matching Tool
An "Automated Zone Matching Tool" (AZM) for performing spatial redistribution of population data was developed with ESRC funding according to a specification by Professor Dave Martin in the School of Geography.
In the UK, Census data is compiled for Enumeration Districts and the results can be mapped using GIS packages. However, to be useful to researchers undertaking spatial analysis, Enumeration Districts are aggregated into larger areas known as Output Areas.
Researchers may wish to combine other types of area according to their own requirements. In deciding which of the smaller areas should be merged, census designers use variables such as the desired population, area shape as well as spatial statistical measures to ensure 'homogeneity' of distributions of age and social class. For areas of any size this task become a complex puzzle. Therefore a number of computer algorithms have been developed by academics over the years.
AZM is a repository of some of these algorithms. It is a solution that currently involves a Visual Basic program and ArcView GIS. The program performs an intensive search for the best overall solution within the constraints specified. It provides a short report indicating how well the average new area fits the targets enabling users to experiment with a range of options.
Tags:
GIS
Spatial Data
Development
Date: 2003-2004
Project: uc0741
Client: School of Geography, University of Southampton
Associates: Prof. David Martin, School of Geography, University of Southampton
Value: £6000
Recently viewed
AZM - Automated Zone Matching Tool
Tweets
Follow
GeoData GIS Training
17 Sep

International Journal of Health Geographics

BioMed Central

Research

Open Access

Geographic analysis of low birthweight and infant mortality in Michigan using automated zoning methodology

Sue C Grady*[1] and Helen Enander[2]

Address: [1]Department of Geography, 130 Geography Building, Michigan State University, East Lansing, Michigan 48824, USA and [2]Department of Geography, 1H Geography Building, Michigan State University, East Lansing, Michigan 48824, USA

Email: Sue C Grady* - gradys@msu.edu; Helen Enander - enander@msu.edu

* Corresponding author

Published: 18 February 2009

International Journal of Health Geographics 2009, **8**:10 doi:10.1186/1476-072X-8-10

Received: 30 September 2008
Accepted: 18 February 2009

This article is available from: http://www.ij-healthgeographics.com/content/8/1/10

Abstract

Background: Infant mortality is a major public health problem in the State of Michigan and the United States. The primary adverse reproductive outcome underlying infant mortality is low birthweight. Visualizing and exploring the spatial patterns of low birthweight and infant mortality rates and standardized incidence and mortality ratios is important for generating mechanistic hypotheses, targeting high-risk neighborhoods for monitoring and implementing maternal and child health intervention and prevention programs and evaluating the need for health care services. This study investigates the spatial patterns of low birthweight and infant mortality in the State of Michigan using automated zone matching (AZM) methodology and minimum case and population threshold recommendations provided by the National Center for Health Statistics and the US Census Bureau to calculate stable rates and standardized incidence and mortality ratios at the Zip Code (n = 896) level. The results from this analysis are validated using SaTScan. Vital statistics birth (n = 370,587) and linked infant death (n = 2,972) records obtained from the Michigan Department of Community Health and aggregated for the years 2004 to 2006 are utilized.

Results: For a majority of Zip Codes the relative standard errors (RSEs) of rates calculated prior to AZM were greater than 20%. Spurious results were the result of too few case and birth counts. Applying AZM with a target population of 25 cases and minimum threshold of 20 cases resulted in the reconstruction of zones with at least 50 births and RSEs of rates 20–22% and below respectively, demonstrating the stability reliability of these new estimates. Other AZM parameters included homogeneity constraints on maternal race and maximum shape compactness of zones to minimize potential confounding. AZM identified areas with elevated low birthweight and infant mortality rates and standardized incidence and mortality ratios. Most but not all of these areas were also detected by SaTScan.

Conclusion: Understanding the spatial patterns of low birthweight and infant deaths in Michigan was an important first step in conducting a geographic evaluation of the State's reported high infant mortality rates. AZM proved to be a useful tool for visualizing and exploring the spatial patterns of low birthweight and infant deaths for public health surveillance. Future research should also consider AZM as a tool for health services research.

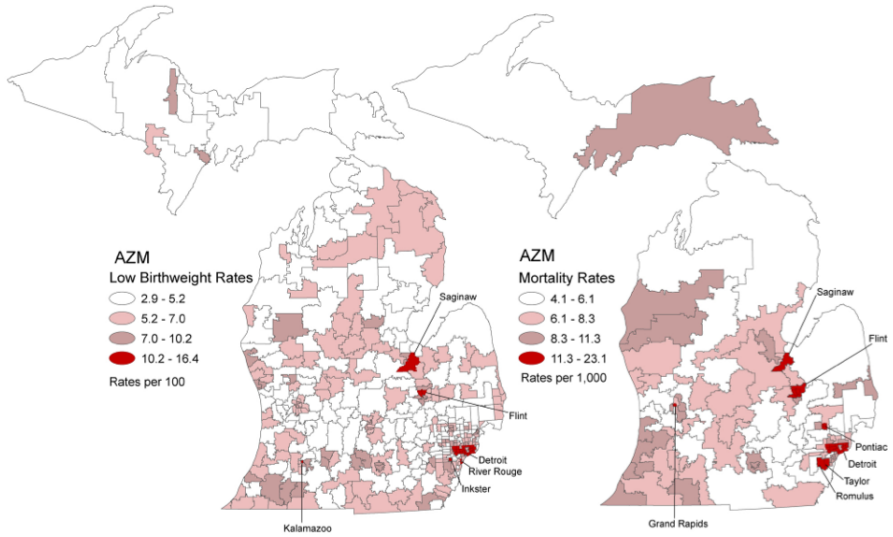


Figure 3
Maps of low birthweight incidence rates and infant mortality rates following AZM, Michigan 2004 to 2006.

not as rare an event as infant mortality; therefore, the size of zones created to meet the TP or minimum case threshold for low birthweight are smaller or more compact than zones created for infant deaths. In future studies these smaller zones may be useful for generating mechanistic hypotheses. For example, in this study we used maternal race as our homogeneity constraints, thus the zones in which low birthweight rates are calculated may be relatively homogeneous on this maternal characteristic in addition to being compact. Future studies of low birthweight that use environmental constraints may identify common risk factors associated with certain living environments. Previous studies [7,12] have shown the usefulness of AZM methodology to identify neighborhood social and built environmental risk factors for poor health outcomes as well as protective characteristics that lead to healthy living [16].

Table 4: SaTScan clusters of low birthweight, Michigan 2002–2004.

Cluster	Relative Risk	Log-Likelihood Ratio	p-value
Detroit	2.0	543.0	0.001
Detroit	1.8	211.1	0.001
Flint	1.8	211.1	0.001
Saginaw	1.7	33.7	0.001
Inkster	1.7	18.9	0.001
River Rouge	1.7	17.8	0.001
Pontiac	1.5	15.2	0.001
Kalamazoo	2.0	14.4	0.001

For infant mortality however, the size of the zones created to meet the TP or minimum case threshold for infant deaths were relatively large making it difficult to explore underlying mechanisms. Instead these zones may be better used to understand the demand for maternal and infant health care services, including neonatal-perinatal health care. Adding health care services as a homogeneity constraint will create zones that are homogeneous within (e.g., zones with high versus low utilization of services) and heterogeneous between on this constraint. Overlaying the health care services onto these zones may be useful in identifying areas that have services but lack utilization versus areas that do not have services but are in need. In future research we will also assess the need for health services within administrative units such as counties in case

hypotheses at a finer geographic scale the birth and linked infant death records would need to be geocoded and aggregated to the census block, block group or census tract. The authors decided that from a surveillance perspective, the Zip Code level of analysis would be conduced first while the records are being geocoded and thereafter, surveillance at a finer geographic scale could follow.

Automated Zoning Methodology

Automated zoning was implemented using the Automated Zone Matching (AZM) 1.0.0 software written by David Martin [37]. This software is freeware and available for public use on David Martin's website http:www2.geog.soton.ac.uk/users/martindj/davehome/soft ware.htm[37]. This software incorporates the principles of automated zone design originally conceptualized by Oppenshaw [38]. In preparation for AZM analysis the ESRI Zip Code boundary file [36] was converted to an ArcInfo coverage and island polygons were removed and slivers and overshoots were removed and undershoots were corrected. After the topology was check and corrected the arc and polygon attribute information was exported for use in the AZM software. The arc files were uploaded in AZM as intersection and contiguity files. These files comprised the Zip Code geography used in subsequent analyses.

The first parameter selected was the population target (PT) and/or minimum population threshold constraint(s). As noted previously, we used 25 low birthweight cases or infant deaths as our ideal target and 20 cases/deaths as our minimum threshold. Thus, all new zones created had at least 20 cases/deaths from which to calculate stable rates (i.e., rates with RSE 20%, respectively). AZM functions by minimizing the squared difference between the target number of cases/deaths and the number of cases/deaths in each Zip Code [37]. Thus if every Zip Code contained exactly 25 cases/deaths this constraint would reach zero. The minimum population threshold was 50 births. This threshold was not specified in AZM but resulted after the aggregation of low birthweight cases/deaths. These TP parameters were held constant throughout the analysis.

The second parameter selected was the shape statistic defined as: $\sum q_k^2 / A_k$

Where q_k is the perimeter of zone k and A_k is its area. AZM functions to minimize the perimeter squared divided by area, which maximizes shape compactness on zones. As outlined in Martin [37] software documentation "irregular shapes may have longer perimeters in relation to their area; thus, squaring the perimeter makes highly irregular shapes less attractive when this constraint is in operation."

The third parameter selected was the homogeneity constraint. The homogeneity constraint promotes homogeneity within zones and heterogeneity between zones by encouraging the aggregation of similar values. In this study we used mother's race (i.e., African Americans versus all others) as our homogeneity constraint. Maternal race was added because of the high levels of racial residential segregation in Michigan's cities and the desire to capture spatial-racial disparities in low birthweight and infant mortality. The IAC for these two racial groups was obtained for each k category and overall K across all categories (k1, k2 = K categories). An IAC of 0.5 implied a reasonable degree of homogeneity [37].

The IAC was calculated as:

$$\delta_k = \frac{\frac{1}{M-1}\sum_{g=1}^{M} N_g(P_{kg}-P_k)^2}{(\bar{N}^*-1)P_k(1-P_k)} - \frac{1}{(\bar{N}^*-1)}$$

Where, $\bar{N}^*$ was the mean case/death size of the M number of Zip Codes, with an adjustment [39] to take into account variation in the case/death size of units. It was expected that $\bar{N}^*$ would be very close to $\bar{N}$. N_g was the case/death size of Zip Code g; P_k was the overall proportion of cases/deaths in category k, and P_{kg} was the proportion in category k in Zip Code g. Thus, the IAC was approximately the ratio of the Zip Code variance to the maternal level variance, and this ratio was divided by the mean case/death size.

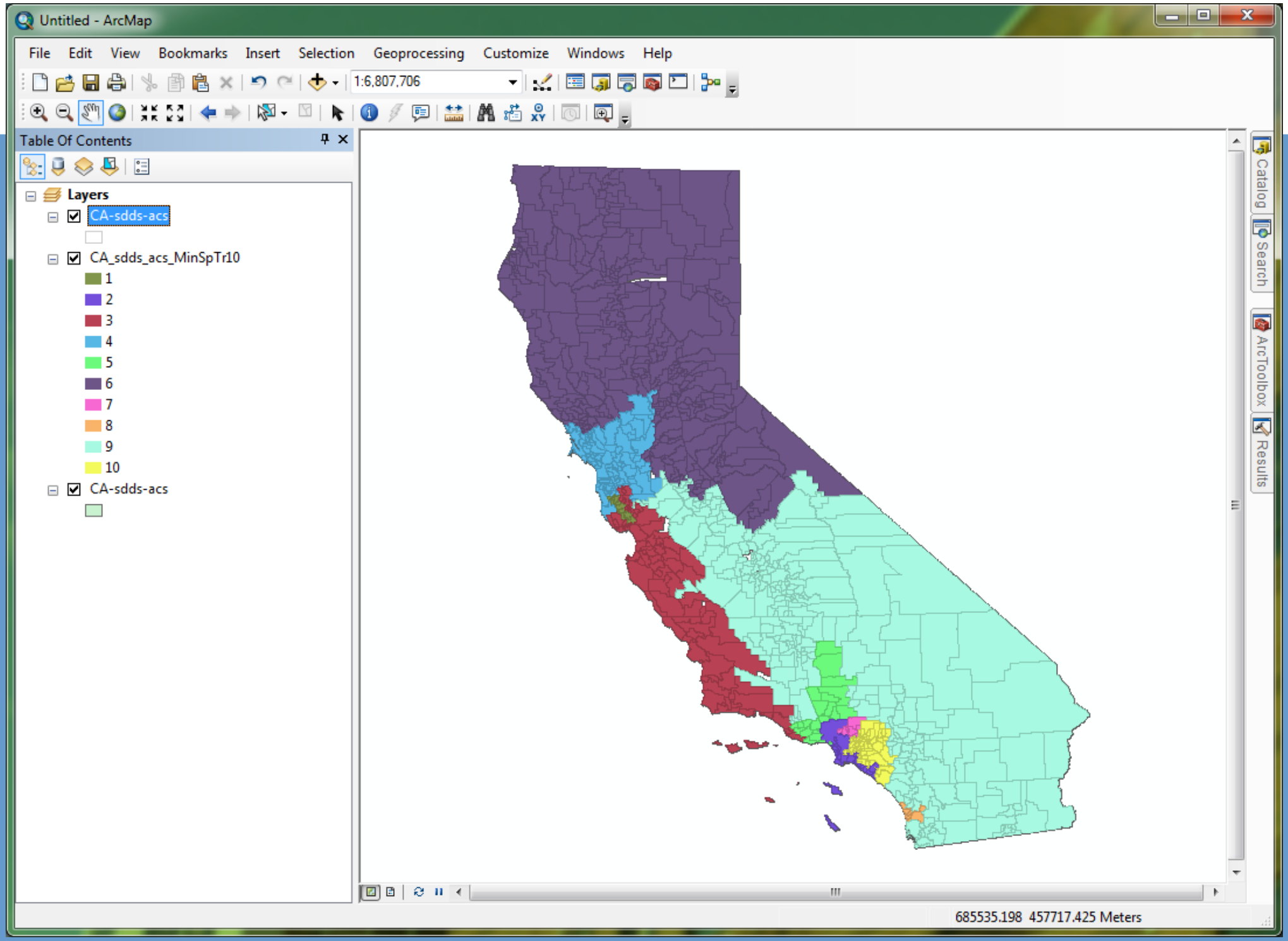
After δ_k was calculated for each race k category the overall IAC, δ for all categories was calculated as:

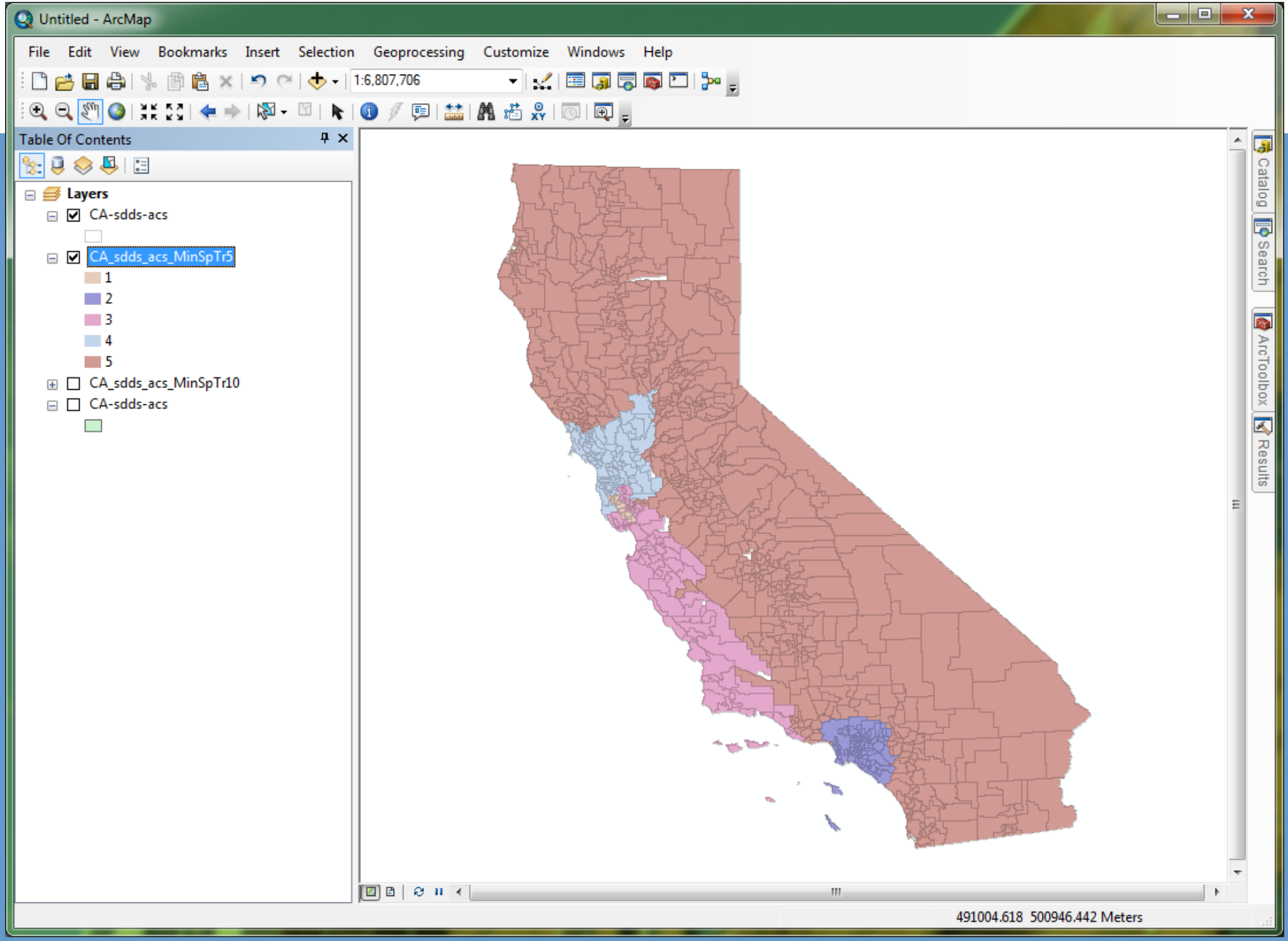
$$\text{IAC} = \frac{1}{k-1}\sum_{k=1}^{K}(1-P_k)\delta_k$$

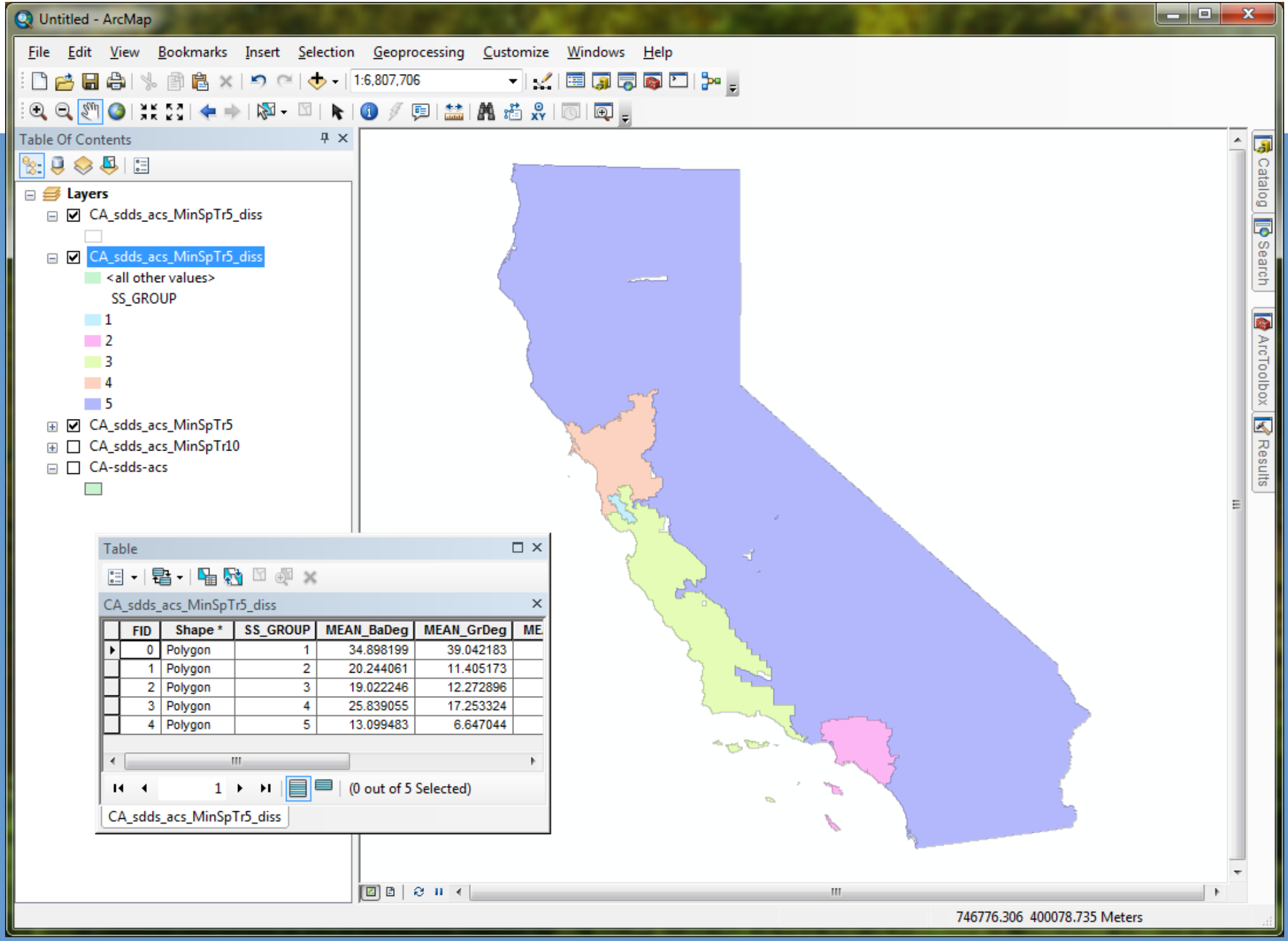
An optional parameter that may be changed for user preferences was the "random number initialization value," which sets the seed value for the pseudorandom number generator prior to the IRA run. Keeping this value constant during multiple restarts of AZM would result in the same zone designs. Changing this seed value would result in a different sequence of pseudorandom decisions, which would alter the zone designs. For the purposes of this surveillance research we kept this parameter constant to eliminate differences between zone designs. With these model parameters the AZM analysis was conducted by running 50 program restarts with 100 iterations each taking the run (i.e., zone design) with the most compact shape, the strongest IAC and lastly the best TP statistic.

Minimum Spanning Tree

- Similar to hierarchical grouping
  - Works with spatial data, via connectivity graph (neighbors)
- Top down methodology
  - All things begin as "connected" and groups are broken off
- Integrates both spatial similarity (weights) and attribute similarity


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:6,807,706
Table Of Contents
Layers
CA-sdds-acs
CA_sdds_acs_MinSpTr10
1
2
3
4
5
6
7
8
9
10
CA-sdds-acs
Catalog
Search
ArcToolbox
Results
685535.198 457717.425 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:6,807,706
Table Of Contents
Layers
CA-sdds-acs
CA_sdds_acs_MinSpTr5
1
2
3
4
5
CA_sdds_acs_MinSpTr10
CA-sdds-acs
Catalog
Search
ArcToolbox
Results
491004.618 500946.442 Meters


Untitled - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:6,807,706
Table Of Contents
Layers
CA_sdds_acs_MinSpTr5_diss
CA_sdds_acs_MinSpTr5_diss
<all other values>
SS_GROUP
1
2
3
4
5
CA_sdds_acs_MinSpTr5
CA_sdds_acs_MinSpTr10
CA-sdds-acs
Table
CA_sdds_acs_MinSpTr5_diss
FID Shape * SS_GROUP MEAN_BaDeg MEAN_GrDeg ME
0 Polygon 1 34.898199 39.042183
1 Polygon 2 20.244061 11.405173
2 Polygon 3 19.022246 12.272896
3 Polygon 4 25.839055 17.253324
4 Polygon 5 13.099483 6.647044
1
(0 out of 5 Selected)
CA_sdds_acs_MinSpTr5_diss
Catalog
Search
ArcToolbox
Results
746776.306 400078.735 Meters

How Many Groups?

- How many groups is the right number of groups?
  - This question has plagued statisticians since the decision to group observations
  - Keep in mind that the fit of the solution will generally increase as number of groups increases

How Many Groups?

- Can be "problem" driven
  - e.g., group observations (Zip Codes) by their nearest hospital
    - Number of groups is number of hospitals
- Often, subjective
  - How many groups "seem" correct for the data/problem
    - e.g., expert opinion

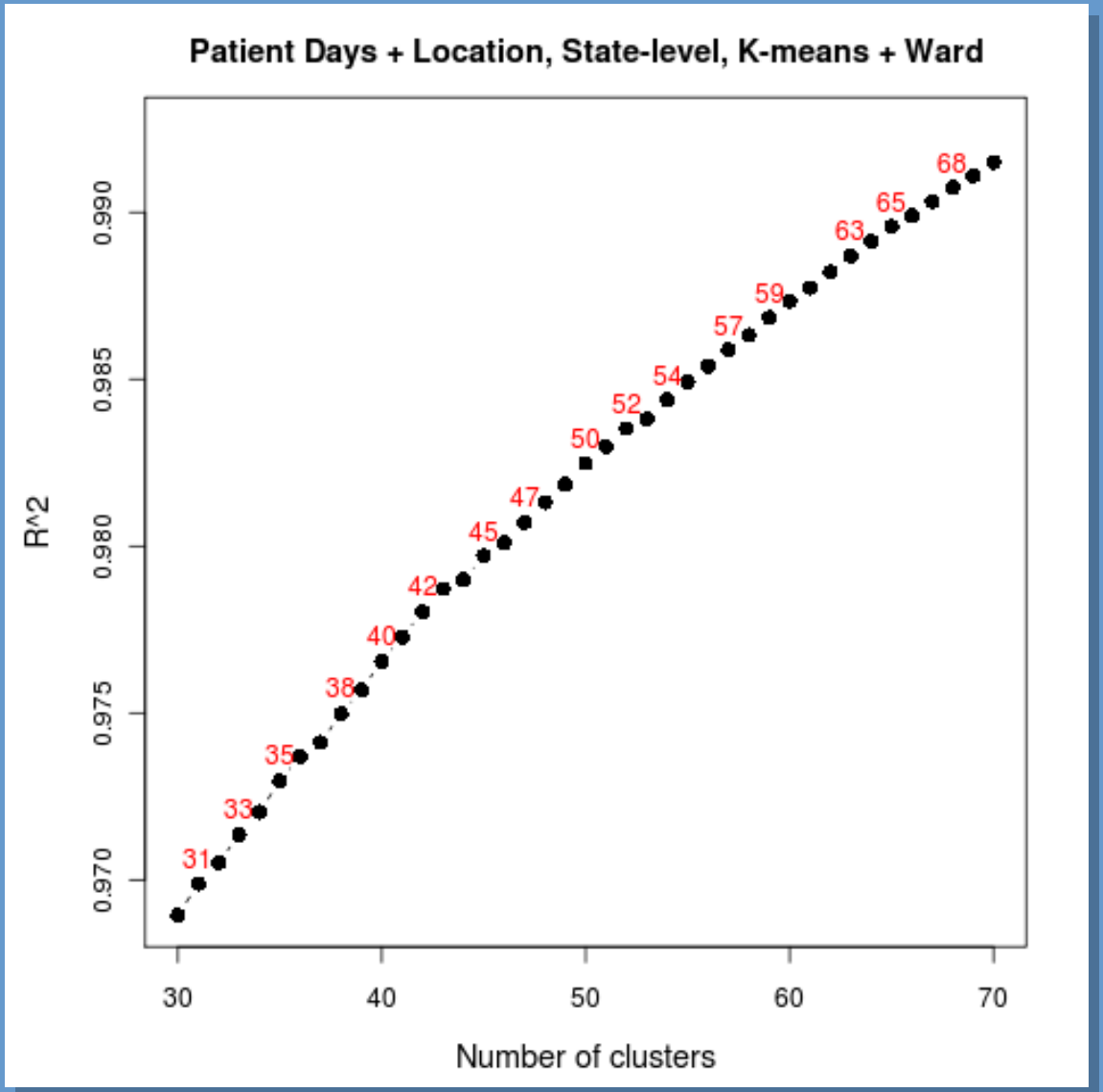
How Many Groups?

- Evaluating Group Solutions
  - Regression-like statistics
  - Grouping is somewhat comparable to fitting a regression line through a set of points
    - Center locations of groups serve as the regression line
    - Calculate basic regression statistics
      - R^2, F score
      - Data has a variance, and the groups explain a portion of it
      - Within SS, Between SS, Total SS

How Many Groups?

- Evaluating Group Solutions
  - Regression-like statistics
  - Incremental F-score (thanks to Dr. Bruce Pigozzi)
    - The gain in model significance (measured as F score) from the addition of a single additional group
    - When plotted for a set of solutions, peaks correspond to solutions that produce "value" with the additional group

R^2 of Group Solutions

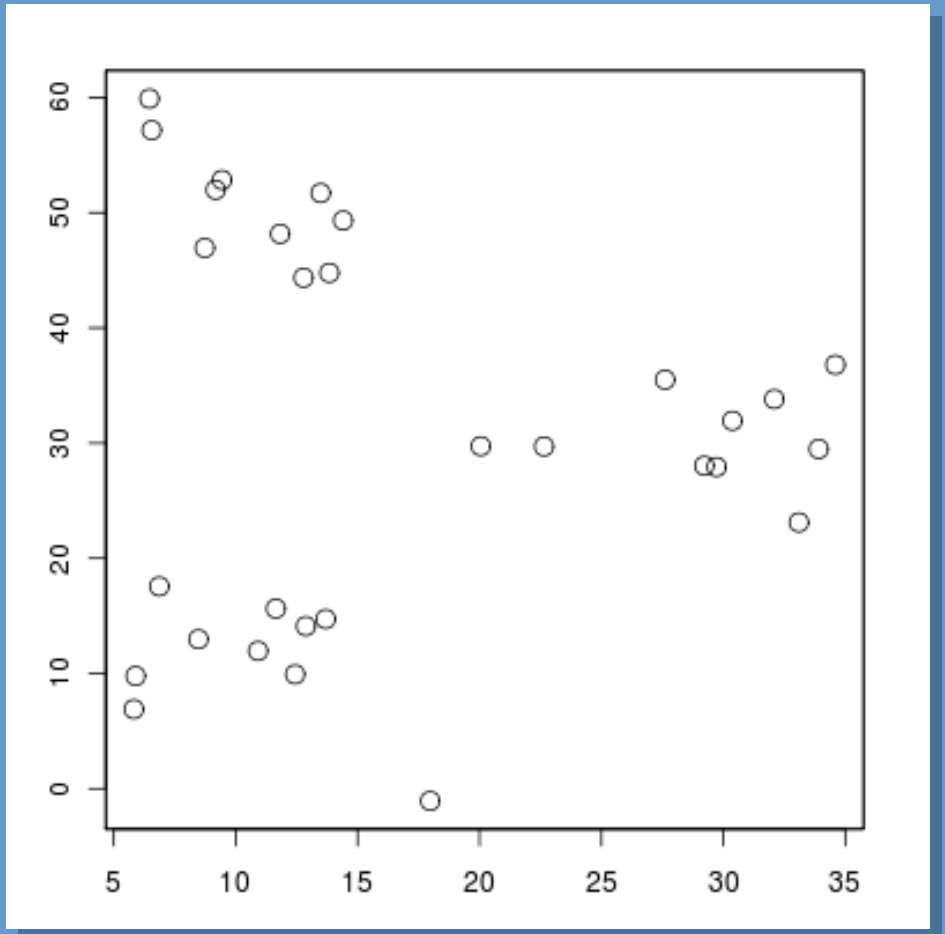


Incremental F score (*incF*)

$$incF_i = \frac{\left(\frac{R^2_i - R^2_{i-1}}{k_i - k_{i-1}} \right)}{\left(\frac{1 - R^2_i}{n - (k_i - 1)} \right)}$$

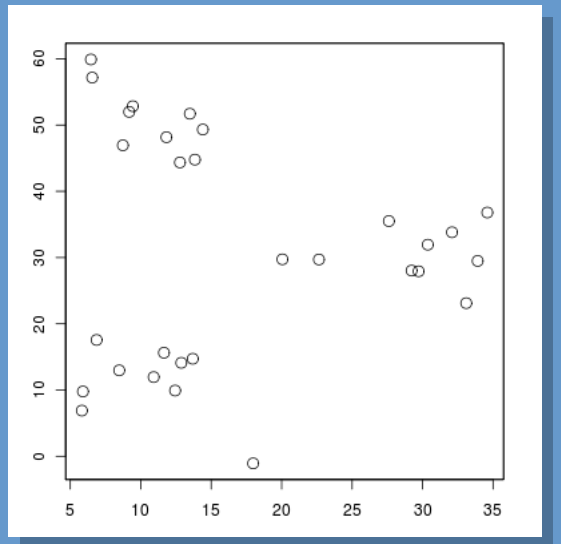
- Gain in "model fit", while also penalizing for "adding" another group to explain the data

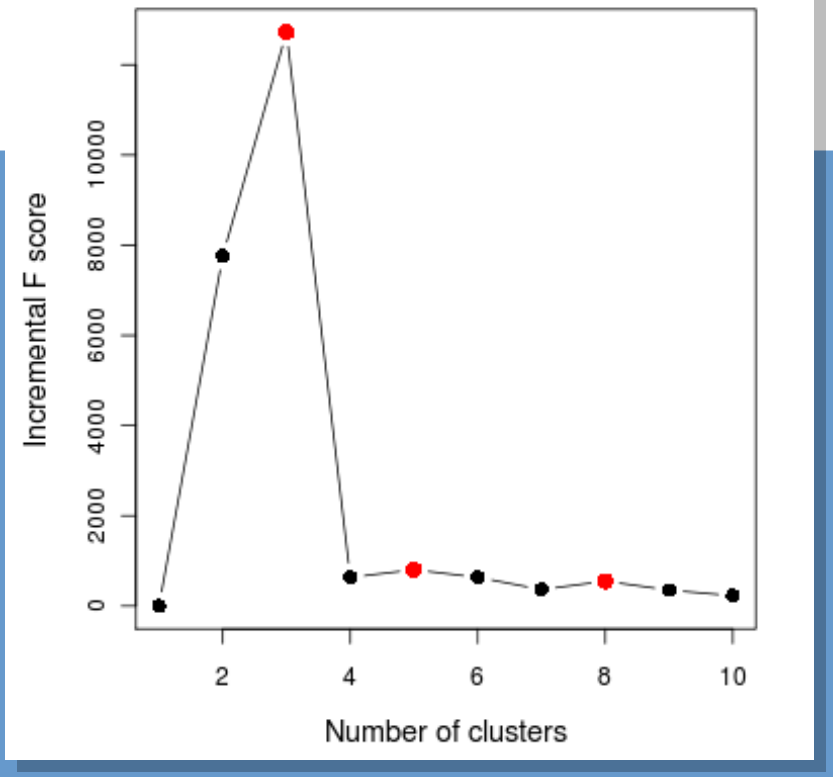
incF Example

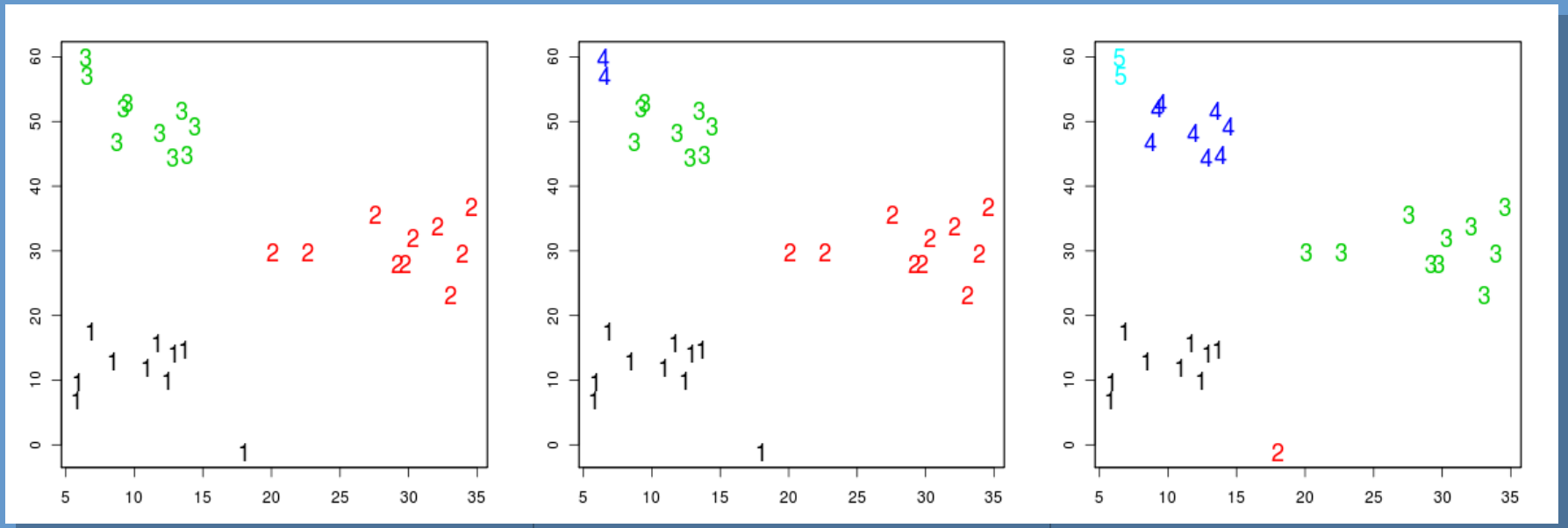


How many groups?

incF Example







Clustering Availability

- In ArcGIS
  - K-means
  - Minimum Spanning Tree
- In GeoDa
  - K-means
  - Hierarchical
  - K-means (with coordinates)
  - Minimum Spanning Tree
  - More!(https://geodacenter.github.io/documentation.html)

Keywords

- Grouping
- Regionalization
- Statistical, spatial, hybrid grouping
- Hierarchical, partitional
- Thiessen polygons
- Space as an attribute