Chapter 13 – Descriptive summary, design, and inference

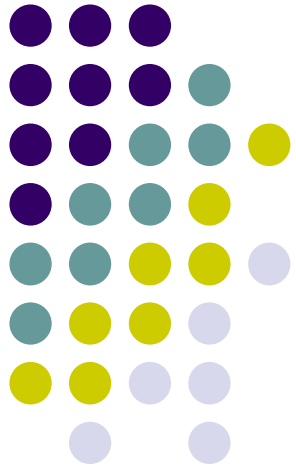
§ 1 More spatial analysis

§ 2 Descritive summaries

§ 3 Optimization

§ 4 Hypothesis testing

§ 5 Conclusion



Learning Objectives

- After reading this chapter you will be able to:
  - Data mining, a new form of analysis that is enabled by GIS and by vast new supplies of data
  - The concept of summarizing a pattern in a few simple statistics
  - Methods that support decisions by enlisting GIS to search automatically across thousands or millions of options
  - The concept of a hypothesis, and how to make inferences from small samples to larger populations

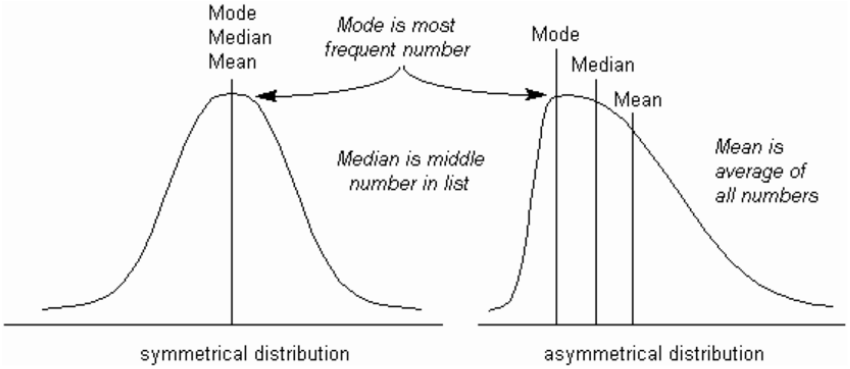
§ 1 More spatial analysis

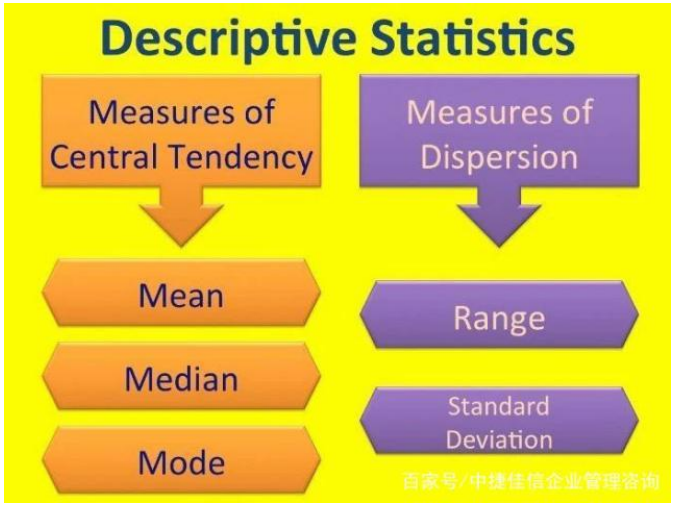
- use more sophisticated framework
- turns to summaries, optimization methods, and inferential methods
- Many date from well before the era of GIS, and the advent of geographic databases and cheap computing technology has turned what were once esoteric methods buried in the academic literature into practical ways of solving everyday problems
- *Data mining* is used to detect anomalies and patterns in vast archives of digital data

§ 2 Descriptive summaries

§ 2.1 Centers

- Numerical Summaries :mean（平均值）, median（中位数）, mode(众数)， which attempt to create a summary description of a series of numbers in the form of a single number
- In spatial case, centers are the two-dimensional equivalent of the mean





§ 2 Descriptive summaries

§ 2.1 Centers（cont.）

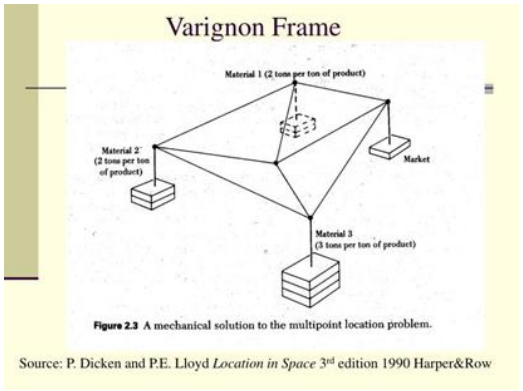
Mean Center (there is also a median center called Manhattan Center)	Identifies the geographic center (or the center of concentration) for a set of features ****Mean sensitive to outliers****	Simply the mean of the X coordinates and the mean of the Y coordinates for a set of points $$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad , \quad \overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$
Weighted Mean Center	Like the mean but allows weighting by an attribute.	Produced by weighting each X and Y coordinate by another variable (Wi) $$\overline{X} = \frac{\sum_{i=1}^{n} wixi}{\sum_{i=1}^{n} wi} \quad \overline{Y} = \frac{\sum_{i=1}^{n} wixi}{\sum_{i=1}^{n} wi}$$

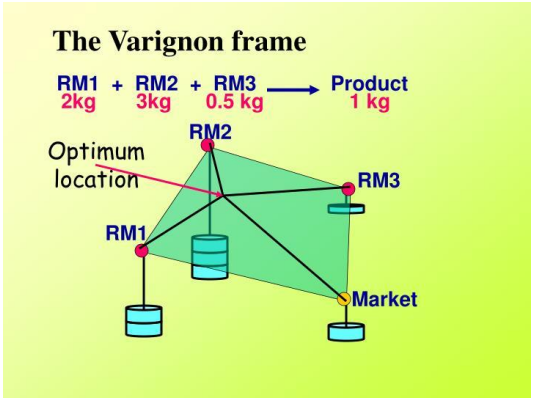
§ 2 Descriptive summaries

§ 2.1 Centers（cont.）

- Minimum aggregate travel（MAT），location that minimizes the sum of distances
- Many methods of spatial analysis are used to make design decisions

Central Distance	Identifies the most centrally located feature for a set of points, polygon(s) or line(s)	Point with the shortest total distance to all other points is the most central feature $D = \sum_{i=1}^{n} \sum_{i=1}^{n-1} \sqrt{(xj - xi)^2 + (yj - yi)^2}$ $Dcentral = \text{minimum}\,(D)$





The Varignon Frame（瓦里翁框架）

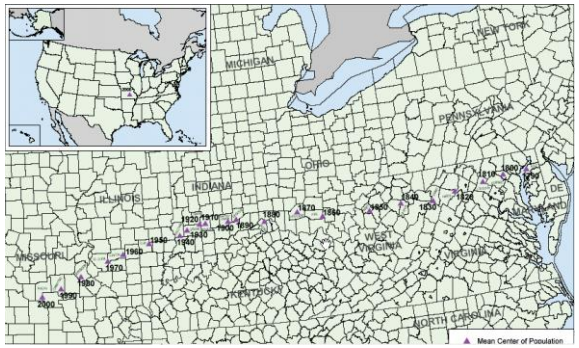


Figure 15.3 The march of the US population westward since the first census in 1790, as summarized in the population centroid (Source: US Bureau of the Census)

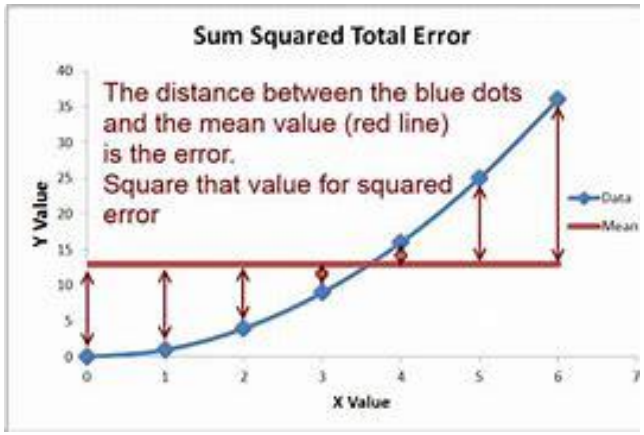
§ 2 Descriptive summaries

§ 2.2 Dispersion(离散)

- *Range, Standard Deviation, Coefficient of Variation*
- *standard deviation*： where *n* is the number of numbers, *s* is the standard deviation, x_i refers to the i_{th} observation, and *x* is the mean of the observations：

$$s = \sqrt{\sum_i w_i (x_i - \overline{x})^2 \Big/ \sum_i w_i} \qquad S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

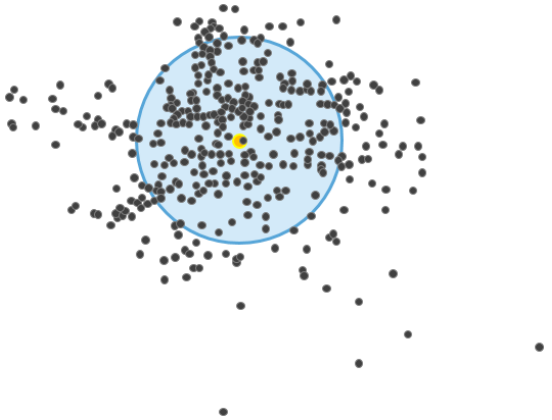
- In spatial case, Mean distance from the center is a useful summary of dispersion



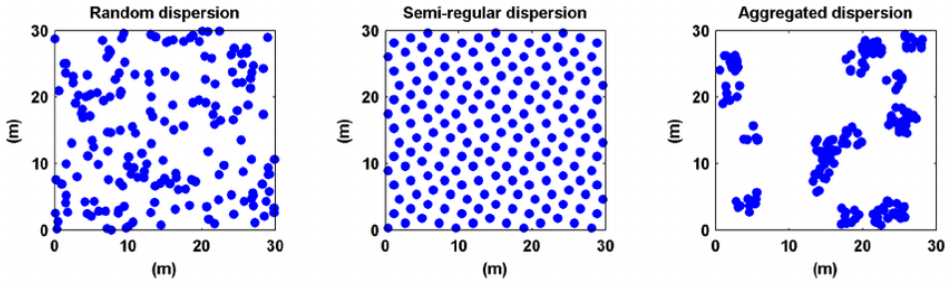
Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

§ 2 Descriptive summaries

Spatial Descriptive Statistic	Description	Calculation
Standard Distance	Measures the degree to which features are concentrated or dispersed around the geometric mean center The greater the standard distance, the more the distances vary from the average, thus features are more widely dispersed around the center Standard distance is a good **single** measure of the dispersion of the points around the mean center, but it doesn't capture the **shape** of the distribution.	Represents the standard deviation of the distance of each point from the mean center: $SD = \sqrt{\frac{\sum_{i=1}^{n}\left(xi-\overline{X}\right)^2}{n} + \frac{\sum_{i=1}^{n}\left(yi-\overline{Y}\right)^2}{n}}$ Where *xi* and *yi* are the coordinates for a feature and $\overline{X}$ and $\overline{Y}$ are the mean center of all the coordinates. **Weighted SD** $SD_w = \sqrt{\frac{\sum_{i=1}^{n} w_i(x_i - \bar{X}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(y_i - \bar{Y}_w)^2}{\sum_{i=1}^{n} w_i}}$ Where *xi* and *yi* are the coordinates for a feature and $\overline{X}$ and $\overline{Y}$ are the mean center of all the coordinates. *wi* is the weight value.



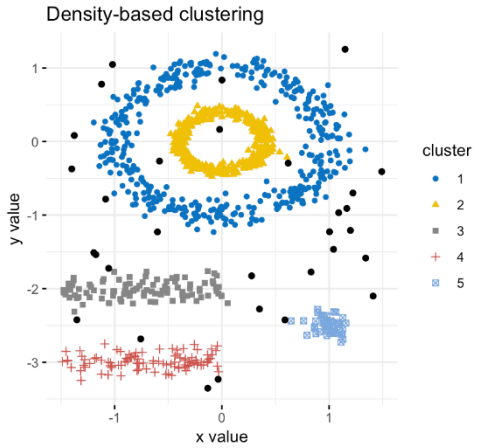
SD output

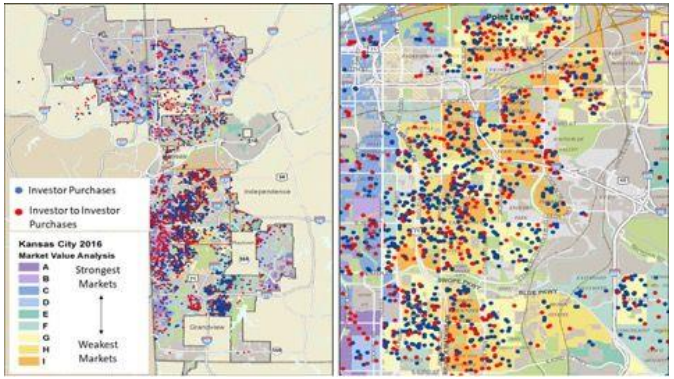


§ 2 Descriptive summaries

§ 2.3 Measures of pattern: unlabeled points

- Point patterns can be identified as clustered, dispersed, or random
- Clustering can be produced by two distinct mechanisms, identified as first-order and second-order
  - *First-order* processes involve points being located independently, but may still result in clusters because of varying point density
  - *Second-order* processes involve interaction between points, and lead to clusters when the interactions are attractive in nature, and dispersion when they are competitive or repulsive
- In general it is not possible to determine whether a given clustered point pattern was created by varying density factors, or by interactions
- On the other hand, dispersed patterns can only be created by second-order processes





§ 2 Descriptive summaries

§ 2.3 Measures of pattern: unlabeled points（cont.）

§ 2.3.1 Ripley's K-function

- Determines whether features, or the values associated with features, exhibit statistically significant clustering or dispersion over a range of distances
- K(d) is defined as the expected number of points within a distance d of an arbitrarily chosen point, divided by the density of points per unit area

$$\hat{L}(d) = \sqrt{K(d)/\pi}$$

- L(d) will equal d for all d in a random pattern, i.e., $K(d)=\pi d^2$
- Clustering at certain distances is indicated by departures of L above the line and dispersion by departures below the line

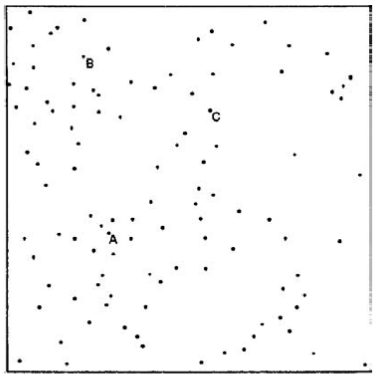


Figure 15.8 Point pattern of individual tree locations. A, B, and C identify the individual trees analyzed in Figure 15.9 (*Source*: Getis A. and Franklin J. 1987 'Second-order neighborhood analysis of mapped point patterns.' *Ecology* **68**(3): 473–477)

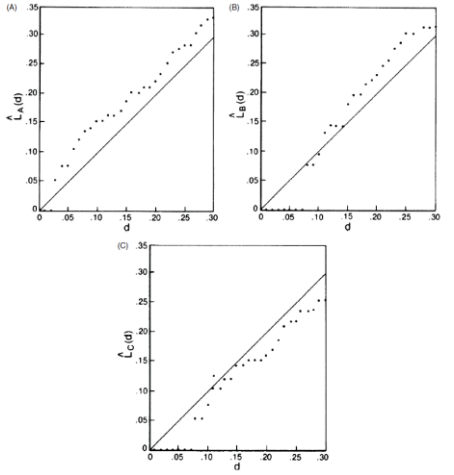


Figure 15.9 Analysis of the local distribution of trees around three reference trees in Figure 15.8 (see text for discussion) (*Source*: Getis A. and Franklin J. 1987 'Second-order neighborhood analysis of mapped point patterns.' *Ecology* **68**(3): 473–477)

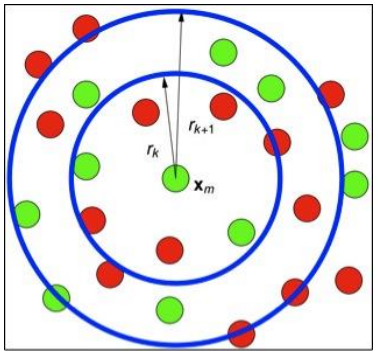
§ 2 Descriptive summaries

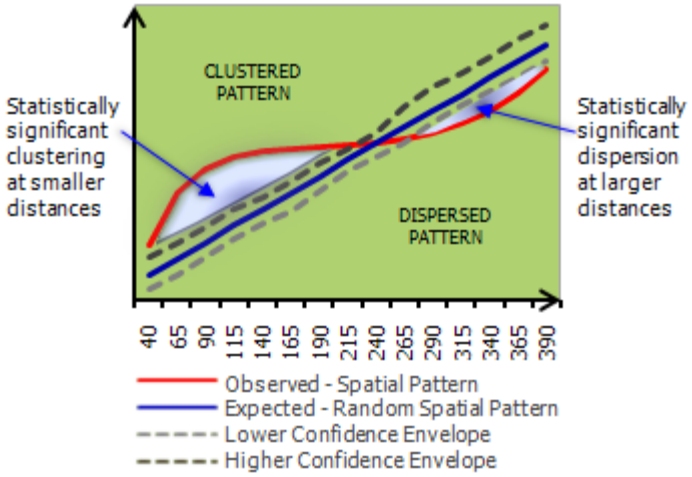
§ 2.3 Measures of pattern: unlabeled points（cont.）

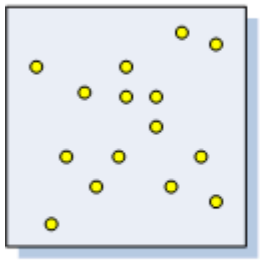
- Used as Multi-Distance Spatial Cluster Analysis

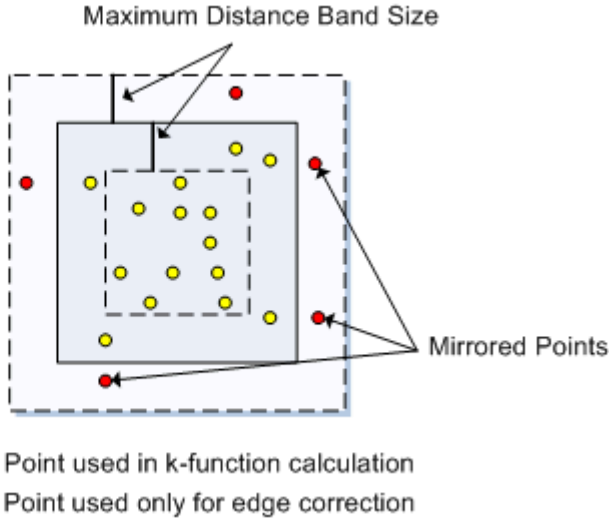
$$L(d) = \sqrt{\frac{A\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} k(i,j)}{\pi N(N-1)}}$$

- where A is area, N is the number of points, d is the distance and k(i, j) is the weight, which is 1 when the distance between i and j is less than or equal to d and 0 when the distance between i and j is greater than d









§ 2 Descriptive summaries

§ 2.3 Measures of pattern: unlabeled points（cont.）

§ 2.3.2 Average Nearest Neighbor(ANN)

- An average nearest neighbor (ANN) analysis measures the average distance from each point in the study area to its nearest point

The Average Nearest Neighbor ratio is given as:

$$ANN = \frac{\bar{D}_O}{\bar{D}_E} \quad (1)$$

where $\bar{D}_O$ is the observed mean distance between each feature and its nearest neighbor:

$$\bar{D}_O = \frac{\sum_{i=1}^{n} d_i}{n} \quad (2)$$

and $\bar{D}_E$ is the expected mean distance for the features given in a random pattern:

$$\bar{D}_E = \frac{0.5}{\sqrt{n/A}} \quad (3)$$

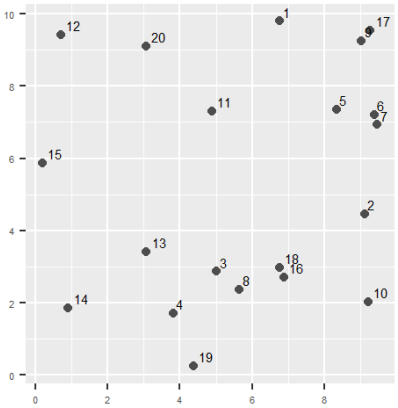
In the above equations, d_i equals the distance between feature i and it's nearest neighboring feature, n corresponds to the total number of features, and A is the area of a minimum enclosing rectangle around all features, or it's a user-specified Area value.

The average nearest neighbor z-score for the statistic is calculated as:

$$z = \frac{\bar{D}_O - \bar{D}_E}{SE} \quad (4)$$

where:

$$SE = \frac{0.26136}{\sqrt{n^2/A}} \quad (5)$$



From	To	Distance	From	To	Distance
1	9	2.32	11	20	2.55
2	10	2.43	12	20	2.39
3	8	0.81	13	4	1.85
4	19	1.56	14	13	2.67
5	6	1.05	15	12	3.58
6	7	0.3	16	18	0.29
7	6	0.3	17	9	0.37
8	3	0.81	18	16	0.29
9	17	0.37	19	4	1.56
10	2	2.43	20	12	2.39

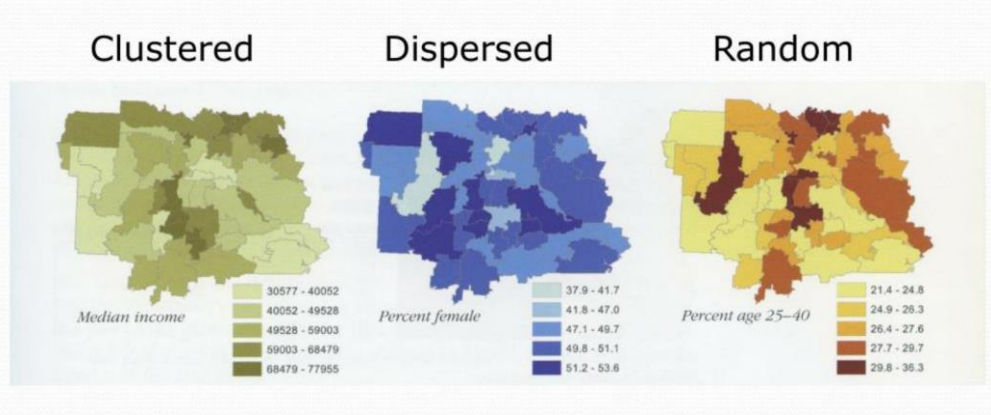
§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points or areas

§ 2.3.3 Moran statistic

- When features are differentiated, such as by values of some attribute, the typical question concerns whether the values are randomly distributed over the features, or whether high extreme values tend to cluster: high values surrounded by high values and low values surrounded by low values
- The Moran statistic is designed precisely for this purpose, to indicate general properties of the pattern of attributes, It distinguishes between
  - positively auto-correlated patterns, in which high values tend to be surrounded by high values, and low values by low values
  - random patterns, in which neighboring values are independent of each other
  - dispersed patterns, in which high values tend to be surrounded by low, and *vice versa*

Patrick Alfred Pierce Moran



§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points

- The Spatial Autocorrelation (Global Moran's I) tool measures spatial autocorrelation based on both feature locations and feature values simultaneously. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I Index value and both a a z-score and p-value to evaluate the significance of that Index

The Moran's I statistic for spatial autocorrelation is given as:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad (1)$$

where z_i is the deviation of an attribute for feature i from its mean $(x_i - \bar{X})$, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features, and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \quad (2)$$

The z_I-score for the statistic is computed as:

$$z_I = \frac{I - \mathrm{E}[I]}{\sqrt{\mathrm{V}[I]}} \quad (3)$$

where:

$$\mathrm{E}[I] = -1/(n-1) \quad (4)$$

$$\mathrm{V}[I] = \mathrm{E}[I^2] - \mathrm{E}[I]^2 \quad (5)$$

莫兰指数公式理解

$$I = \frac{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S_0}}{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

image.png

于是莫兰指数可以看成上下两个部分：

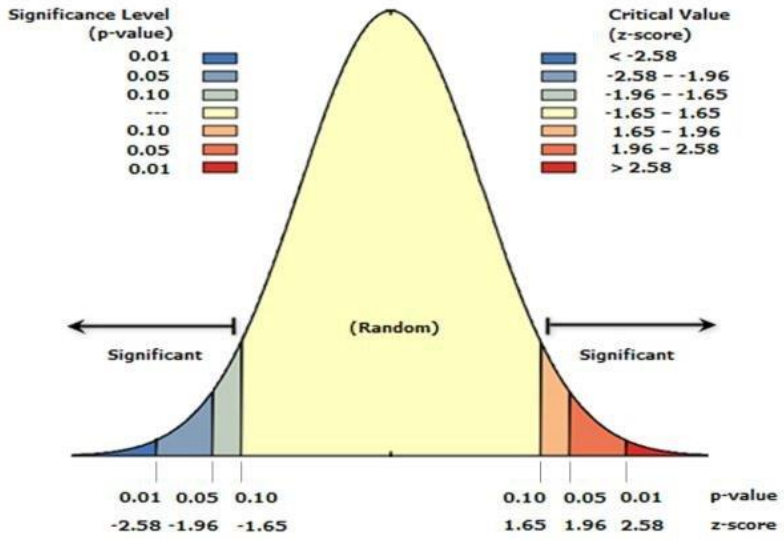
上部分（蓝框）：考虑地理相邻关系的数据X的方差

下部分（红框）：仅是简单的统计学关于数据X的方差

§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points

- The Most statistical tests begin by identifying a null hypothesis. The null hypothesis for the pattern analysis is Complete Spatial Randomness (CSR). The z-scores and p-values tell you whether you can reject that null hypothesis or not.
- The p-value is the probability that the observed spatial pattern was created by some random process. When the p-value is very small, it means it is very unlikely (small probability) that the observed spatial pattern is the result of random processes, so you can reject the null hypothesis
- Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below



§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points

- Local Moran statistic was developed by Luc Anselin(1995) as a *local indicator of spatial association* or LISA statistic

The Local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{X}) \quad (1)$$

where x_i is an attribute for feature i, $\bar{X}$ is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{X})^2}{n - 1} \quad (2)$$

with n equating to the total number of features.

The z_{I_i}-score for the statistics are computed as:

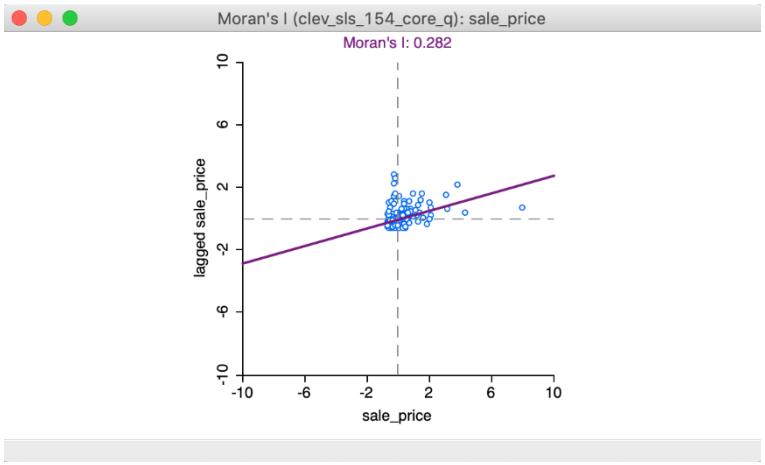
$$z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{\mathrm{V}[I_i]}} \quad (3)$$

where:

$$\mathrm{E}[I_i] = -\frac{\sum_{j=1, j \neq i}^{n} w_{ij}}{n - 1} \quad (4)$$

$$\mathrm{V}[I_i] = \mathrm{E}[I_i^2] - \mathrm{E}[I_i]^2 \quad (5)$$

Z_i	$\sum_{j \neq i}^{n} w_{ij} Z_j$	I_i
>0	>0	>0
<0	<0	>0
<0	>0	<0
>0	<0	<0



§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points

社会地理计算联合研究中心
Geocomputation Center for Social Science

中文 English

网站首页 | 中心简介 | 研究方向 | 项目介绍 | 科研成果 | 学术交流 | 相关资源 | 联系我们

中心概况 | 中心领导

中心主任

龚健雅 院士

龚健雅，武汉大学教授，博士生导师。现任武汉大学遥感信息工程学院院长、测绘遥感信息工程国家重点实验室主任。1957年4月生于江西省樟树市。1982年毕业于华东地质学院测量系，1992年于武汉测绘科技大学获博士学位。2011年当选中国科学院院士。国家杰出青年基金获得者、973项目首席科学家、国家自然科学基金创新群体学术带头人、国家测绘局科技领军人才、国务院第六届学科评议组测绘学科组召集人，国际摄影测量与遥感学会第六委员会主席。

中心共同主任

Professor Luc Anselin

Luc Anselin is the Stein-Freiler Distinguished Service Professor of Sociology and the College at the University of Chicago, where he is also Director of the Center for Spatial Data Science, Chair of the Committee on Geographical Sciences, and Senior Fellow NORC. Anselin received his PhD in Regional Science from Cornell University. He was elected to the U.S. National Academy of Sciences in 2008 and the American Academy of Arts and Sciences in 2011. He is also a Fellow of the Spatial Econometric Society and of the University Consortium for Geographic Information Science. In 2005 he was honored with the Walter Isard Prize, and in 2006, he received the prestigious Alonso Memorial Prize for Innovative Work in Regional Science.

鲍曙明 教授

美国克莱姆森大学经济学博士。现任中国数据研究所所长，武汉大学社会地理计算联合研究中心海外共同主任，历任密西根大学中国数据中心和空间数据中心主任、社会研究院（ISR）/高校政治与社会科学研究联盟（ICPSR）研究员、和中国研究中心研究员，江西师范大学鄱阳湖流域与湿地研究教育部重点实验室创始主任和华东理工大学城市与区域分析实验室创始主任等职。目前担任国际华人地理信息科学协会秘书长、中国留美经济学会常务执行主任、international Journal of Emerging Markets高级编辑等职。在地理信息系统，区域经济，空间数据分析等领域发表了80多篇文章。

§ 2 Descriptive summaries

§ 2.3 Measures of pattern: labeled points

- Given a set of weighted features, identifies statistically significant hot spots(High-high), cold spots(Low-low), and spatial outliers

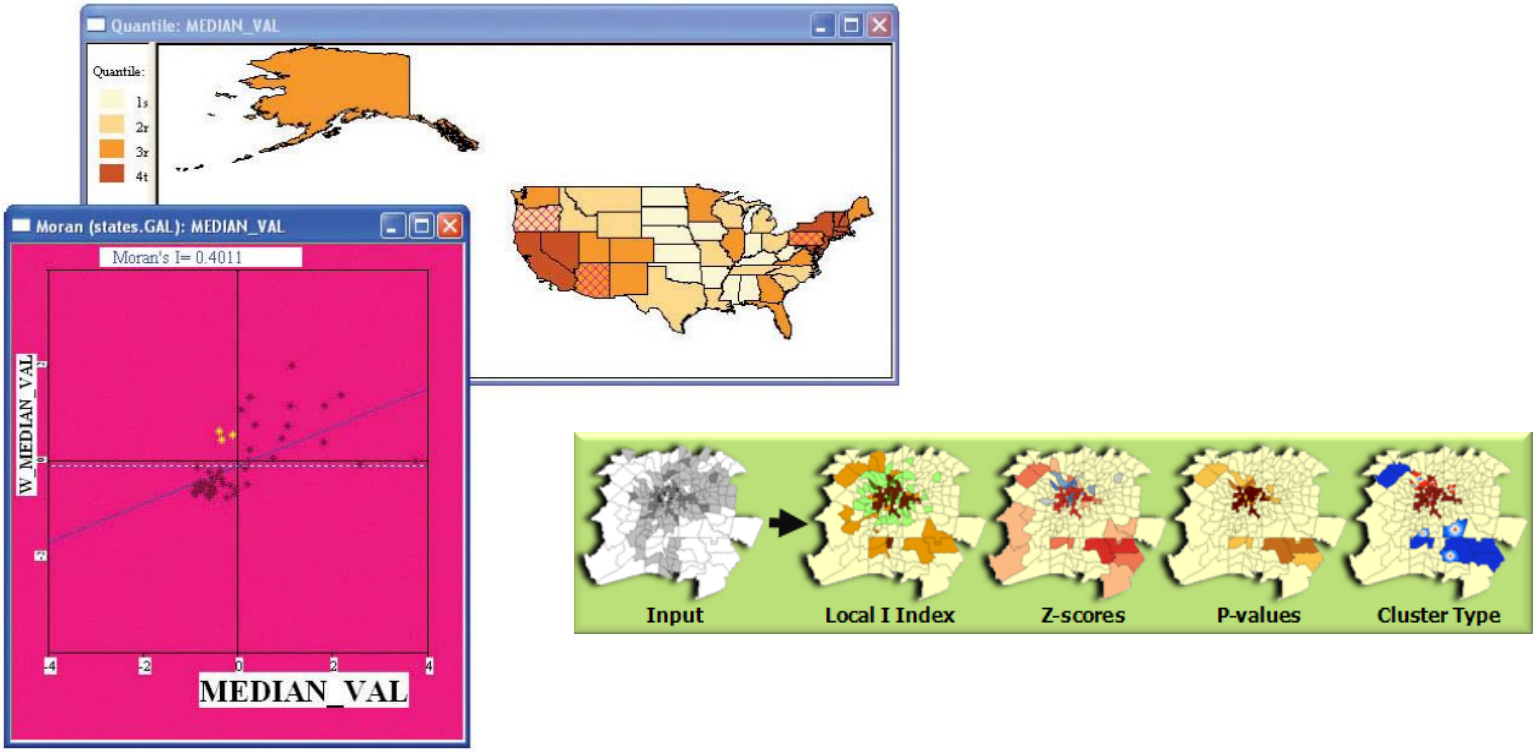


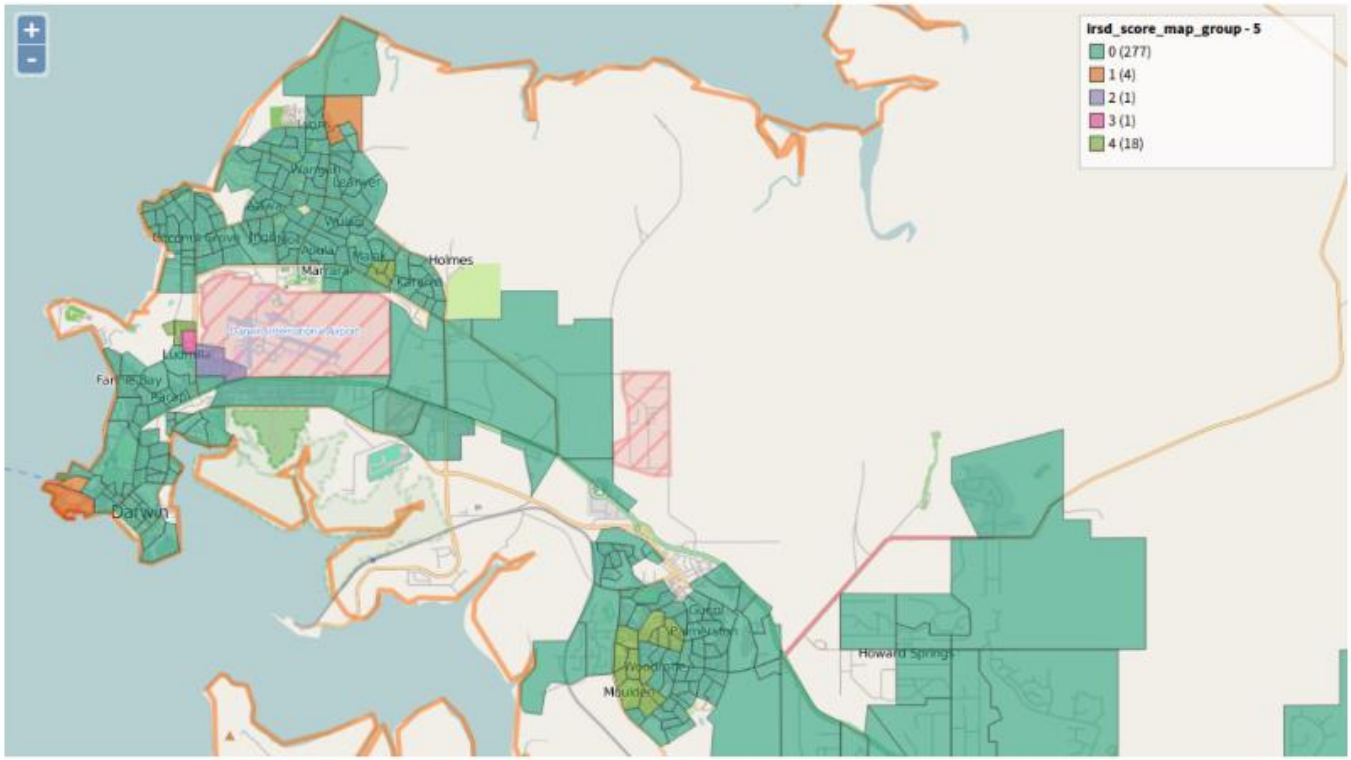
Figure 15.10 The Local Moran statistic, applied to describe local aspects of the pattern of housing value among US states. In the map window the states are colored according to median value, with the darker shades corresponding to more expensive housing. In the scatterplot window the median value appears on the *x* axis, while on the *y* axis is the weighted average of neighboring states. The three points colored yellow are instances where a state of below-average housing value is surrounded by states of above-average value. The windows are linked (see Figure 14.8 for details of this GeoDa software), and the three points are identified as Oregon, Arizona, and Pennsylvania. The global Moran statistic is also shown (+0.4011, indicating a general tendency for clustering of similar values)

§ 2 Descriptive summaries

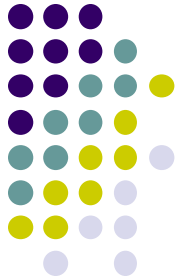
§ 2.3 Measures of pattern: labeled points

- Examples

- Orange indicates areas of high IRSD index scores surrounded by other high scores (low disadvantage),
- Purple indicates high scores surrounded by low scores,
- Pink indicates low scores surrounded by high scores,
- Light-green indicates low areas surrounded by low scores, and
- Dark-green represents areas with no statistical significance.



§ 3 Optimization

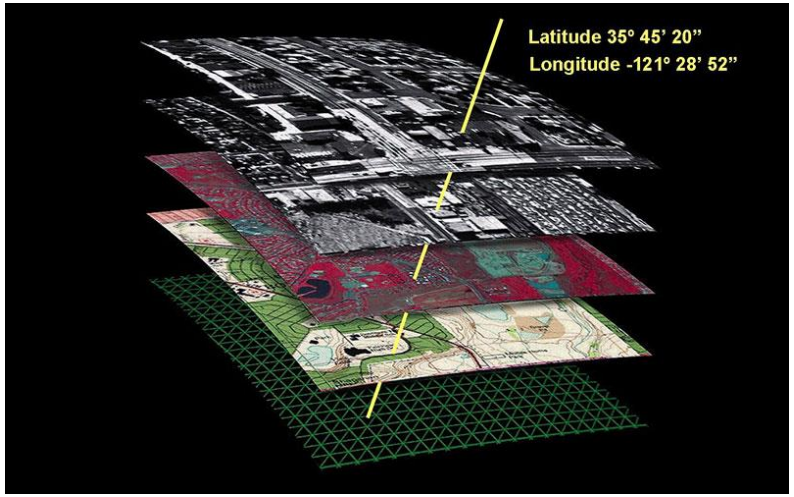


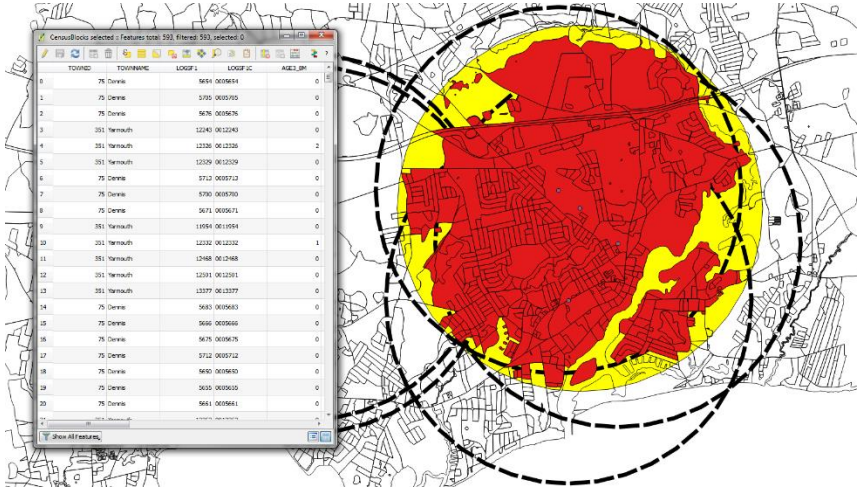
- The basic idea is to analyze patterns not for the purpose of discovering anomalies or testing hypotheses about process, as in previous sections, but with the objective of creating improved designs
- Normative methods apply well-defined objectives to the design of systems
- Design methods are often implemented as components of systems built to support decision making – so-called *spatial-decision support systems*, or SDSS
- Complex decisions are often contentious, with many *stakeholders* interested in the outcome and arguing for one position or another
- SDSS are specially adapted GIS that can be used during the decision-making process, to provide instant feedback on the implications of various proposals and the evaluation of ‘what-if’ scenarios

§ 3 Optimization(cont.)

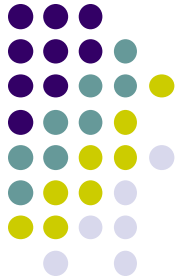
§ 3.1 Point location

- MAT problem, 1-median
- p-median problem: seeks optimum locations for any number *p* of central facilities such that the sum of the distances between each weight and the *nearest* facility is minimized
- *Location-allocation* Problems: where to *locate*, and how to *allocate* demand for service to the central facilities





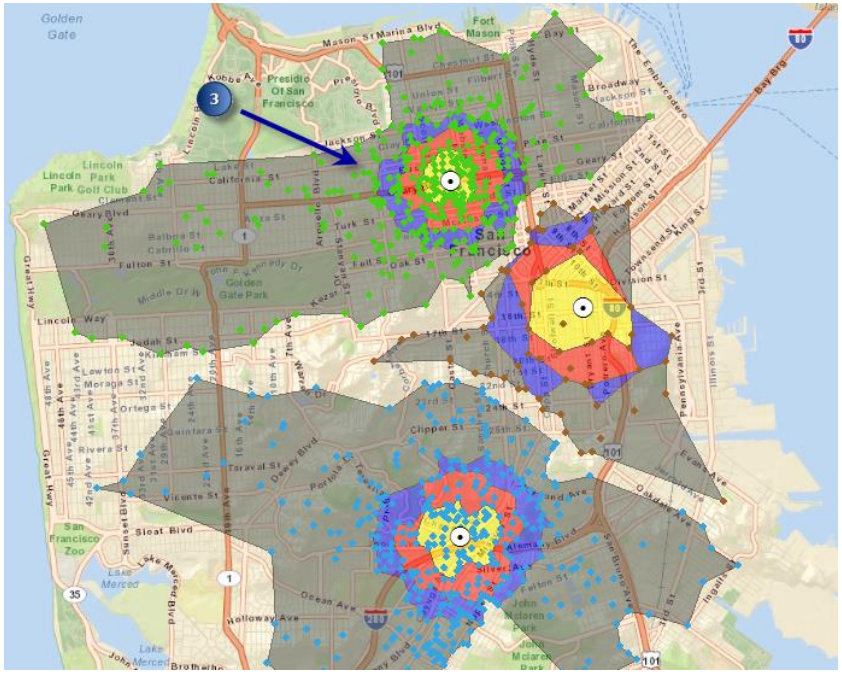
Location-allocation problems

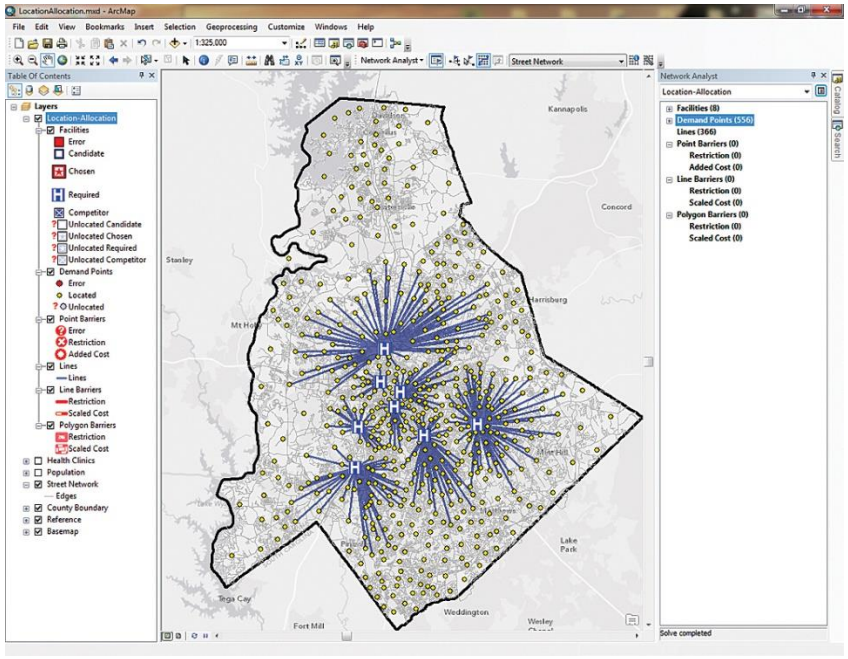


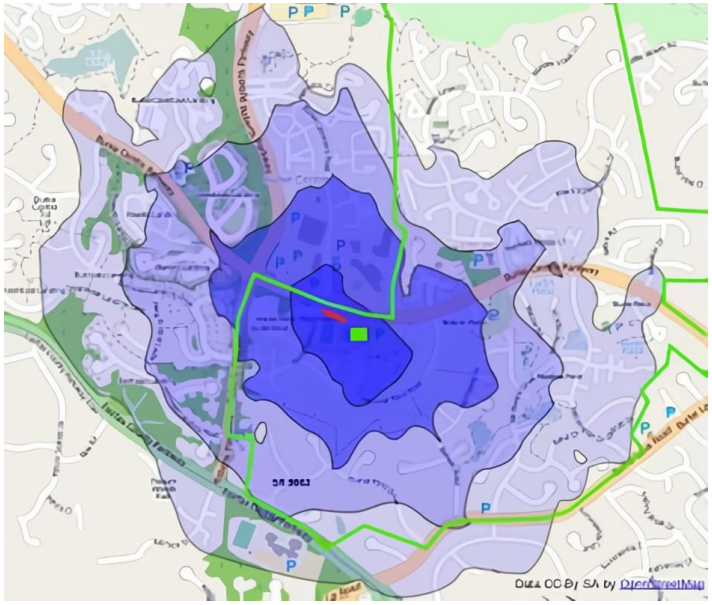
- concern the provision of a service to satisfy a spatially dispersed demand
- demand for the service exists at a large number of widely dispersed sites
  - impossible to provide the service everywhere
    - e.g. every household needs a source of groceries, but impossible to provide a grocery store at each household
- for reasons of cost (economies of scale) service must be provided from a few, centralized locations ("sites")
  - sometimes the number of sites is known in advance, e.g. McDonalds wishes to locate 3 restaurants in city x
  - in other cases the optimum number of sites is one aspect of the solution
- two elements to the problem:
  - 1. Location
    - where to put the central facilities (and possibly how many, how big)
  - 2. Allocation
    - which subsets of the demand should be served from each site ("trade areas", "service areas")

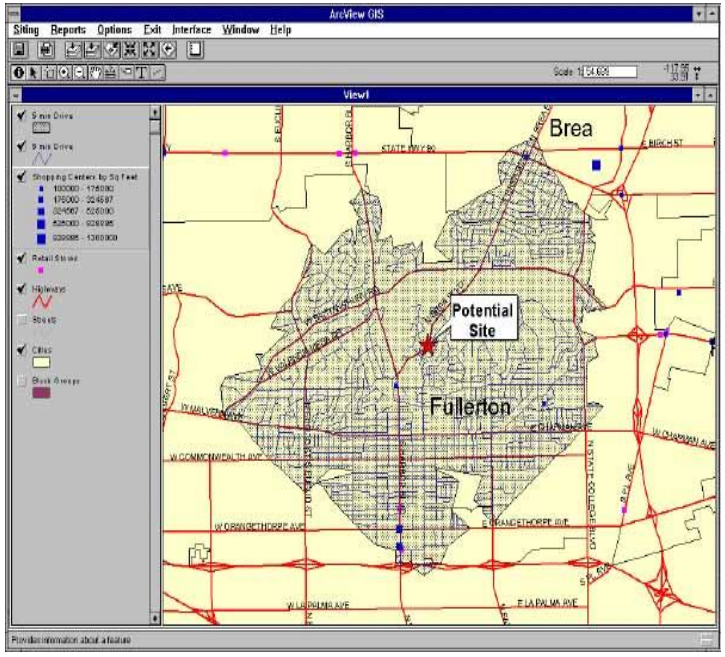
Network analysis in Infrastructure Planning

The GIS expert uses network analysis to select the optimal location for the new facilities





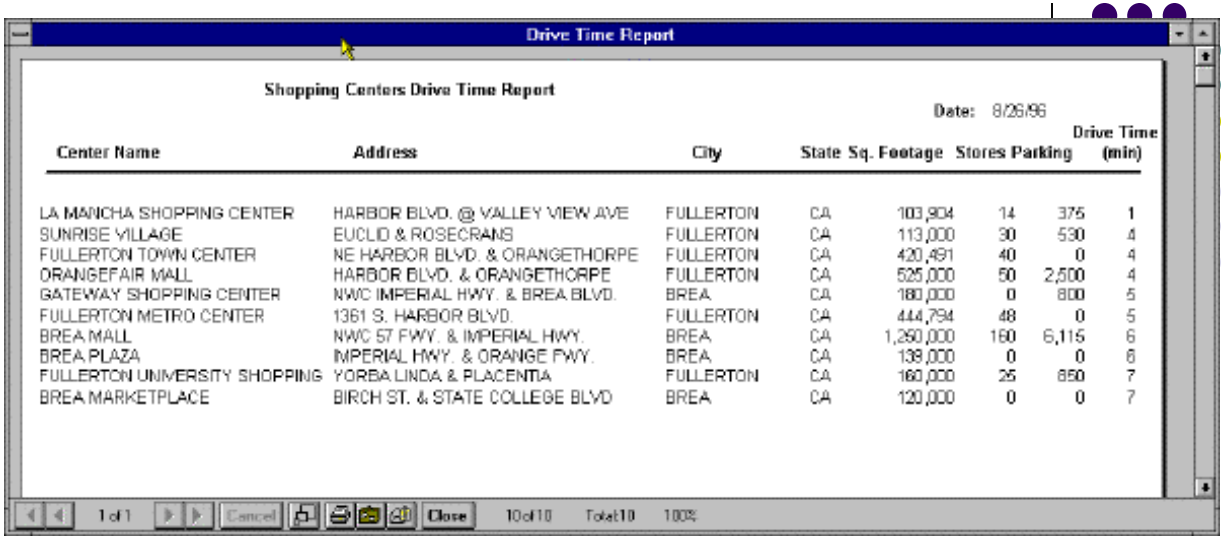


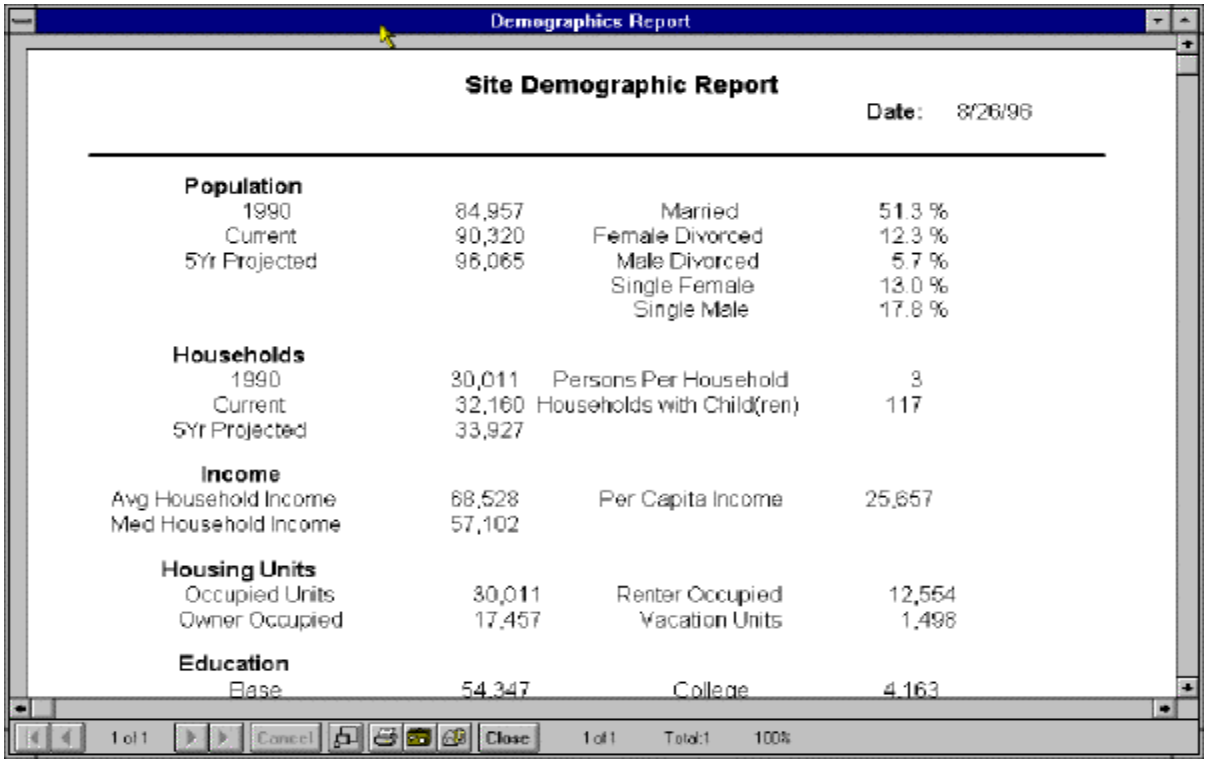


Travel Time

this example shows a much more precise estimation of store visits based on a 5 minute drive time. This estimate is based on travel times along streets, a much more accurate method than simply drawing a circle around a site

In addition, the analysis shown below also provides two printed reports. First is a report showing the 10 closest other stores and their drive times to the proposed facility. The second report is a demographic breakdown of potential customers within the 5 minute drive time.





§ 3 Optimization(cont.)

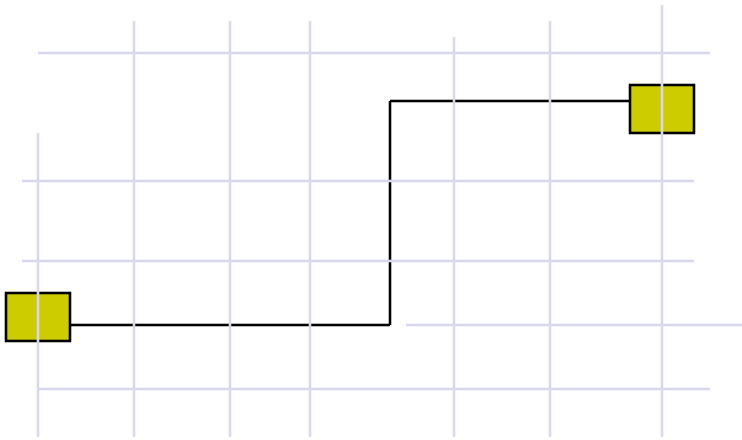
§ 3.2 Routing problems

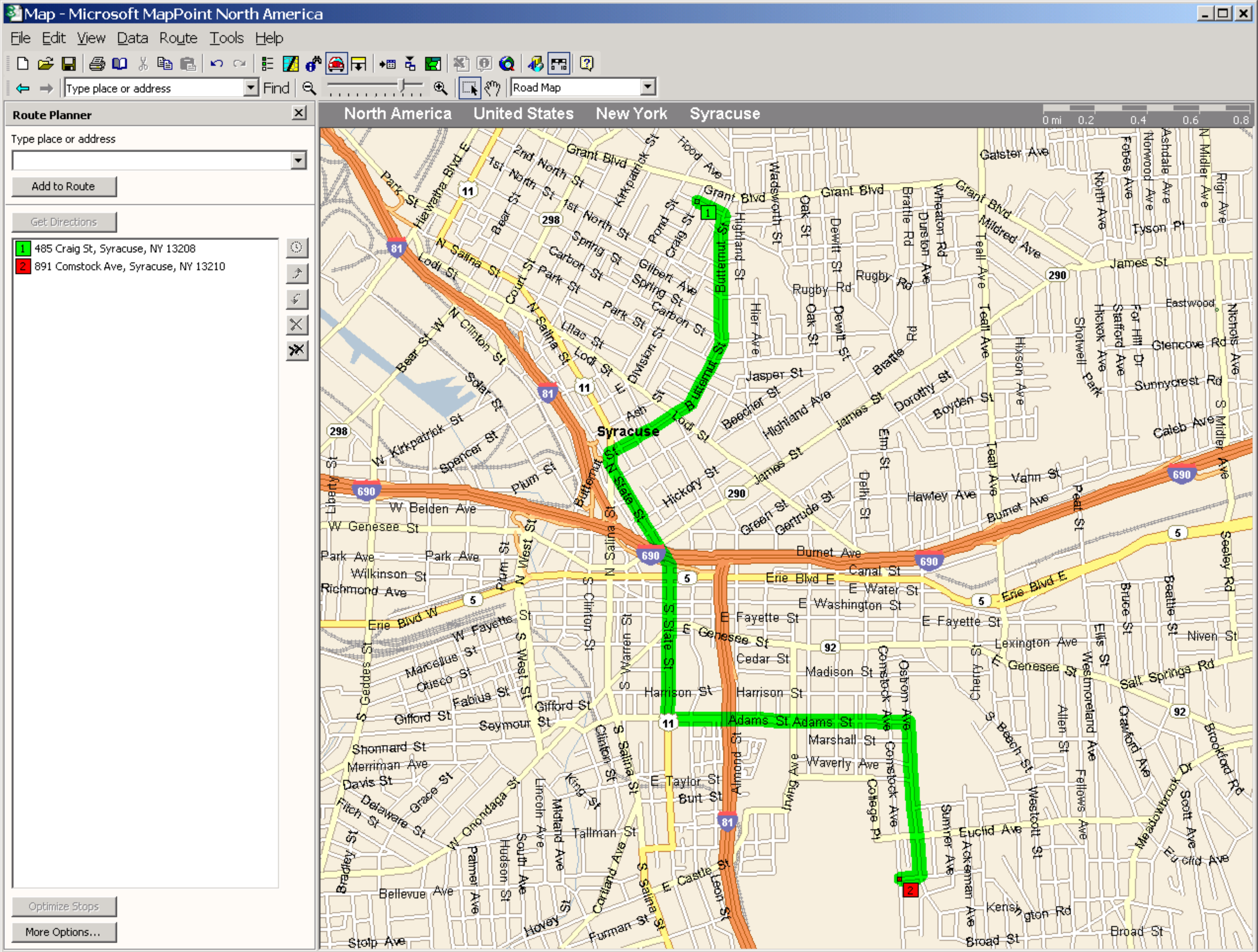
- A GIS can be very effective at solving routing problems because it is able to examine vast numbers of possible solutions quickly
- The simplest routing problem with multiple destinations is the so-called traveling-salesman problem or TSP
- In this problem, there are a number of places that must be visited in a tour from the depot and the distances between pairs of places are known

```
function Dijkstra(Graph, source):

    create vertex set Q

    for each vertex v in Graph:             // Initialization
        dist[v] ← INFINITY                  // Unknown distance from source to v
        prev[v] ← UNDEFINED                 // Previous node in optimal path from source
        add v to Q                          // All nodes initially in Q (unvisited nodes)

    dist[source] ← 0                        // Distance from source to source

    while Q is not empty:
        u ← vertex in Q with min dist[u]    // Source node will be selected first
        remove u from Q

        for each neighbor v of u:           // where v is still in Q.
            alt ← dist[u] + length(u, v)
            if alt < dist[v]:               // A shorter path to v has been found
                dist[v] ← alt
                prev[v] ← u

    return dist[], prev[]
```

Shortest Path

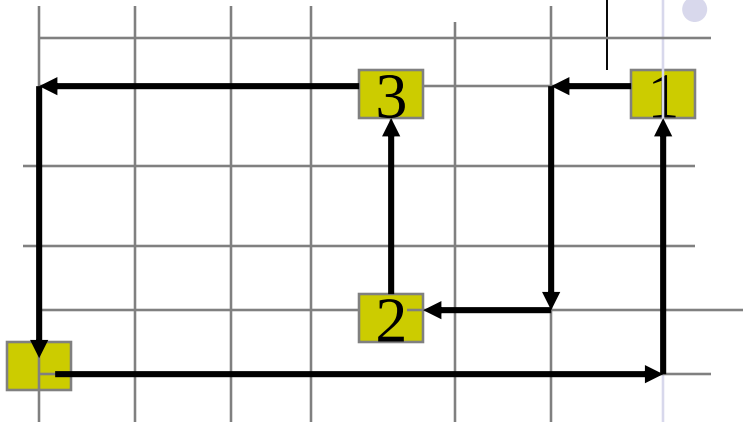
- The shortest path problem is the most basic routing solution. Essentially, we are interested in finding the least cost method to get from point A to point B on a connected graph
- The problem is generally very easy to solve from a mathematical standpoint. In fact, most Sophomore or Junior college students obtaining a degree in operations research are often required write a shortest path algorithm as a class project
- The following examples show a shortest path route in an example graph, and one on a real street network.
- This is an example of a simple path

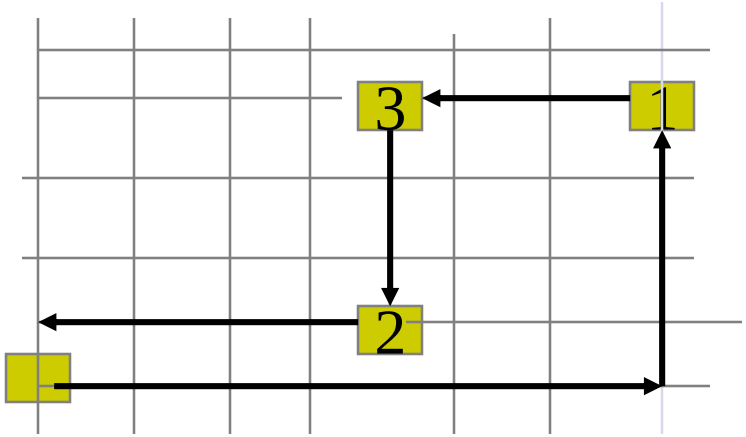


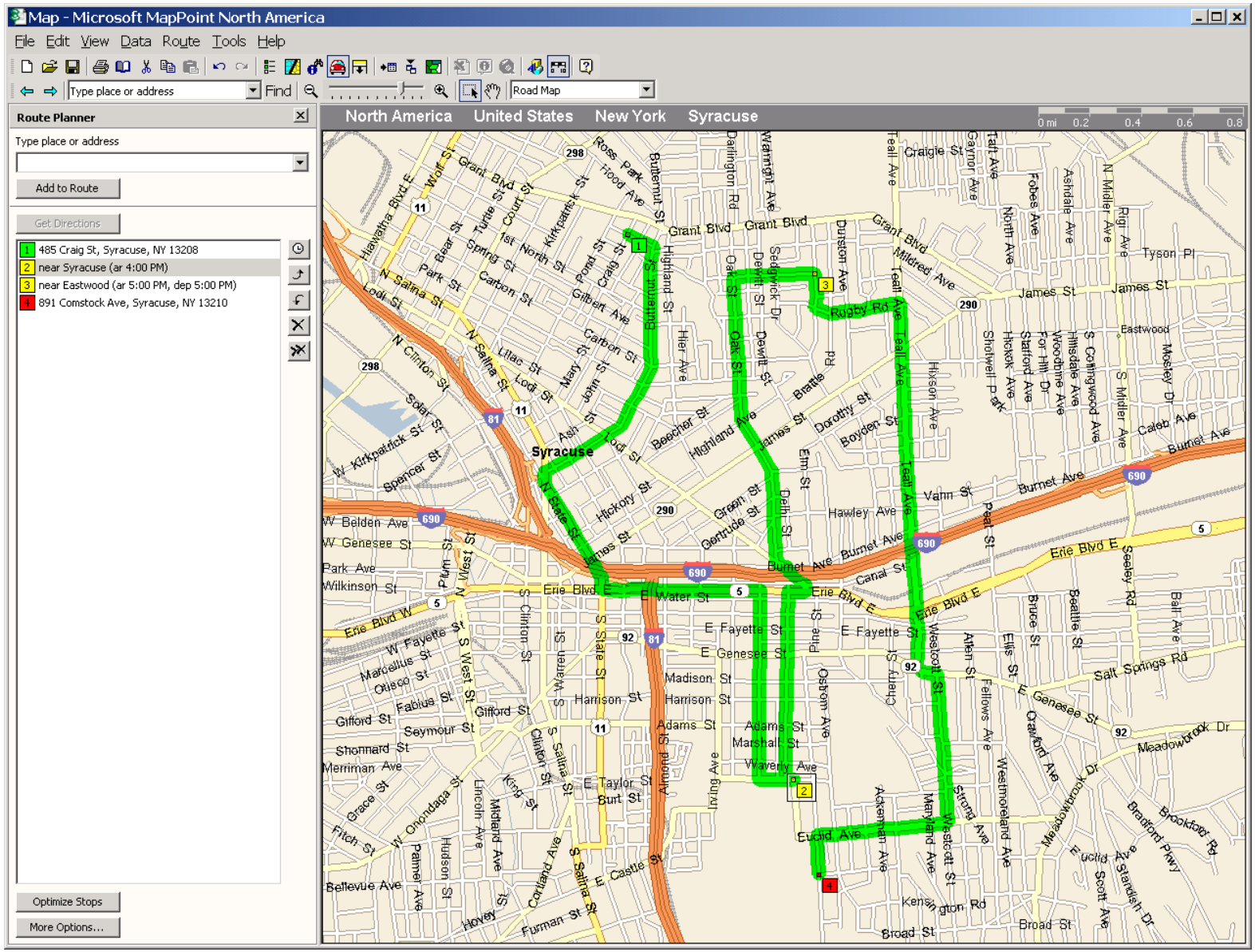

Map - Microsoft MapPoint North America
File Edit View Data Route Tools Help
Type place or address
Find
Road Map
Route Planner
Type place or address
Add to Route
Get Directions
1 485 Craig St, Syracuse, NY 13208
2 891 Comstock Ave, Syracuse, NY 13210
Optimize Stops
More Options...
North America United States New York Syracuse
0 mi 0.2 0.4 0.6 0.8
Syracuse

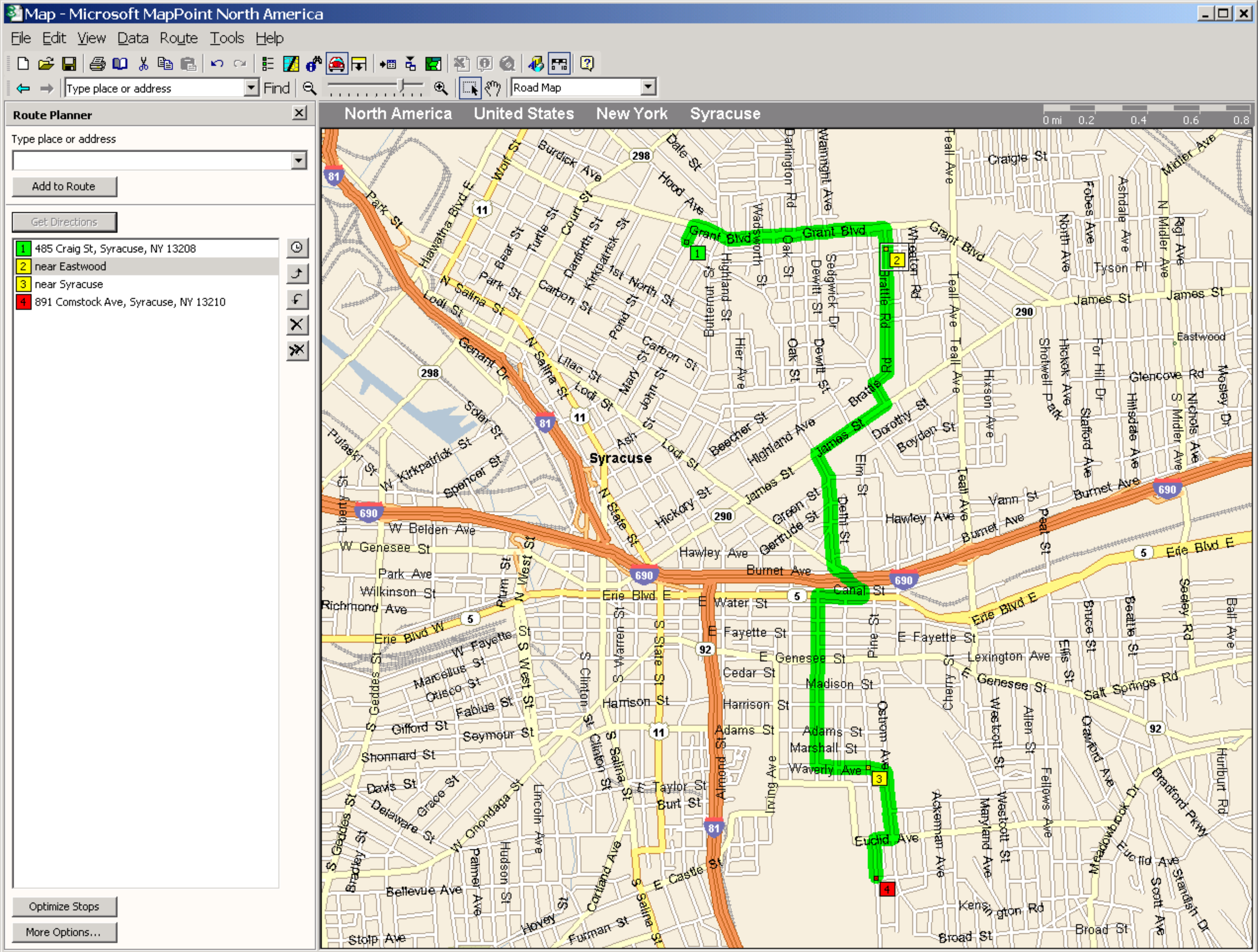
Shortest Path

- Just because you have a GIS that can compute shortest path doesn't mean you are doing routing. Hardly, as you will see, it gets more complex, way more complex
- For example, if we added three more stops (#1, #2, #3) and just ran a shortest path algorithm, are answer would be less than ideal.
- By just coming up with shortest paths from 1 -> 2 -> 3 and then back home, is not the best solution
- What we really want is the best overall tour through a graph



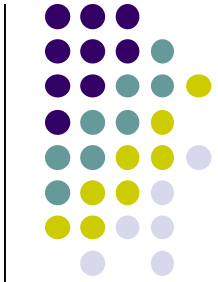



Map - Microsoft MapPoint North America
File Edit View Data Route Tools Help
Type place or address
Find
Road Map
Route Planner
Type place or address
Add to Route
Get Directions
1 485 Craig St, Syracuse, NY 13208
2 near Syracuse (ar 4:00 PM)
3 near Eastwood (ar 5:00 PM, dep 5:00 PM)
4 891 Comstock Ave, Syracuse, NY 13210
Optimize Stops
More Options...
North America
United States
New York
Syracuse
0 mi 0.2 0.4 0.6 0.8
Syracuse
Grant Blvd
Rugby Rd
Teall Ave
James St
Burnet Ave
Erie Blvd E
Erie Blvd W
E Water St
E Fayette St
E Genesee St
Madison St
Harrison St
Adams St
Marshall St
Waverly Ave
Euclid Ave
Westcott St
Salt Springs Rd
Meadowbrook Dr
Broad St
Kensington Rd
Hawley Ave
Vann St
Canal St
Hickory St
Ash St
Lodi St
Spencer St
W Kirkpatrick St
W Belden Ave
W Genesee St
Park Ave
Wilkinson St
W Fayette St
Gifford St
Seymour St
Shonnard St
Merriman Ave
Bellevue Ave
Furman St
Hovey St
E Castle St
E Taylor St
Burt St
Fitch St
W Onondaga St
Grace St
Lincoln Ave
Midland Ave
Hudson St
Palmer Ave
Cortland Ave
King St
Fabius St
Marcellus St
Otisco St
Ostrom Ave
Fellows Ave
Allen St
Ellis St
Bruce St
Beattie St
Bair Ave
Seeley Rd
Crawford Ave
Westmoreland Ave
Maryland Ave
Ackerman Ave
Strong Ave
Brookfoor Rd
Bradford Pkwy
Standish Dr
Scott Ave
Caleb Ave
Mosley Dr
Eastwood
Tyson Pl
Craigie St
Taft Ave
Gaynor Ave
Forbes Ave
Ashdale Ave
N Midler Ave
Rigi Ave
Millard Ave
Durston Ave
Sedgwick Dr
Dewitt St
Oak St
Highland Ave
Hier Ave
Beecher St
Todd St
Carbon St
Gilbert Ave
Pond St
Craig St
Kirkpatrick St
Turtle St
Court St
Bear St
1st North St
Spring St
Park St
N Salina St
N Clinton St
Solar St
Wolf St
Grant Blvd St
Ross Park
Hood Ave
Buttermut St
Darlington Rd
Wainright Ave
Hiawatha Blvd E
Lilac St
Mary St
John St
Ash St
Todd St
Beecher St
Highland Ave
James St
Dorothy St
Boyden St
Hixson Ave
Shotwell Park
Hickok Ave
Stafford Ave
Fox Hill Dr
Woodbine Ave
Hillsdale Ave
S Collingwood Ave
S Midler Ave
Delhi St
Elm St
Green St
Gertrude St
Hickory St
N State St
S State St
S Warren St
S West St
S Clinton St
S Salina St
S Franklin St
Almond St
Ivory Ave
Chery St
Osborn Ave
Peat St
Burnet Ave
Canal St
Erie Blvd E
Erie Blvd W
E Fayette St
E Genesee St
Madison St
Harrison St
Adams St
Marshall St
Waverly Ave
Euclid Ave
Westcott St
Salt Springs Rd
Meadowbrook Dr
Broad St
Kensington Rd
Hawley Ave
Vann St
Canal St
Hickory St
Ash St
Lodi St
Spencer St
W Kirkpatrick St
W Belden Ave
W Genesee St
Park Ave
Wilkinson St
W Fayette St
Gifford St
Seymour St
Shonnard St
Merriman Ave
Bellevue Ave
Furman St
Hovey St
E Castle St
E Taylor St
Burt St
Fitch St
W Onondaga St
Grace St
Lincoln Ave
Midland Ave
Hudson St
Palmer Ave
Cortland Ave
King St
Fabius St
Marcellus St
Otisco St
Ostrom Ave
Fellows Ave
Allen St
Ellis St
Bruce St
Beattie St
Bair Ave
Seeley Rd
Crawford Ave
Westmoreland Ave
Maryland Ave
Ackerman Ave
Strong Ave
Brookfoor Rd
Bradford Pkwy
Standish Dr
Scott Ave
Caleb Ave
Mosley Dr
Eastwood
Tyson Pl
Craigie St
Taft Ave
Gaynor Ave
Forbes Ave
Ashdale Ave
N Midler Ave
Rigi Ave
Millard Ave
Durston Ave
Sedgwick Dr
Dewitt St
Oak St
Highland Ave
Hier Ave
Beecher St
Todd St
Carbon St
Gilbert Ave
Pond St
Craig St
Kirkpatrick St
Turtle St
Court St
Bear St
1st North St
Spring St
Park St
N Salina St
N Clinton St
Solar St
Wolf St
Grant Blvd St
Ross Park
Hood Ave
Buttermut St
Darlington Rd
Wainright Ave
Hiawatha Blvd E
Lilac St
Mary St
John St
Ash St
Todd St
Beecher St
Highland Ave
James St
Dorothy St
Boyden St
Hixson Ave
Shotwell Park
Hickok Ave
Stafford Ave
Fox Hill Dr
Woodbine Ave
Hillsdale Ave
S Collingwood Ave
S Midler Ave
Delhi St
Elm St
Green St
Gertrude St
Hickory St
N State St
S State St
S Warren St
S West St
S Clinton St
S Salina St
S Franklin St
Almond St
Ivory Ave
Chery St
Osborn Ave
Peat St
Burnet Ave
Canal St
Erie Blvd E
Erie Blvd W
E Fayette St
E Genesee St
Madison St
Harrison St
Adams St
Marshall St
Waverly Ave
Euclid Ave
Westcott St
Salt Springs Rd
Meadowbrook Dr
Broad St
Kensington Rd



Map - Microsoft MapPoint North America
File Edit View Data Route Tools Help
Type place or address
Find
Road Map
Route Planner
Type place or address
Add to Route
Get Directions
1 485 Craig St, Syracuse, NY 13208
2 near Eastwood
3 near Syracuse
4 891 Comstock Ave, Syracuse, NY 13210
Optimize Stops
More Options...
North America
United States
New York
Syracuse
0 mi 0.2 0.4 0.6 0.8
Syracuse
Eastwood

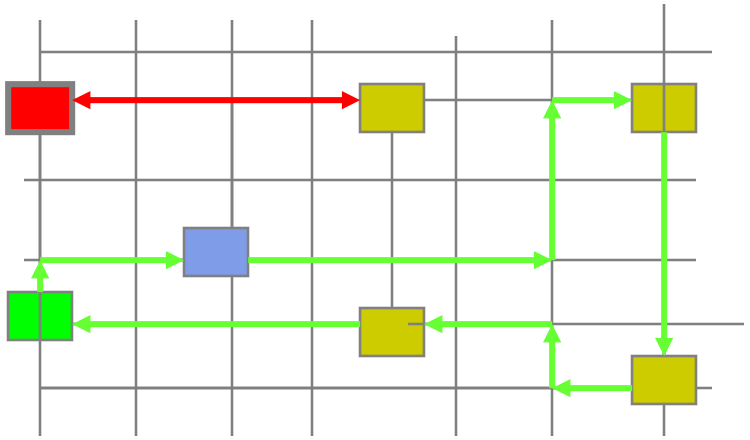
Traveling Salesman Problem



- You can see that our second example actually created a tour, rather than multiple paths. This is often referred to as the traveling salesman problem
- In theory, a salesman must find the most efficient way to visit all of the appointments in his territory. And, he doesn't want to go back to an appointment he already made
- In reality, once you add more than a few links and a few appointments (nodes) the problem is impossible to solve. The true optimal solution requires too many calculations, even with modern computers
- Therefore, a heuristic (a fancy word for an educated guess) finds an approximate best tour, and for the most part is usually very close to the optimal solution
- So, you can see that this is quite more complicated than a shortest path – hang on, we're not out of the woods yet…

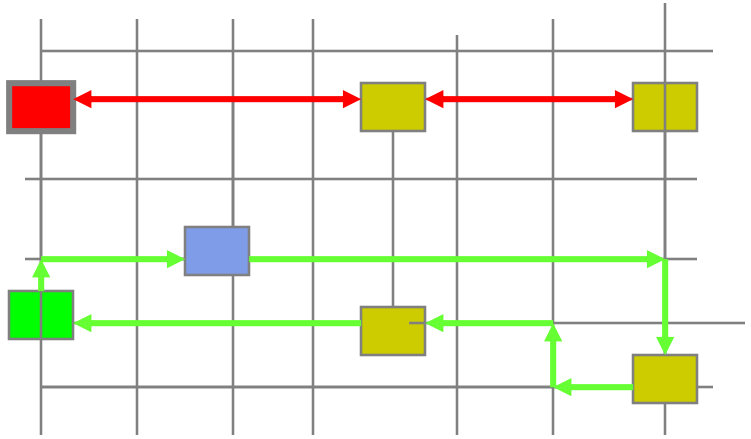
Multiple Traveling Salesman Problem

- We just modified our shortest path problem to create a tour for a single salesman. But, how many companies do you know that have only one salesman?
- The reality is, just because the GIS software you are thinking of buying can solve the traveling salesman problem, doesn't mean its going to meet your need
- That adds to the complexity a little more, doesn't it?
- The example on the right determines the least cost tour for two salesmen and five delivery points



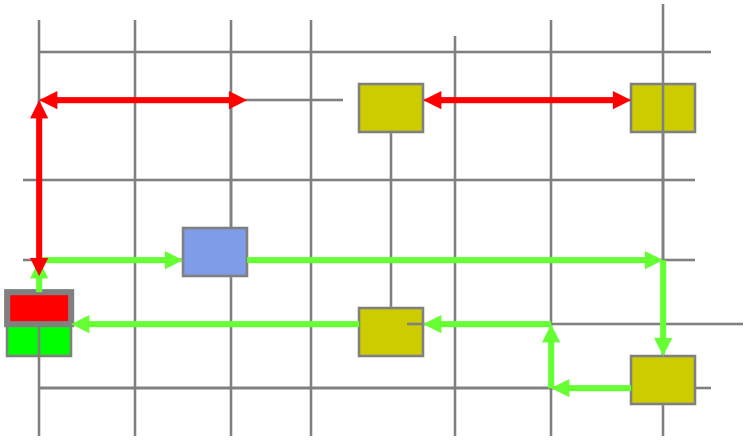
Multiple Traveling Salesman – Balanced workloads

- You might be happy with our solution, IF YOU WERE THE RED SALESMAN!!!
- Obviously just finding the least cost tour is not enough here. One salesman is working four times harder than the other one
- This is a classic example of what happens in the sanitation or meter reading industries. One crew may be able to finish really early and go home (with pay), while the other crew has to work overtime (with time and a half pay). This amounts to a lot of money wasted on inefficiencies
- The important consideration is to not just solve an optimal tour, but also meet a new constraint *Cw* (worktime) . That is, trying to get everyone to nominally work an 8 hour day

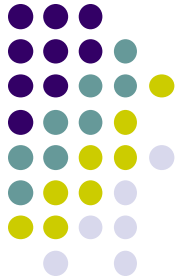


Single Depot, Multiple Vehicle Routing

- This is a takeoff on the last example. But here, there is a single depot that the workers start at, and a single depot where they drop their truck off. This is a fairly typical scenario for a company that has delivery trucks housed at their warehouse
- We up-the-ante on this scenario by adding multiple depots and multiple salesmen
- Each one of these scenarios, though having their basis in the shortest path, shortest tour, actually have some significant additions that you may not find in GIS software



§ 4 Hypothesis testing



- The testing of hypotheses and the drawing of inferences
- Much work in statistics is *inferential(推论统计)* – it uses information obtained from samples to make general conclusions about a larger population, on the assumption that the sample came from that population
- Methods of inference reason from information about a sample to more general information about a larger population

§ 4 Hypothesis testing

- Hypothesis testing is the process of making a choice between two conflicting hypotheses. The null hypothesis, H0, is a statistical proposition stating that there is no significant difference between a hypothesized value of a population parameter and its value estimated from a sample drawn from that population. The alternative hypothesis, H1 or Ha, is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. When the null hypothesis is tested, a decision is either correct or incorrect. An incorrect decision can be made in two ways: We can reject the null hypothesis when it is true (Type I error) or we can fail to reject the null hypothesis when it is false (Type II error). The probability of making Type I and Type II errors is designated by alpha and beta, respectively. The smallest observed significance level for which the null hypothesis would be rejected is referred to as the p-value. The p-value only has meaning as a measure of confidence when the decision is to reject the null hypothesis. It has no meaning when the decision is that the null hypothesis is true

§ 4 Hypothesis testing（cont.）

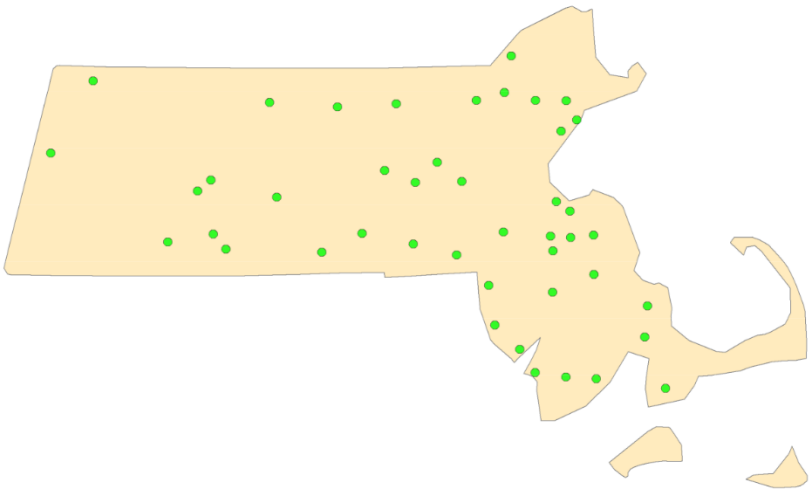
§ 4.1 Hypothesis tests on geographic data

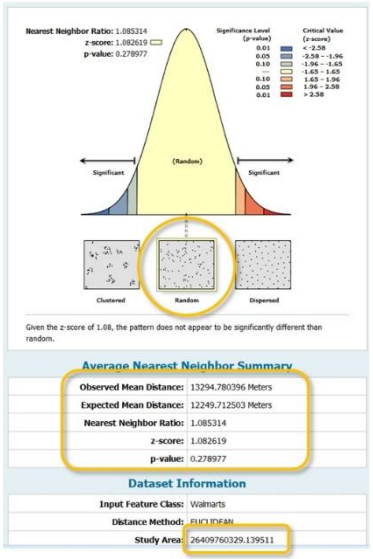
- A GIS project often analyzes all the data in a given area, rather than a sample
  - Data systematically collected
  - Not independent because of first law of geography
- The Earth's surface is very heterogeneous（异质的）, making it difficult to take samples that are truly representative of any large region

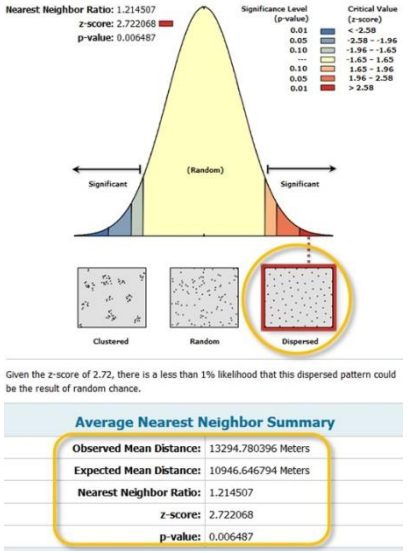
§ 4 Hypothesis testing（cont.）

§ 4.2 Hypothesis tests——an example

- Objective: Test Walmart distribution in Massachusetts
- Testing for complete spatial randomness(CSR) with ANN tool
  - Different Results without or with explicit boundary of the state



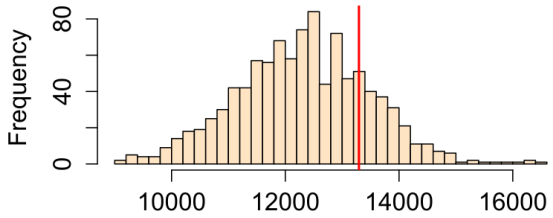




§ 4 Hypothesis testing（cont.）

§ 4.2 Hypothesis tests——an example

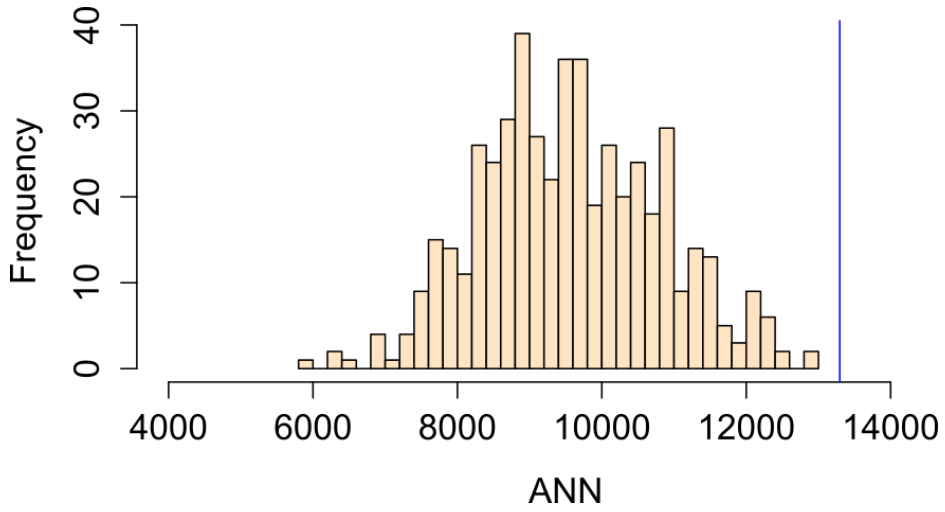
- A better approach: a Monte Carlo test
- First, we postulate a process–our null hypothesis, H0. For example, we hypothesize that the distribution of Walmart stores is consistent with a completely random process (CSR)
- Next, we simulate many realizations of our postulated process and compute a statistic (e.g. ANN) for each realization
- Finally, we compare our observed data to the patterns generated by our simulated processes and assess (via a measure of probability) if our pattern is a likely realization of the hypothesized process
- plot all $ANN_{expected}$ values using a histogram, then compare our observed ANN value of 13,294 m to this distribution

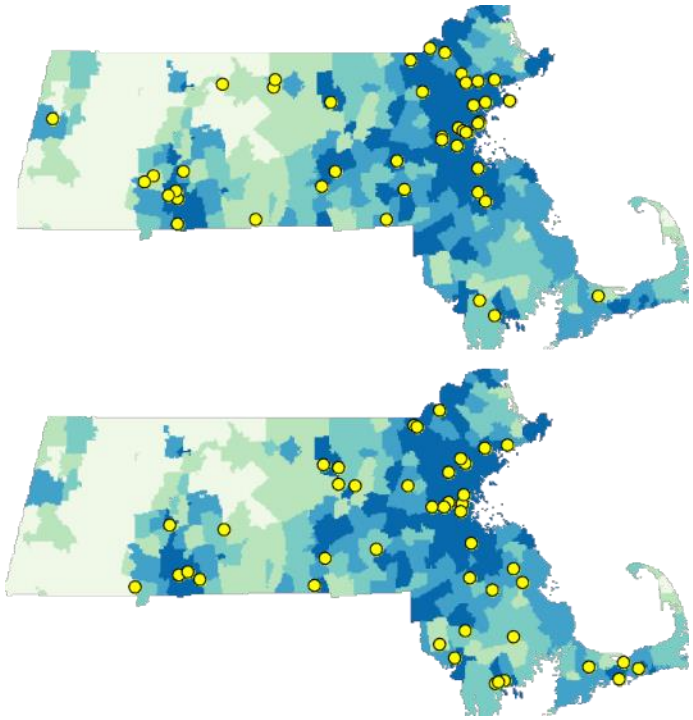


§ 4 Hypothesis testing（cont.）

§ 4.2 Hypothesis tests——an example

- Testing by simulating the placement of Walmart stores using the population density layer

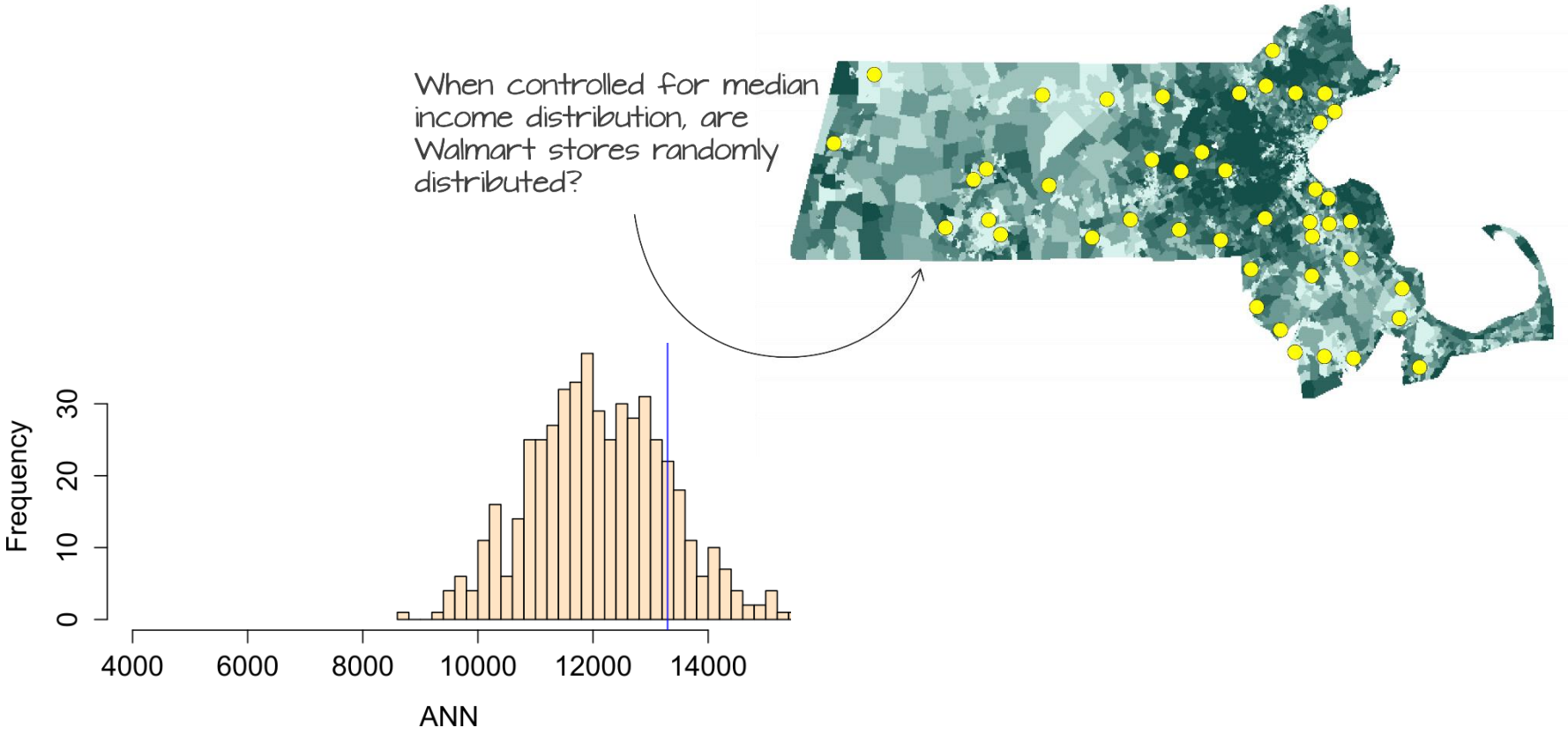




§ 4 Hypothesis testing（cont.）

§ 4.2 Hypothesis tests——an example

- Testing by simulating the placement of Walmart stores using the median household income



§ 4 Hypothesis testing（cont.）

§ 4.2 Hypothesis tests——an example

- Testing for a covariate effect: on a density based approach to point pattern analysis: The Poisson point process model
- we may want to assess if the Poisson point process model that pits the placement of Walmarts as a function of population distribution (the alternate hypothesis) does a better job than the null model that assumes homogeneous intensity
- p-value which gives us the probability we would be wrong in rejecting the null. Here p=0.039 suggests that there is an 3.9% chance that we would be remiss to reject the base model in favor of the alternate model
- put another way, the alternate model may be an improvement over the null model

§ 5 Conclusion

- has covered the conceptual basis of many of the more sophisticated techniques of spatial analysis
- in particular raised some fundamental issues associated with applying methods and theories that were developed for non-spatial data to the spatial case
- Spatial analysis is clearly not a simple and straightforward extension of non-spatial analysis, but instead raises many distinct problems, as well as some exciting opportunities
- The two chapters on spatial analysis have only scratched the surface of this large and rapidly expanding field