UNT's course evaluation system (SPOT - Student Perceptions of Teaching) opens on **Monday, April 16** and runs through **Thursday, May 3**.

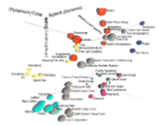
You should receive an email on April 16 providing guidance on how to respond.

Please do respond: I need and value your feedback on what worked well this semester, and what can be improved.

Thanks in advance for your helpful comments

Week 11

Cluster Analysis



Cluster Analysis

- This is another tool we can use to simplify a very complex, multivariate database
- However, unlike factor analysis, this method operates specifically in spatial terms: group data (observations) in "space"
  - As we discussed with MDS, "space" can be geographical, or perhaps another kind of space we conceptualize (such as a "perception space" or a "similarity space")

Cluster Analysis

- Steps in cluster analysis ("CA")
  - 1. take points, areas, or objects ("observations") and measure the "distance" between each pair
  - 2. analyze this distance data to uncover the latent grouping structure embodied in the dataset
- Some kind of measure of similarity/ dissimilarity is needed to do this analysis

---

Cluster Analysis

- Purpose of CA: see trends, generate hypotheses (highly exploratory)
  - Important advantage: CA does not need normality or linearity (non-parametric), so cluster analysis can be widely used
  - There are more opportunities for use of CA than actual implementations: CA is a method to be aware of for its potential for innovation

---

Cluster Analysis

- Key Idea: cluster analysis usually does not focus on geographic space
  - Clusters are often defined in non-geographic terms: "space" in some other sense
  - Focus is the creation of a classification system: clusters (in this context) = groupings
    - Groupings of people: based on health and lifestyle factors
    - Groupings of forests: based on vegetation types and climate characteristics
    - Groupings of cities: based on major industries or other socioeconomic characteristics

Cluster Analysis

- Emphasis in this class: one specific approach to clustering called *hierarchical clustering*
  - Hierarchical clustering: provides information on clustering at multiple levels of complexity
    - With hierarchical clustering, one analysis gives you information to cluster a database into 2 groupings, 3 groupings, 4 groupings, etc. (max. groupings = # of records in dataset)
    - You don't need to know in advance how many groupings (clusters) you want to produce
    - Hierarchical clustering gives you insight to help you select how many clusters you wish to identify

Cluster Analysis

- Alternative path to a solution: another approach called *k-means clustering*
  - K-means clustering: efficient method for producing a *specified number* of clusters
    - "Efficient" in terms of computer run-time
    - However, with k-means clustering you need to know how many clusters are appropriate for your dataset (or at least, how many you want to see)
    - You could do k-means clustering multiple times to compare different levels of cluster systems, but that negates its time efficiency

Cluster Analysis

- Let's look at a dataset of nine values to see the basic cluster idea (hierarchical)

Cell identifier

Actual data value

A 50	B 20	C 18
D 7	E 3	F 34
G 71	H 80	I 86

Imagine each cell value as a data observation for a given geographic area (9 observations for 9 areas in a 3x3 grid)

Cluster Analysis

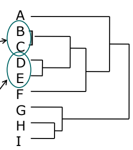
- Let's look at a dataset of nine values to see the basic cluster idea (hierarchical)
  - CA generates a dendrogram chart to show the hierarchical structure in this table

A 50	B 20	C 18
D 7	E 3	F 34
G 71	H 80	I 86

These cells link first because of smallest distance (20-18=2)

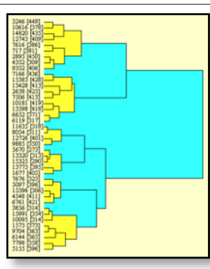
These cells link next (7-3=4)

And so on …

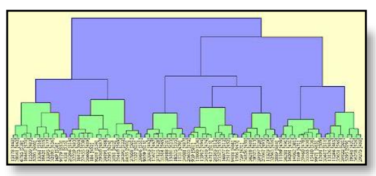


Cluster Analysis

A slightly more complex dendrogram example



Cluster Analysis

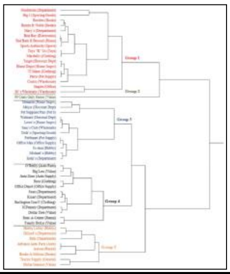


And another…
(they can be much bigger yet)

Cluster Analysis

One real-world dendrogram example from research in business geography

Joseph, Lawrence (2016) The Geographic Exposure to Lifestyles by U.S. Retail Chains, *The Professional Geographer*, DOI: 10.1080/00330124.2016.1140497



Cluster Analysis

- Big idea of the dendrogram: shows the order/structure of the joins (hierarchy)
  - Most similar cells join first, less similar cells join later, until all cells are joined

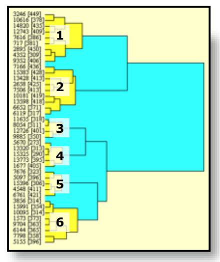
Cluster Analysis

- **Key issues: how do clusters emerge from a dendrogram, and many clusters are appropriate?**
  - Let's look at some simple examples to gain insight into these basic questions and how we can address them

Cluster Analysis

Here the analyst has identified six clusters (the six yellow areas)

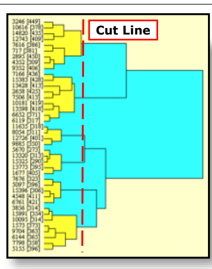
How exactly are these six clusters defined?



Cluster Analysis

The concept of a "cut line" helps here

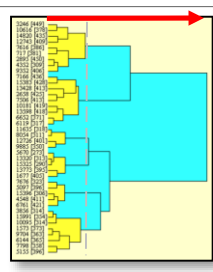
A cut line defines the distance at which the user chooses to create clusters



Cluster Analysis

The concept of a "cut line" helps here

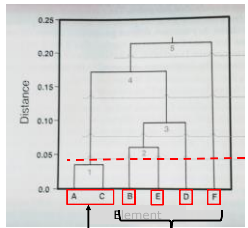
The cut line is the distance you select as the maximum you wish to consider for cluster creation



Distance Axis: the distance between elements (each cluster has an associated distance)

Here's another simple example that develops the use of a cut line

Six Elements: A, B, C, D, E, and F

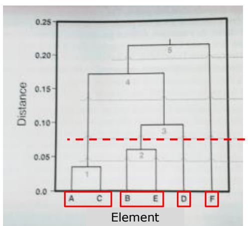


5-cluster cut line

The first actual cluster to form

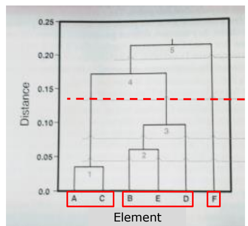
Four individual elements

Six Elements: A, B, C, D, E, and F



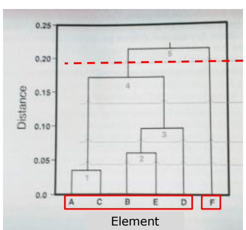
4-cluster cut line

Six Elements: A, B, C, D, E, and F



3-cluster cut line

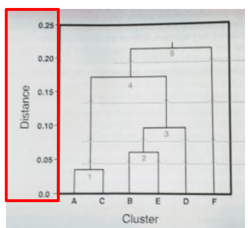
Six Elements: A, B, C, D, E, and F



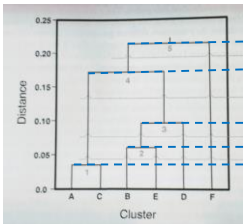
2-cluster cut line

Q: Does this dendrogram give any clues as to what might be the most appropriate number of clusters?

Check out the <u>distance axis</u>: how far each element is apart (and at what distance does each cluster form)



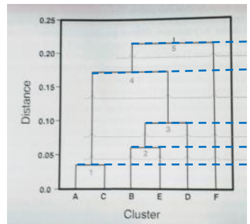
Q: What does the distance information here tell us?



0.22
0.17
0.10
0.06
0.03

Cluster formation distances

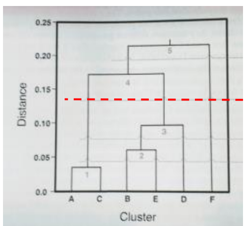
Q: What does the distance information here tell us?



0.22
0.17
0.10
0.06
0.03

Largest jump

So it looks like a 3-cluster solution is best for this situation



Clusters that are much more weak (4, 5)

3-cluster cut line

Strongest clusters (1-3)

Spatial Cluster Analysis Examples

- We will explore more complex situations than this in a few minutes
  - However, right now let's try to get a better grasp of the power of cluster analysis from a <u>spatial perspective</u>: what can you do?
  - The following is one of the most widely-used applications of CA today

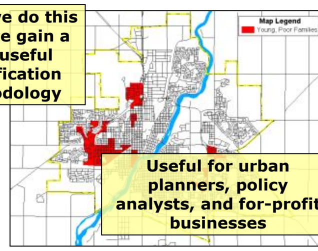
---

Spatial Cluster Analysis Examples

- **Application: CA for geodemographic analysis**
  - Goal: identify <u>uniform subareas</u> within cities
  - Identify the <u>number</u> and <u>kinds</u> of neighborhoods that exist across a city, state, or country
  - See where each of these kinds of neighborhoods can be found across the city
  - Many practical applications

Spatial Cluster Analysis Examples

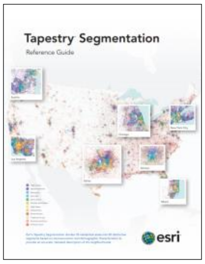
When we do this well, we gain a very useful classification methodology



Useful for urban planners, policy analysts, and for-profit businesses

Spatial Cluster Analysis Examples

See the Esri "Tapestry Segmentation Reference Guide" that I placed on our course website for a full example of this neighborhood-level application



Spatial Cluster Analysis Examples

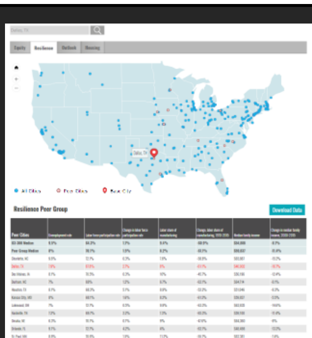
Another great example: Use cluster analysis to identify <u>groups of similar cities</u> across the country

Peer cities are cities that are experiencing <u>similar trends</u> or challenges. To do this well, we need to account for trends in <u>multiple areas</u> of urban life.

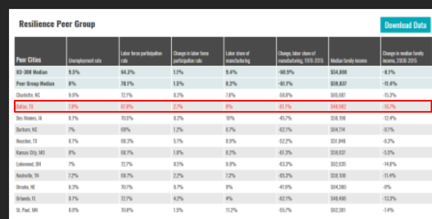
https://www.chicagofed.org/region/community-development/data/pcit

City Example: Search for peer cities for **Dallas**

This map shows Dallas' peer cities across the country, while the table shows the data used to create Dallas' peer grouping



The data table shows the list of variables used to create Dallas' grouping, and how closely Dallas compares to the other cities in its particular peer group (cluster)



Cluster Analysis Extensions

- It is of course possible to do this kind of grouping with a <u>single grouping mechanism</u>
  - For example, clusters based on *income only*
- It is easy to form groupings in that simple case: just identify <u>all the high income cities</u>
- However, real world problems are seldom so simple
  - The <u>complex peer groupings</u> we see in this city example (and in the Esri Tapestry example) provide a much better comparison cluster than one based on a single indicator

Cluster Analysis Extensions

- **To analyze such situations properly, we need to give thought to our data**
  - In a real-world geodemographic analysis application, CA creates city groupings based on <u>dozens of variables</u>
    - Age groups, occupations, education levels, income levels, mobility levels, ethnic backgrounds, housing characteristics, …
  - A truly <u>reliable</u> way of gaining <u>deep insight</u> into local neighborhoods across the city
  - Insights you simply could not obtain by "looking at the data"

Cluster Analysis Extensions

- **Aside: when we use many variables, one key issue to consider is variable standardization**
  - Standardization puts all variables on the same scale (similar to beta values in regression)
  - Standardization is necessary when we deal with variables with different value scales
    - Age groups (0-120 years),
    - Income levels ($0-$20 million)
    - Dwelling size (100-40,000 square feet)
  - **Q:** what issue could arise if we simply put all of these variables into a single analysis?

Cluster Analysis Extensions

- Urban analysis is just one example of the use of CA in a multivariate setting
- Any time you have a complex dataset of many observations involving multiple variables, CA can help you understand what's going on
  - Archaeological sites, hurricane deposition zones, soil samples, air samples, survey results, …

Cluster Analysis: The Details

- Given the power of such a multivariate CA application, how do we actually do this stuff?
  - **First question:** how do you measure "distances" between observations when each observation includes multiple variables?

Cluster Analysis: The Details

- **Multivariate CA Example: An Agricultural Census**

Group the Counties by Similarity of Farm Outputs

Farm Product	County W	County X	County Y	County Z
Wheat	6	5	10	8
Hay	1	2	3	4
Oats	5	5	1	2

Can't just do a simple subtraction: W-X, or X-Y

Cluster Analysis: The Details

- **Distance Metrics**
  - Euclid $d_{xy} = \sqrt{\sum_i (x_i - y_i)^2}$
  - So, for example, between counties X and Y in the previous table

Farm Product	County X	County Y	X-Y	(X-Y)²
Wheat	5	10	-5	25
Hay	2	3	-1	1
Oats	5	1	4	16
Total				**42**

$d_{xy} = \sqrt{42} = 6.5$ So, county X is 6.5 units from county Y

Cluster Analysis: The Details

- **Distance Metrics**
  - Squared Euclid
    - Same as Euclid, except squared (duh!)
    - So, in the previous example, county X is 42 units from county Y (not 6.5)
    - Idea: puts a much larger penalty on large distances, the groups it identifies tend to be very similar

Cluster Analysis: The Details

- **Distance Metrics**
  - Manhattan (or City Block)

$$d_{xy} = \sum_i |(x_i - y_i)|$$

  - In the previous county X/Y example, $d_{xy}=10$

Farm Product	County X	County Y	X-Y	|X-Y|
Wheat	5	10	-5	5
Hay	2	3	-1	1
Oats	5	1	4	4
Total				10

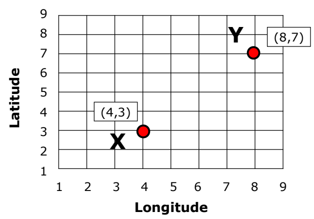
Cluster Analysis: The Details

- **Distance Metrics**
  - Manhattan (or City Block)

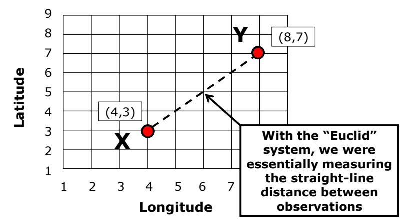
$$d_{xy} = \sum_i |(x_i - y_i)|$$

  - Why "Manhattan"?
  - Think of a spatial example of clustering: cluster observations based on their spatial coordinates
    - Two variables for each observation: longitude and latitude

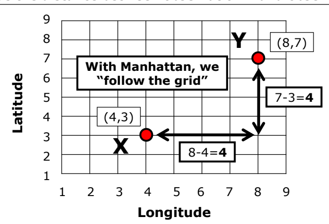
What's the distance between observation X and observation Y?



What's the distance between observation X and observation Y?



What's the distance between observation X and observation Y?



$$d_{xy} = \sum_i |(x_i - y_i)| = (8-4) + (7-3) = 4 + 4 = 8$$

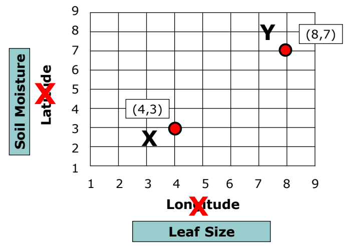
$$\boxed{d_{xy} = 8}$$

What's the distance between observation X and observation Y?

Now that we've defined this distance measurement system, we could cluster <u>any number of points</u> (observations) using their "Manhattan" differences as the starting point

Also note that we can use <u>any number and type of variables</u> for our distance calculations, not just the <u>two</u> longitude and latitude variables used here

What's the distance between observation X and observation Y?

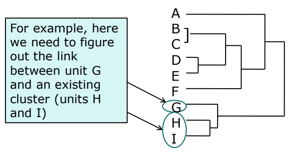


Cluster Analysis: The Details

- **Another question:** following from what we just did
  - We now know how to calculate distances between <u>two different, individual units</u> (like counties or census tracts)
  - How do we calculate distances between an <u>individual unit</u> and a <u>cluster that's already been created</u>?
  - And, how do we calculate distance between <u>two existing clusters</u>?

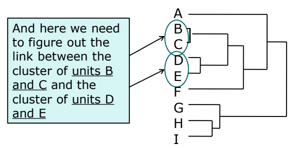
Cluster Analysis: The Details

- This question will come up with <u>every cluster system</u> we try to set up



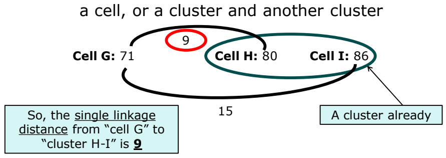
Cluster Analysis: The Details

- This question will come up with <u>every cluster system</u> we try to set up



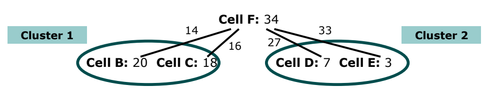
Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 1. Single Linkage (Nearest Neighbors)
  - The <u>smallest distance</u> between a cluster and a cell, or a cluster and another cluster



Cluster Analysis: The Details

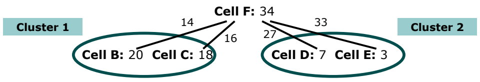
- **Methods of Defining Distances When Clusters are Involved**
- 1. Single Linkage (Nearest Neighbors)
  - The <u>smallest distance</u> between a cluster and a cell, or a cluster and another cluster



Q: What are the single linkage distances here (F to Cluster 1, and F to Cluster 2), and which cluster does F link to?

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 1. Single Linkage (Nearest Neighbors)
  - The smallest distance between a cluster and a cell, or a cluster and another cluster



We can use this methodology to calculate and then compare distances for multiple clusters

Cluster Analysis: The Details

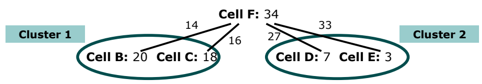
- **Methods of Defining Distances When Clusters are Involved**
- 2. Complete Linkage (Furthest Neighbors)
  - The largest distance between a cluster and a cell, or a cluster and another cluster
  - Pretty much the same idea as what we just went through, except substitute "largest" for "smallest" in your distance calculations

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 2. Complete Linkage (Furthest Neighbors)
  - However: you still cluster based on smallest distances
  - "Largest" only applies to the distance calculation, not the clustering

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 2. Complete Linkage (Furthest Neighbors)
  - Below, what are the complete linkage distances between cell F and the two clusters?

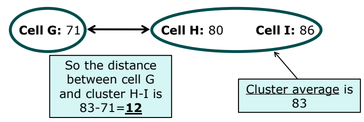


Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 3. Centroid Linkage
  - The distance between a cell and a cluster is the difference between the cell value and the average of the cluster
  - The distance between two cells is the difference between their averages

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 3. Centroid Linkage
  - So, for example



Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Between Groups)
  - Examine <u>all pairs of points</u> (cell-cluster, cluster-cluster) in the distance calculation
  - Reduces to the same as centroid linkage when just dealing with a cell-cluster distance calculation

Cluster Analysis: The Details

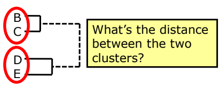
- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Between Groups)

$$d = \frac{\sum diff}{n}$$

  - *d* = the total calculated inter-cluster distance (the actual average linkage)
  - *diff* = individual cell pair differences
  - *n* = number of cell pairs

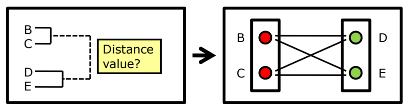
Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Between Groups)
  - Example



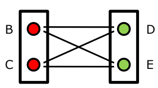
Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Between Groups)
  - Example



Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Between Groups)
  - Example



$d_{BD} = 13$
$d_{BE} = 17$

$d_{CD} = 11$
$d_{CE} = 15$

d = Σ diff / n
= 56/4
= 14

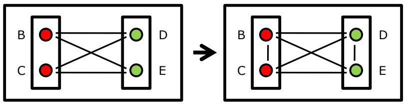
So, the average linkage distance between B-C and D-E is 14.

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (<u>Within</u> Groups)
  - Our last method for discussion is the same as what we just discussed, except that the <u>within group option</u> also takes into account <u>intra-group distances</u> in the calculation

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- 4. Average Linkage (Within Groups)
  - Symbolically, this means



Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- In general, average linkage methods are best because they use all the data points
  - However, there is a place for all methods in the appropriate circumstance

Cluster Analysis: The Details

- **Methods of Defining Distances When Clusters are Involved**
- Other methods you may run across include Ward's, median
  - Some restrictions exist on the use of certain distance methods, depending on the nature of your database
  - But for the ones you've just seen here, no big problems