Clusters

Unsupervised Learning

Luc Anselin

http://spatial.uchicago.edu

curse of dimensionality

principal components

multidimensional scaling

classical clustering methods

Curse of Dimensionality

- Curse of dimensionality

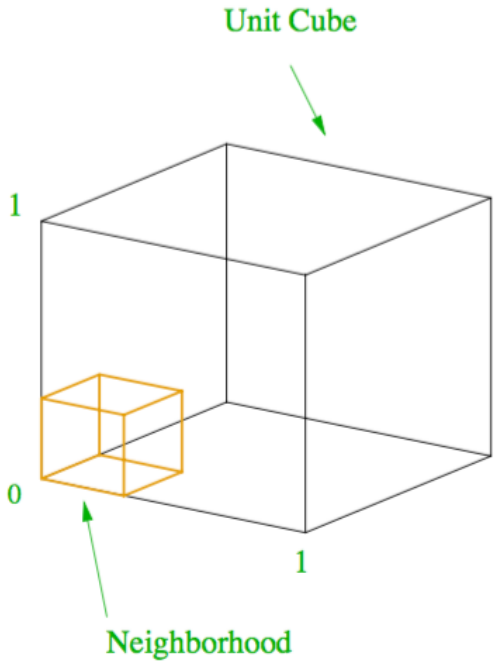
  in a nutshell

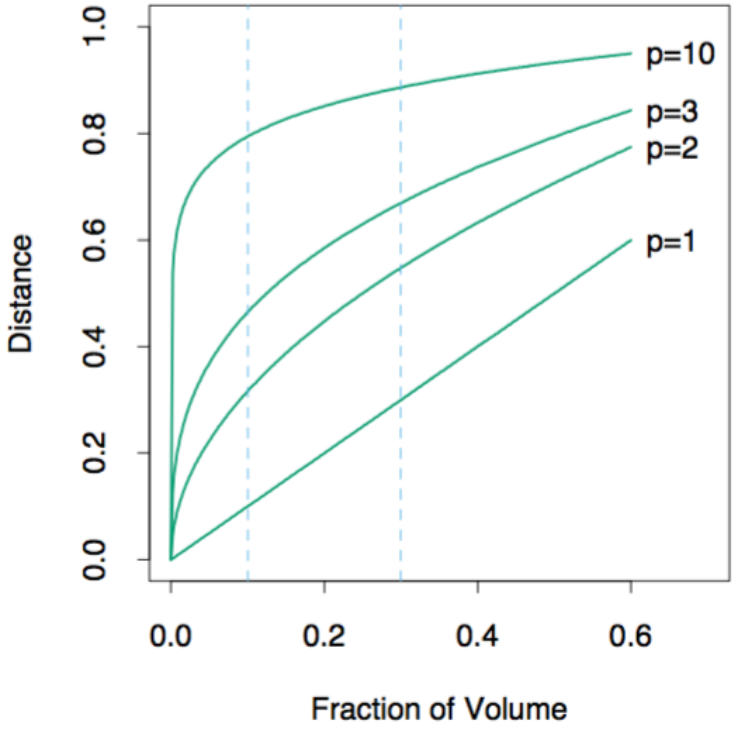
  low dimensional techniques break down in high dimensions

  complexity of some functions increases exponentially with variable dimension

Example 1

change with p (variable dimension) of distance in unit cube required to reach fraction of total data volume

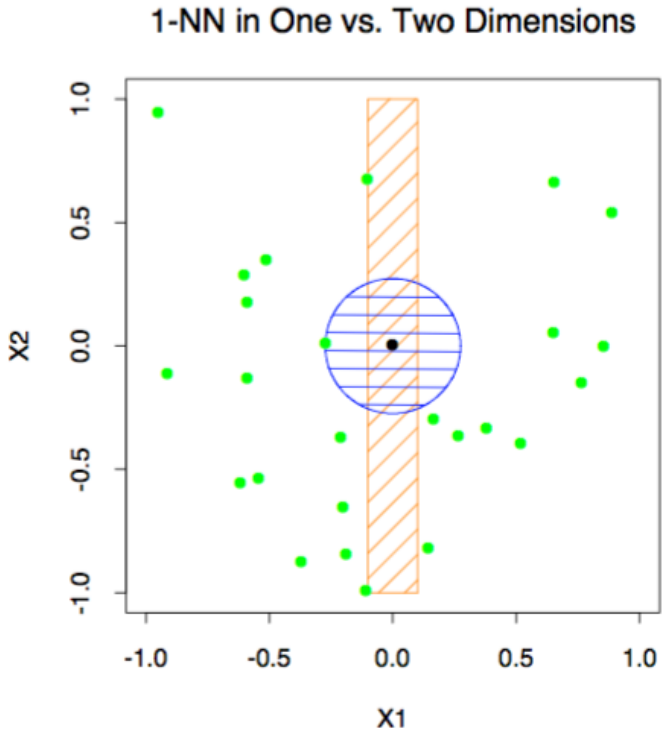




Source: Hastie, Tibshirani, Friedman (2009)

Example 2

nearest neighbor distance in one vs 2 dimensions



Source: Hastie, Tibshirani, Friedman (2009)

- Dimension reduction

  reduce multiple variables into a smaller number of functions of the original variables

    principal component analysis (PCA)

  visualize the multivariate similarity (distance) between observations in a lower dimension

    multidimensional scaling (MDS)

    projection pursuit

Principal Components

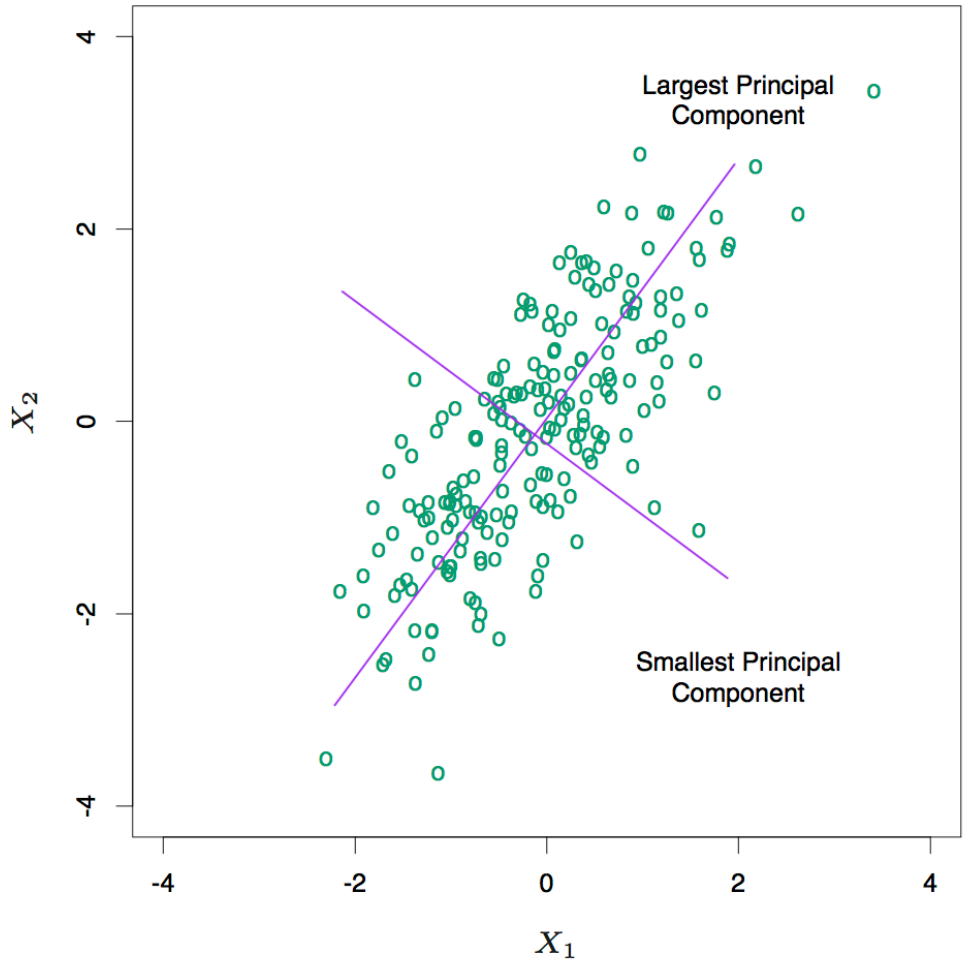
- Principle

  capture the variance in the p by p covariance matrix X'X through a set of k principal components with k << p

  principal components capture most of the variance

  principal components are orthogonal

  principal component coefficients (loadings) are scaled



principal components variance decomposition

Source: Hastie, Tibshirani, Friedman (2009)

- More formally

$c_i = a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{ip} x_p$

each principal component is a weighted sum of the original variables

$c_i' c_j = 0$

components are orthogonal to each other

$\sum_k a_{ik}^2 = 1$

the sum of the squared loadings equals one

components in direction of greatest variation

closest linear fit to the data points

- Eigenvalue Decomposition

  eigenvalue and eigenvectors of square matrix A

  $Av = \gamma v$ with γ as eigenvalue and

  v as eigenvector

  v is a special vector such that translation by A (both a rotation and a shift) remains on v

  eigen decomposition of covariance matrix X'X (with X n by p, X'X is p by p)

  X'X = VGV' a p by p matrix

  V is p by p matrix of eigenvectors

  G is diagonal matrix of eigenvalues

- Singular Value Decomposition (SVD)

  applied to a n by p matrix X (not square X'X)

  $X = UDV'$

  U is orthogonal n by p, $U'U = I$

  V is orthogonal p by p, $V'V = I$

  D is diagonal p by p

  SVD applied to X'X (covariance matrix)

  $X'X = VDU'UDV' = VD^2V'$

  since $U'U = I$

  D^2 is a diagonal matrix with the eigenvalues

- Principal Components and Eigenvectors

  principal components are Xv

  $c_{im} = \Sigma_k x_{ik} v_{km}$ for each eigenvector m

  each variable multiplied by its loading in the eigenvector

  signs are arbitrary (v and -v are equivalent)

  $\Sigma_k v_{km}^2 = 1$

- Typical results of interest

  loadings for each principal component

    the contribution of each of the original variables to that component

  principal component score

    the value of the principal component for each observation

  variance proportion explained

    the proportion of the overall variance each principal component explains

Illustration

- Andre-Michel Guerry -
  Moral Statistics of France (1833)

  social indicators for 86 French departements

  (one of the) first multivariate statistical analyses with geographic visualization

  available as R package Guerry by Friendly and Dray (see also Friendly 2007 for an illustration)

    converted to a GeoDa sample data set

  re-analyzed by Dray and Jombart (2011)

- Moral Statistics

  crime against persons (pop/crime) - Crm_prs

  crime against property (pop/crime) - Crm_prp

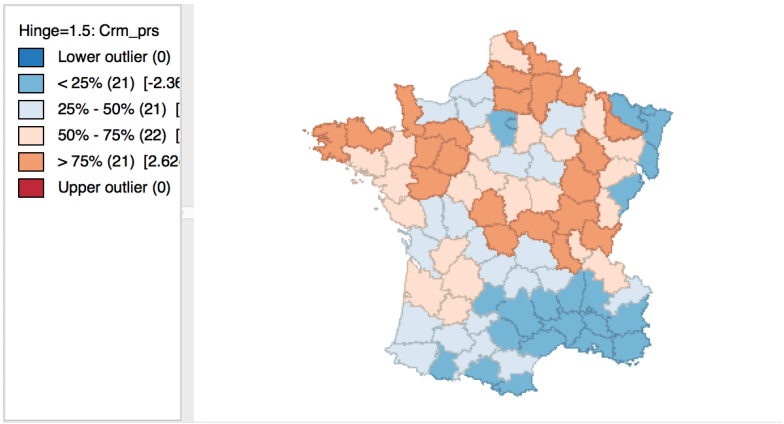
  literacy - Litercy

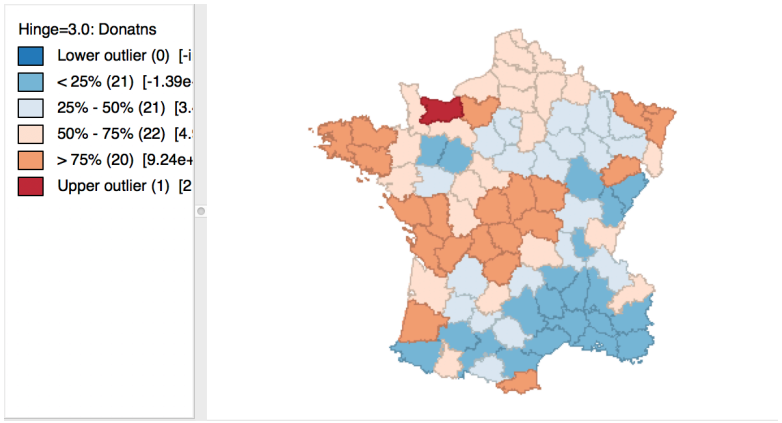
  donations - Donatns

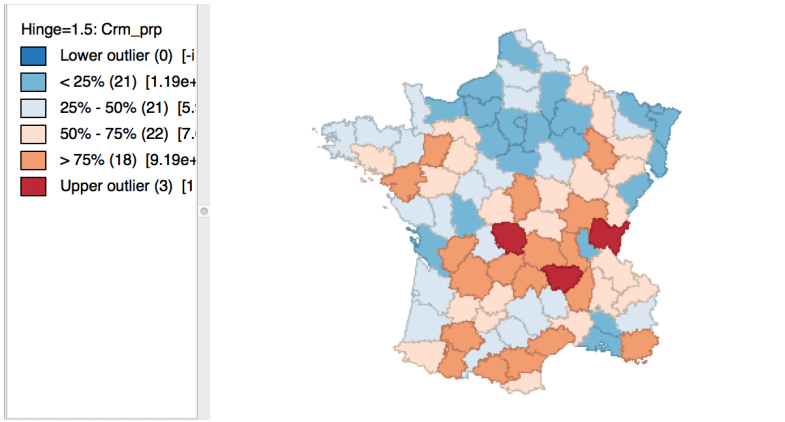
  births out of wedlock (pop/births) - Infants

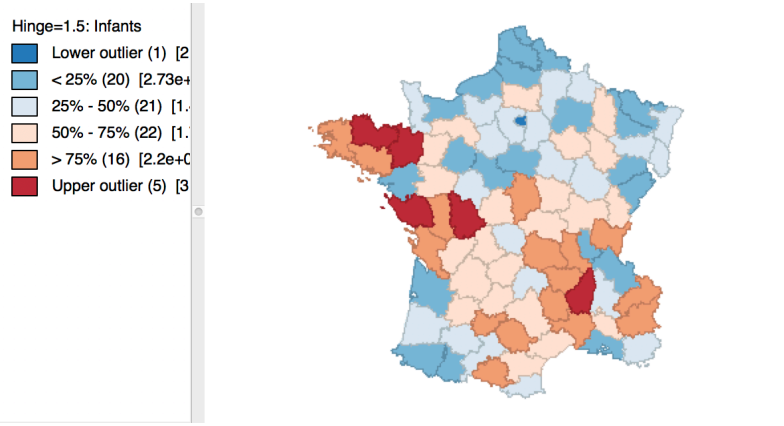
  suicides (pop/suicides) - Suicids

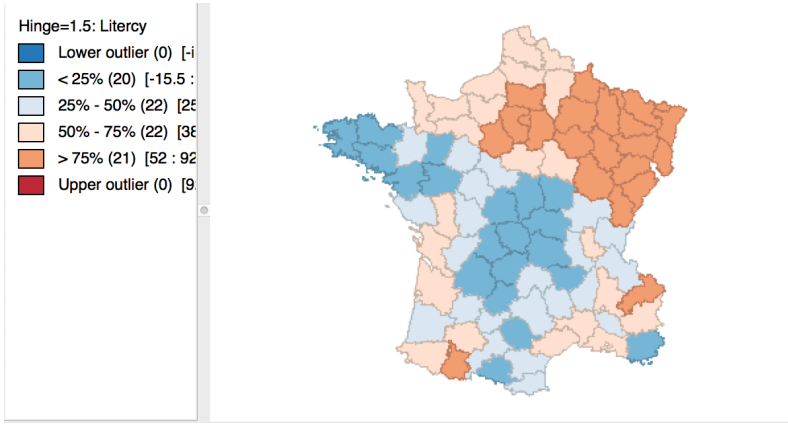
    indicators inverted such that larger values correspond to "better" outcomes

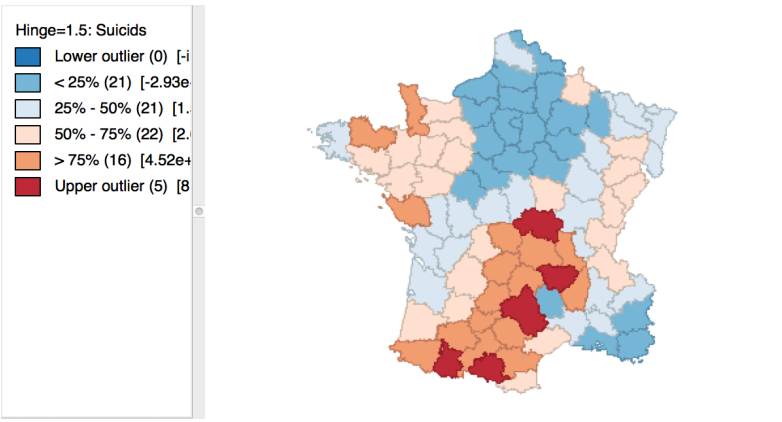












box maps

Table 1: Correlations between the six variables

	Cr. pers.	Cr. prop.	Lit.	Don.	Inf.	Suic.	Moran's I
Crime persons	1.000	0.255	-0.021	0.134	-0.027	-0.134	0.412
Crime property		1.000	-0.363	-0.082	0.278	0.523	0.264
Literacy			1.000	-0.196	-0.412	-0.374	0.718
Donations				1.000	0.159	-0.035	0.353
Infants					1.000	0.289	0.229
Suicides						1.000	0.402

correlation matrix

```
Standard deviation:
1.463034 1.095819 1.049784 0.816680 0.740726 0.583971

Proportion of variance:
0.356745 0.200137 0.183675 0.111161 0.091446 0.056837

Cumulative proportion:
0.356745 0.556882 0.740556 0.851717 0.943163 1.000000

Kaiser criterion: 3.000000

95% threshold criterion: 5.000000
```

```
Eigenvalues:
 2.14047
 1.20082
 1.10205
0.666966
0.548674
0.341022
```

eigenvalue = variance

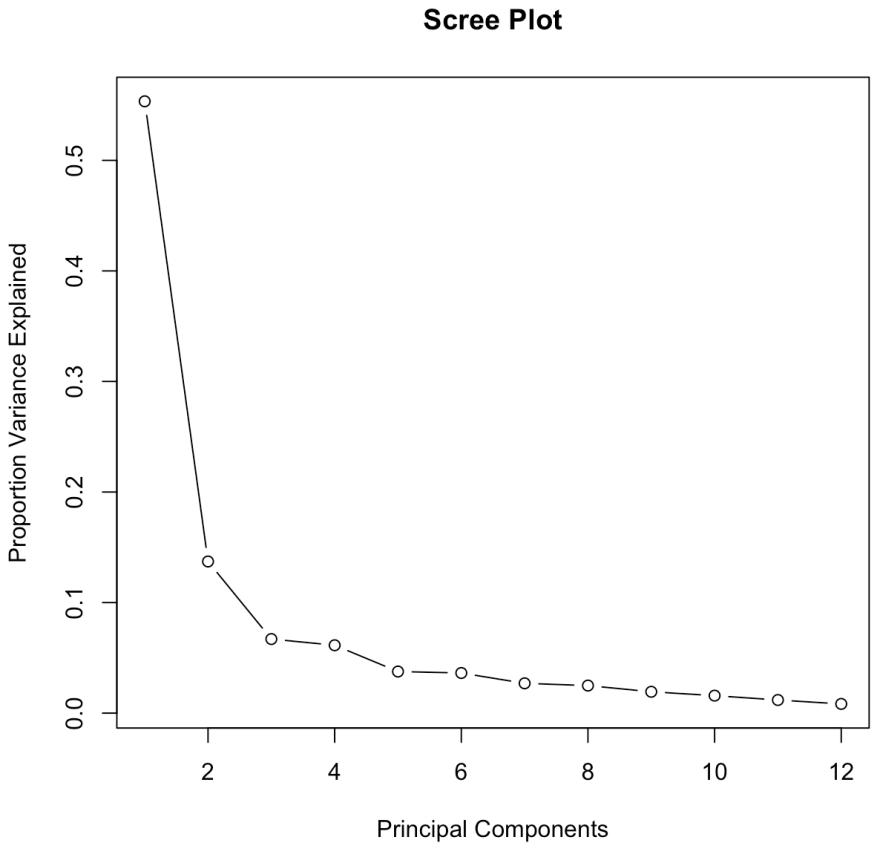
standard deviation = square root of eigenvalue

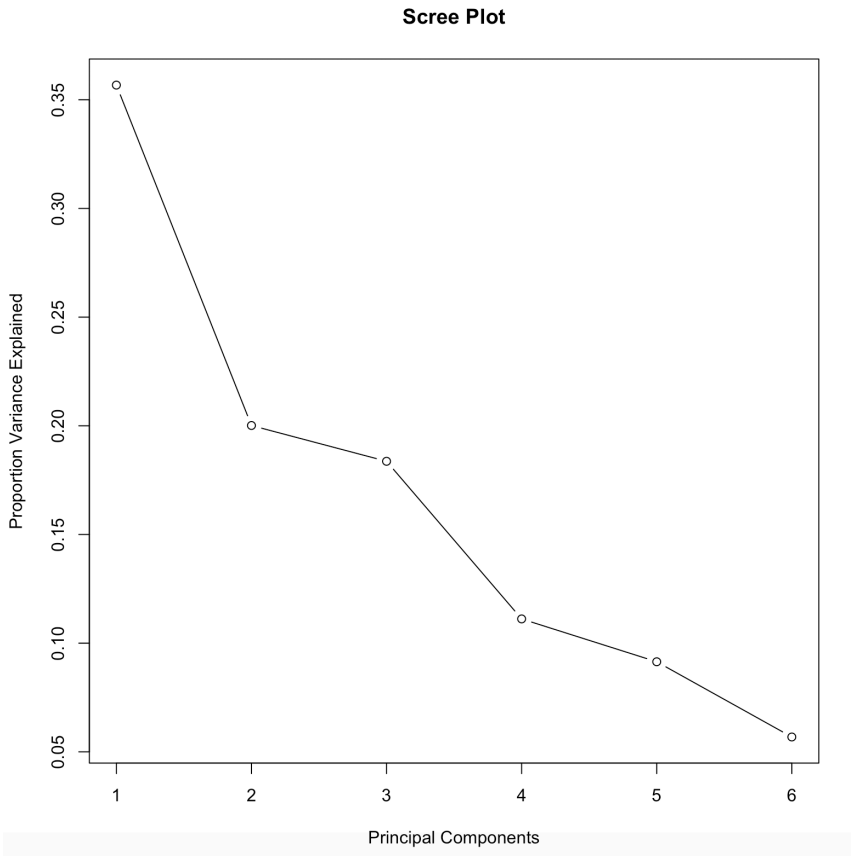
sum of eigenvalues = total variance

2.1405 + 1.2008 + 1.1021 + 0.6667 + 0.5487 + 0.3410 = 5.9997

variance proportion = 2.1405/5.9997 = 0.357

Principal Components and Variance Decomposition

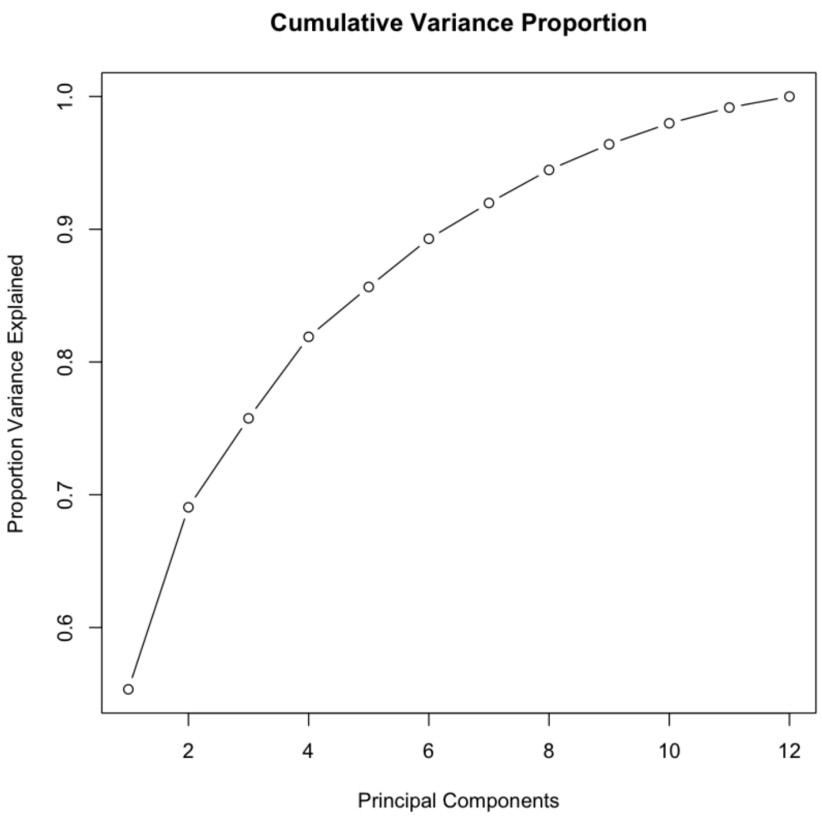


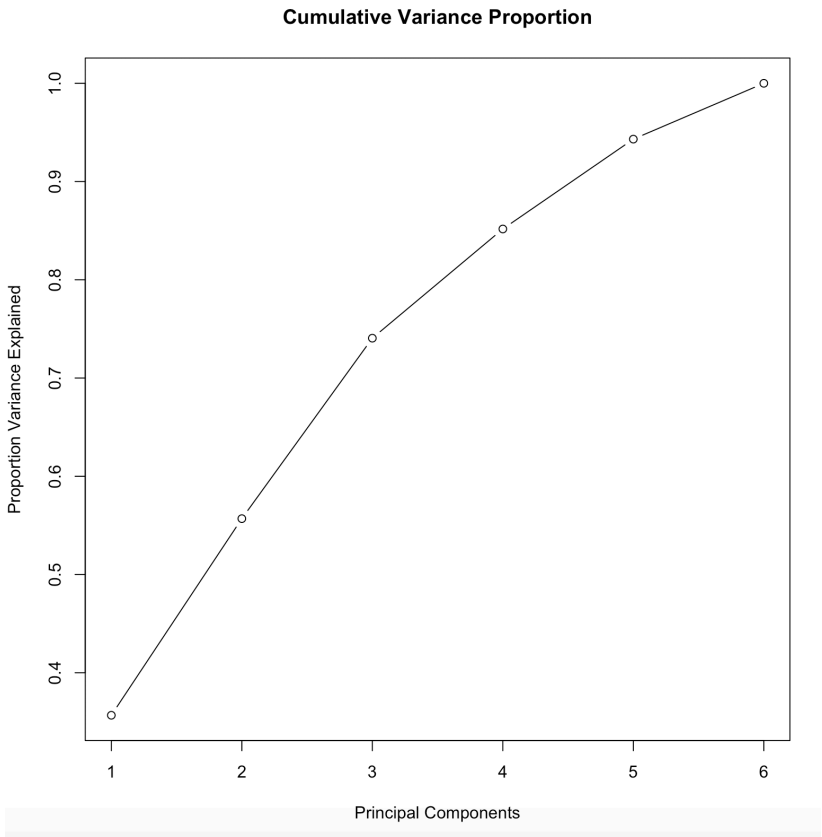


ideal shape

Guerry data

how many components?
elbow in scree plot





ideal shape

Guerry data

cumulative scree plot

Eigenvectors/Variable Loadings:

	PC1	PC2	PC3	PC4	PC5	PC6
crm_prs	-0.0658689	-0.590598	-0.673189	0.139729	-0.010177	-0.417188
crm_prp	-0.512326	0.0883682	-0.476541	-0.0986062	0.138057	0.68836
litercy	0.511753	0.129361	-0.209028	0.00796684	0.821269	0.0559975
donatns	-0.106195	-0.698998	0.413397	-0.472984	0.274207	0.174136
infants	-0.451337	-0.103313	0.323808	0.730309	0.377589	-0.0696003
suicids	-0.50627	0.356902	-0.0168522	-0.462195	0.297643	-0.560189

principal component

$$c_i = a_{i1}\, x_1 + a_{i2}\, x_2 + \ldots + a_{ip}\, x_p$$

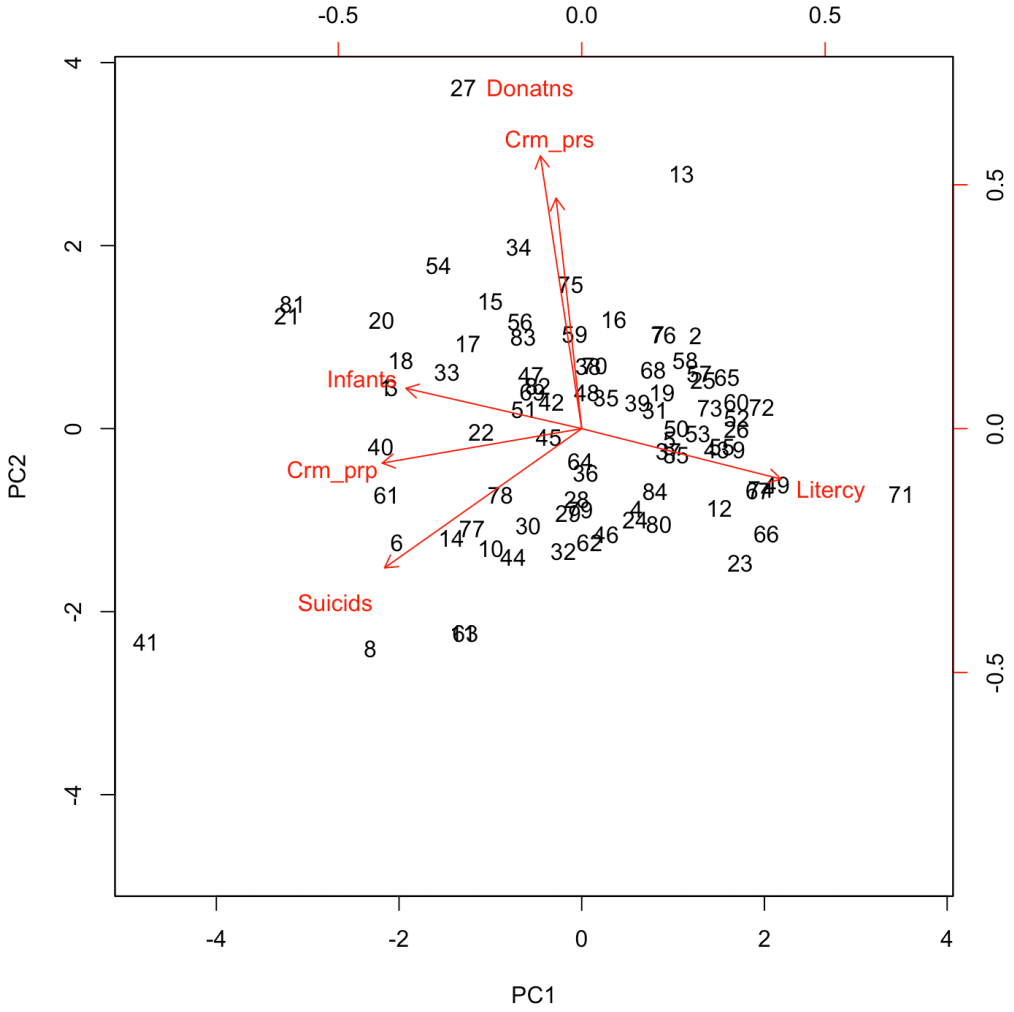
sum of squared loadings = 1

$$0.0043 + 0.2625 + 0.2619 + 0.0113 + 0.2037 + 0.2563 = 1$$

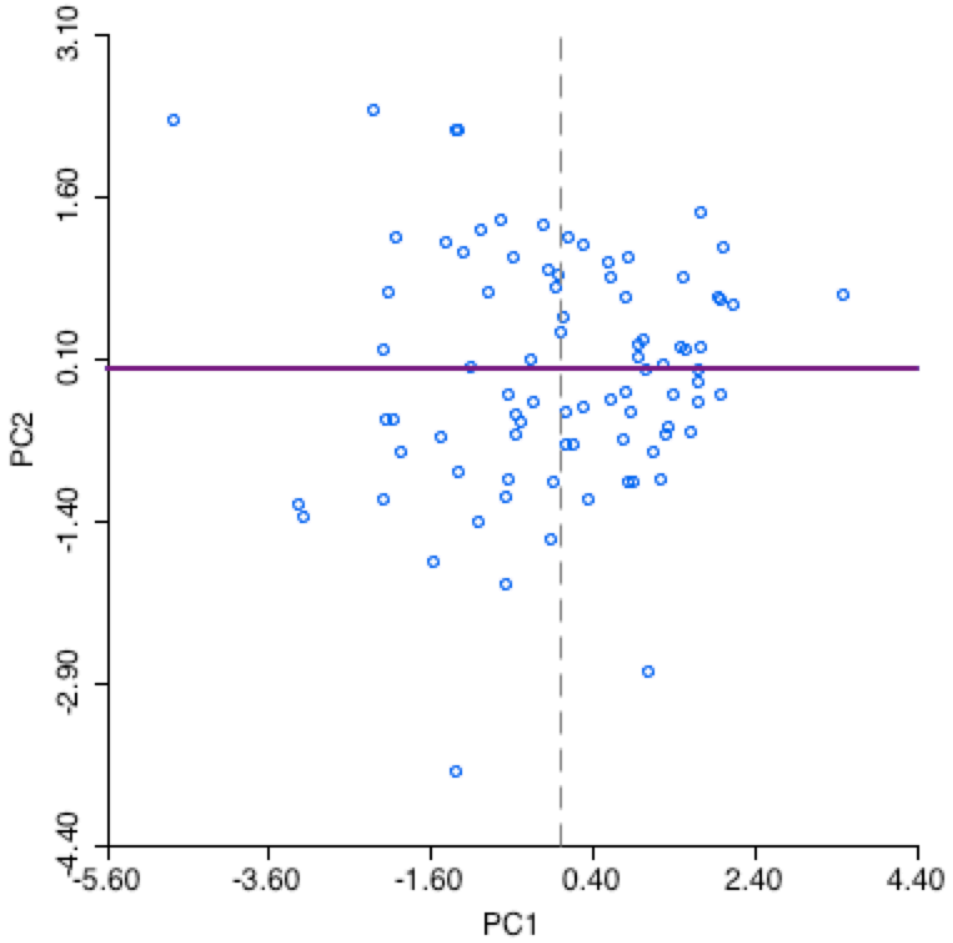
Squared correlations:

	PC1	PC2	PC3	PC4	PC5	PC6
crm_prs	0.00928688	0.418852	0.499429	0.0130218	5.68229e-05	0.0593533
crm_prp	0.561826	0.00937713	0.250265	0.00648504	0.0104576	0.161589
litercy	0.56057	0.020095	0.0481516	4.23341e-05	0.370072	0.00106934
donatns	0.0241388	0.586719	0.188337	0.149209	0.0412544	0.010341
infants	0.436026	0.012817	0.115552	0.355727	0.0782265	0.00165196
suicids	0.548623	0.152959	0.000312987	0.14248	0.0486079	0.107017

squared correlations between variable and pca
= proportion variance of variable explained by each pc



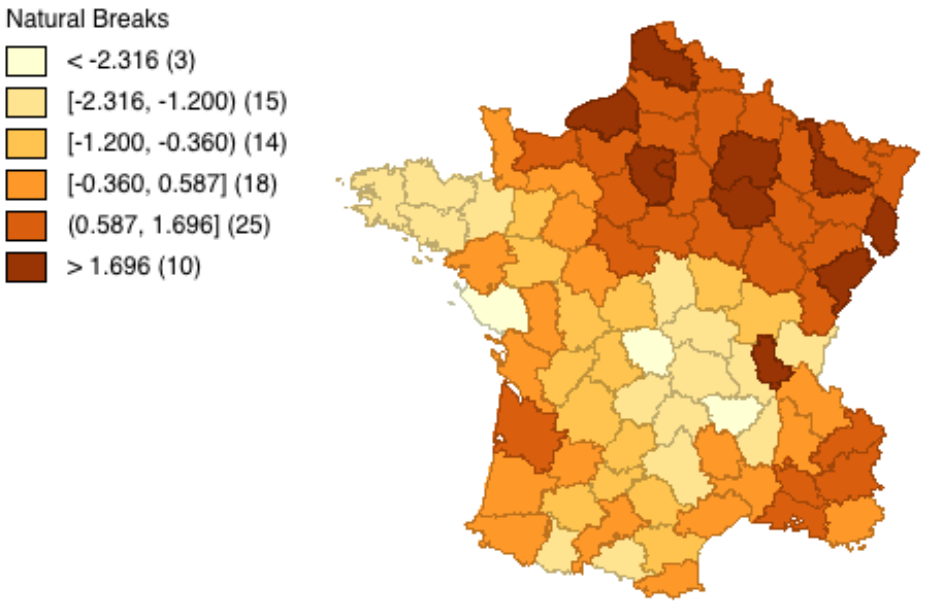
principal components biplot
relative loadings of individual variables

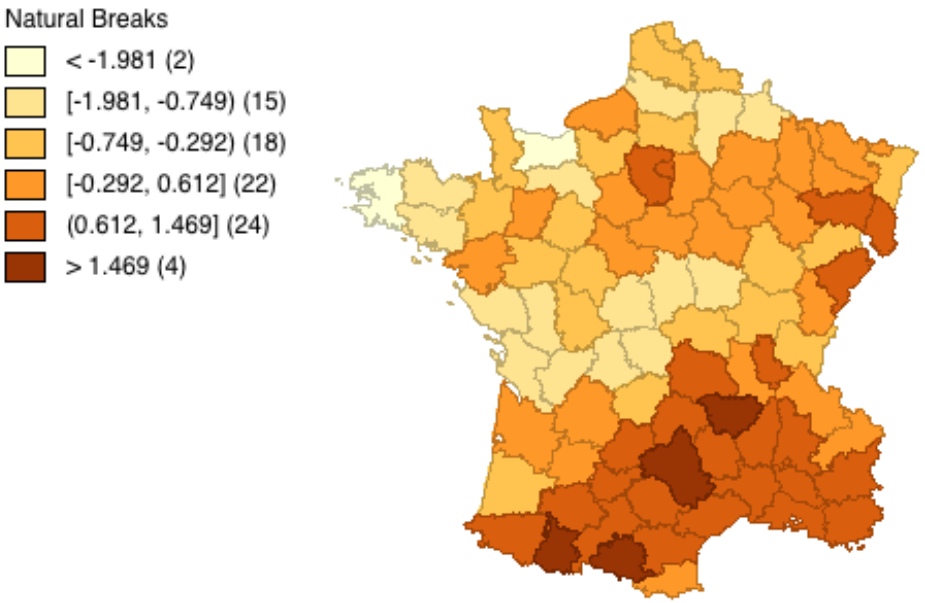


#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
85	0.000	0.000	0.120	0.000	1.000	-0.000	0.082	-0.000	1.000

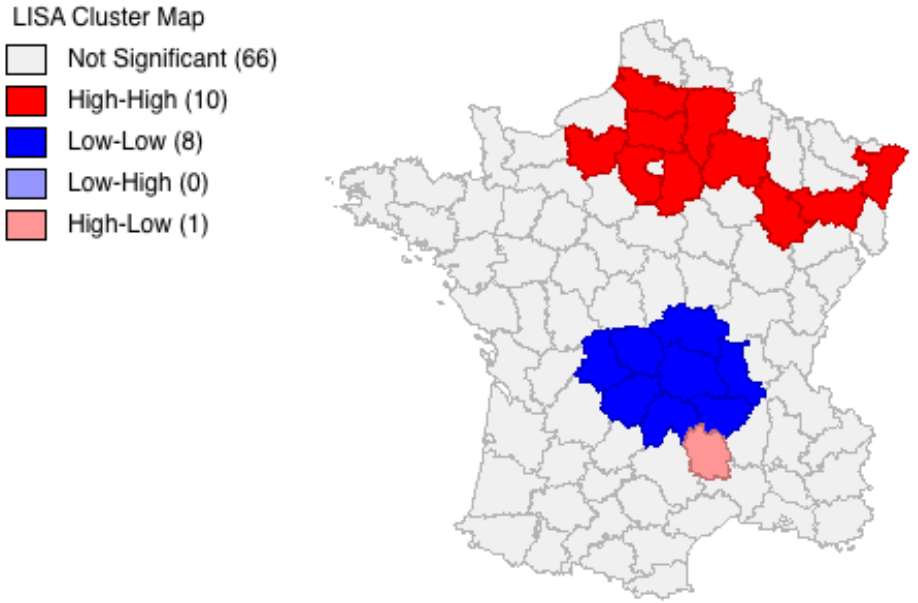
principal components are uncorrelated

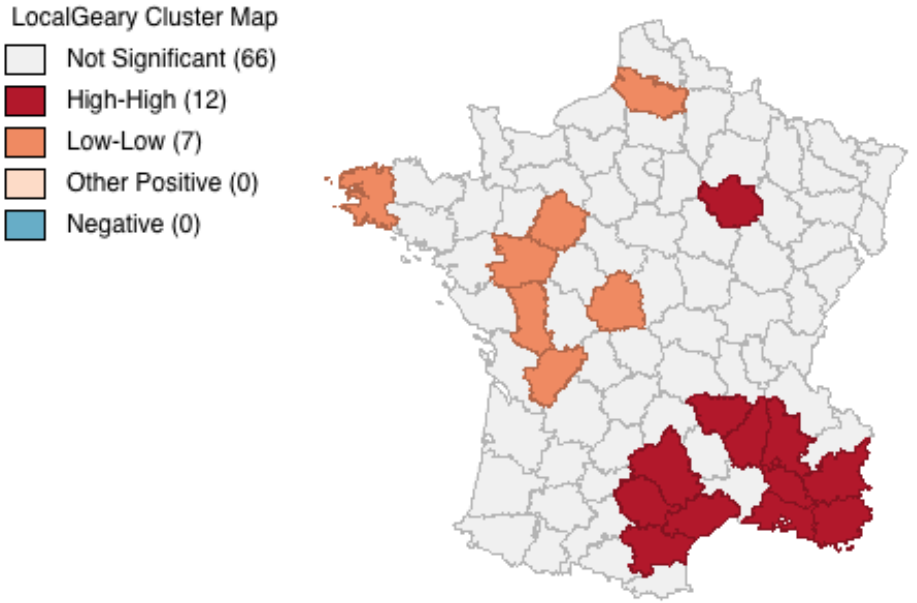
Spatializing the PCA





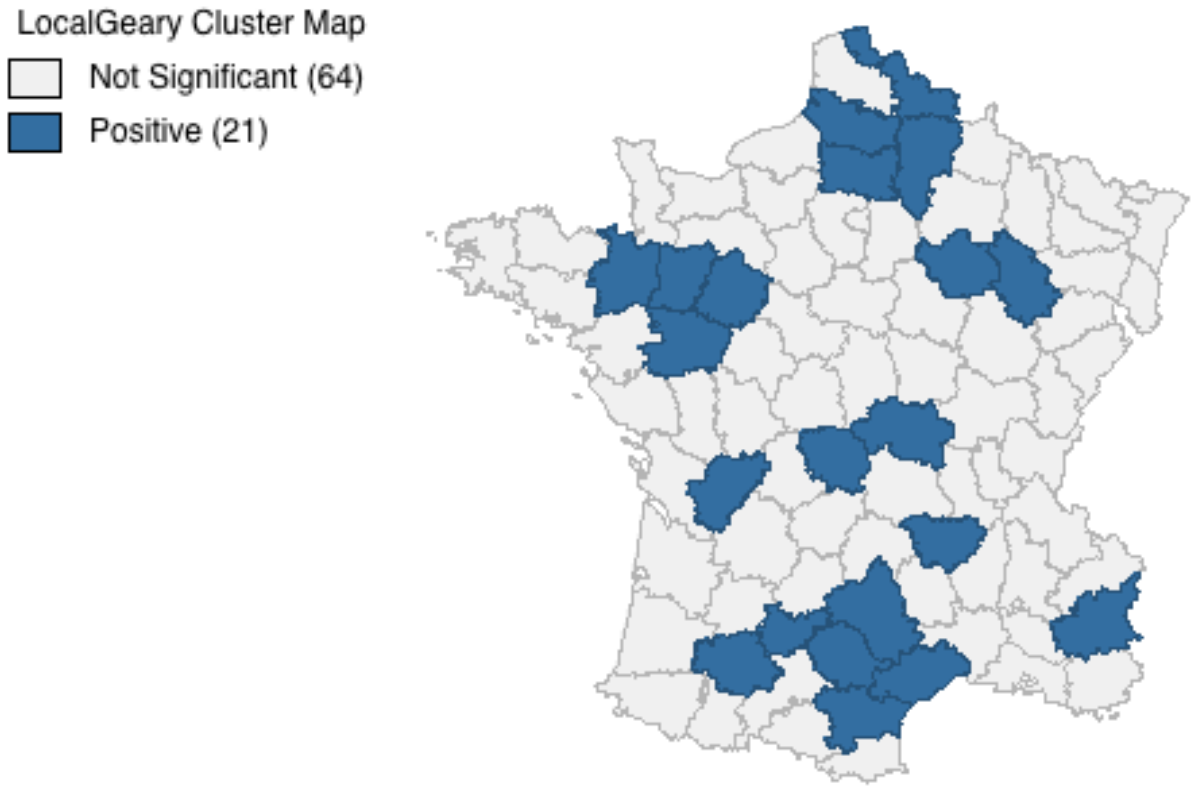
choropleth maps for principal components





pc1 pc2

local Moran and local Geary cluster maps



bivariate local Geary, pc1-pc2

Multidimensional Scaling

- Principle

  n observations are points in a p-dimensional data hypercube

  p-variate distance or dissimilarity between all pairs of points
  (e.g., Euclidean distance in p dimensions)

  represent the n observations in a lower dimensional space (at most p - 1) while respecting the pairwise distances

- More formally

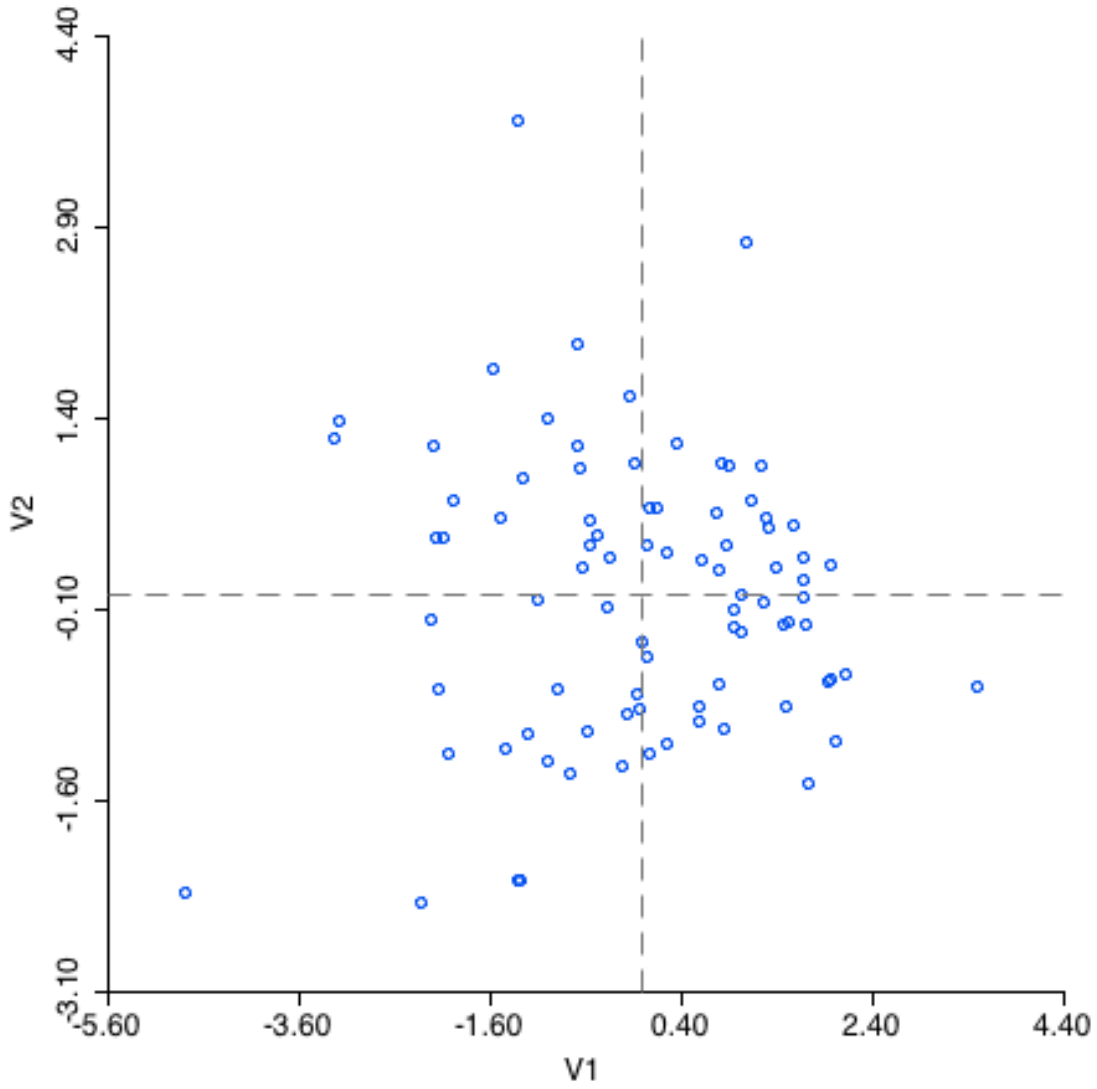
n by n distance or dissimilarity matrix D

$d_{ij} = ||x_i - x_j||$ Euclidean distance in p dimensions

find values $z_1, z_2, \ldots z_n$ in k-dimensional space (with $k << p$) that minimize the stress function

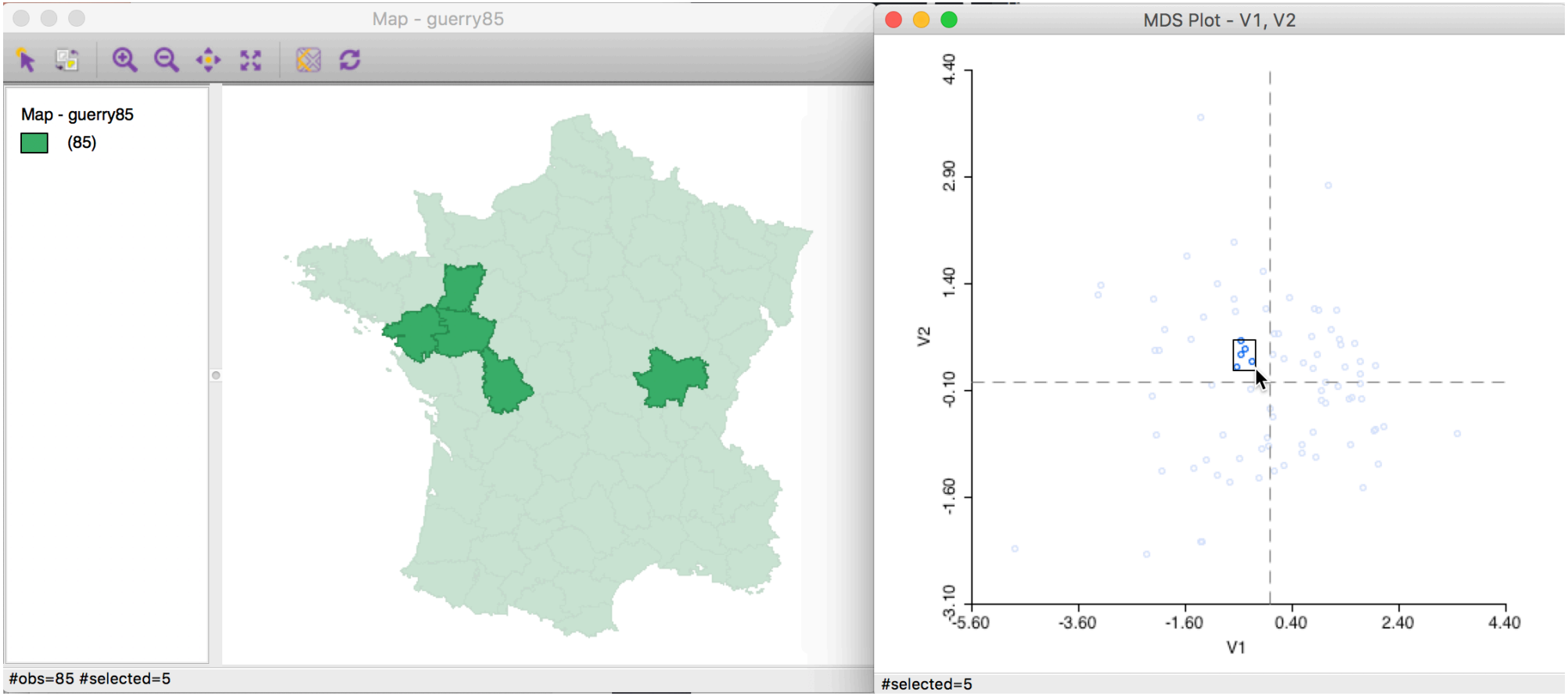
$$S(z) = \Sigma_{i,j} (d_{ij} - ||z_i - z_j||)^2$$

least squares or Kruskal-Shephard scaling

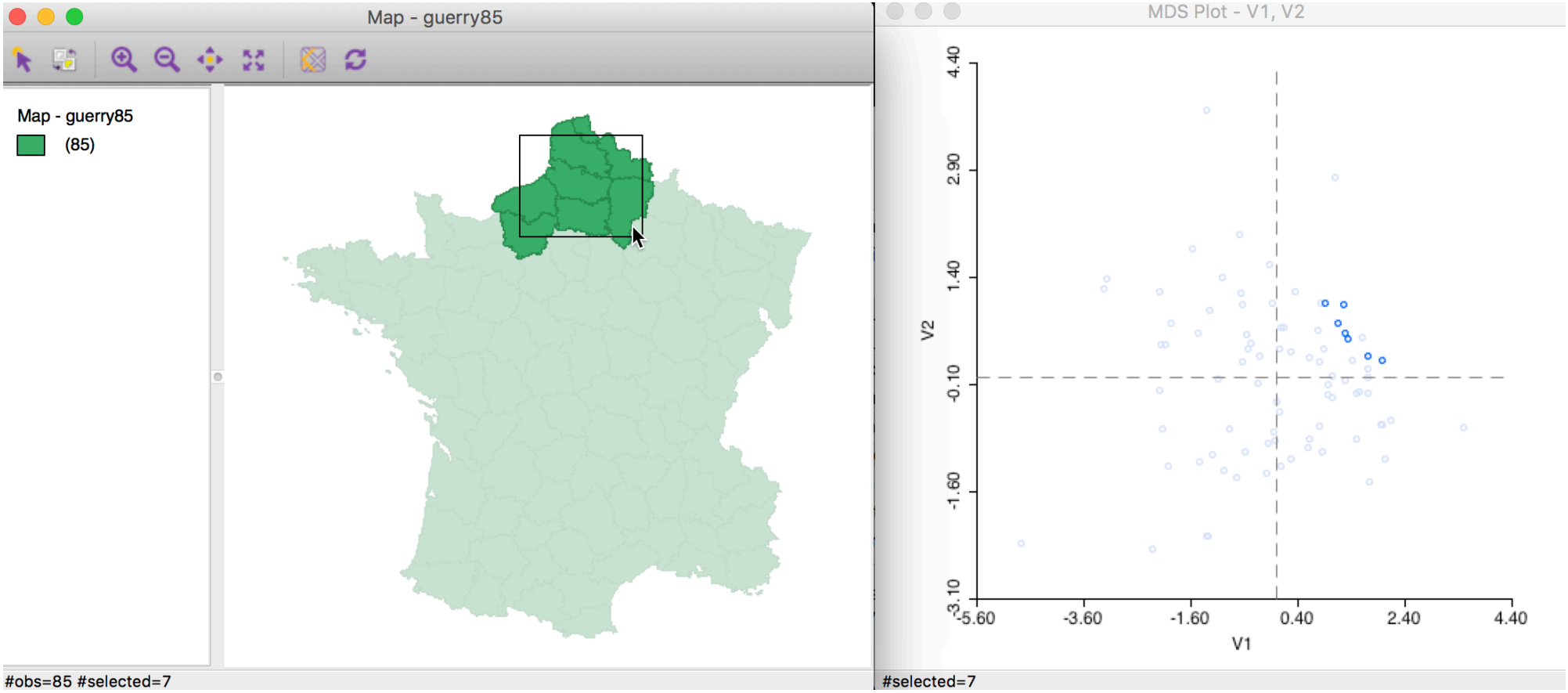


MDS plot

Spatializing MDS



neighbors in multivariate space are not always neighbors in geographic space



brushing the MDS plot

Classical Clustering Methods

- Principle

  grouping of similar observations

  maximize within-group similarity

  minimize between-group similarity

  or, maximize between-group dissimilarity

  each observation belongs to one and only one group

- Issues
  - similarity criterion
    - Euclidean distance, correlation
  - how many groups
    - many "rules of thumb"
  - computational challenges
    - combinatorial problem, NP hard
    - n observations in k groups
      - k^n possible partitions
      - k = 4 with n = 85, $k^n = 1.5 \times 10^{51}$
    - no guarantee of a global optimum

- Dissimilarity

  data x_{ij} with variable j at observation i

  distance by attribute

  - $d_j(x_{ij},x_{i'j}) = (x_{ij} - x_{i'j})^2$

    typically Euclidean distance

  distance between objects

  $D(x_i,x_{i'}) = \Sigma_j\ d_j(x_{ij},x_{i'j})$

  weighted distance between objects

  $D(x_i,x_{i'}) = \Sigma_j\ w_j\ d_j(x_{ij},x_{i'j})$ with $\Sigma_j\ w_j = 1$

  reflect relative importance of attributes

  equal weighting = inverse attribute variance

  is automatic for standardized variables

- Two main approaches
  - hierarchical clustering
    - start from bottom
    - determine number of clusters later
  - partitioning clustering (k-means)
    - start with random assignment to k groups
    - number of clusters pre-determined
  - many clustering algorithms

Hierarchical Clustering

- Algorithm

  find two observations that are closest (most similar)

    they form a cluster

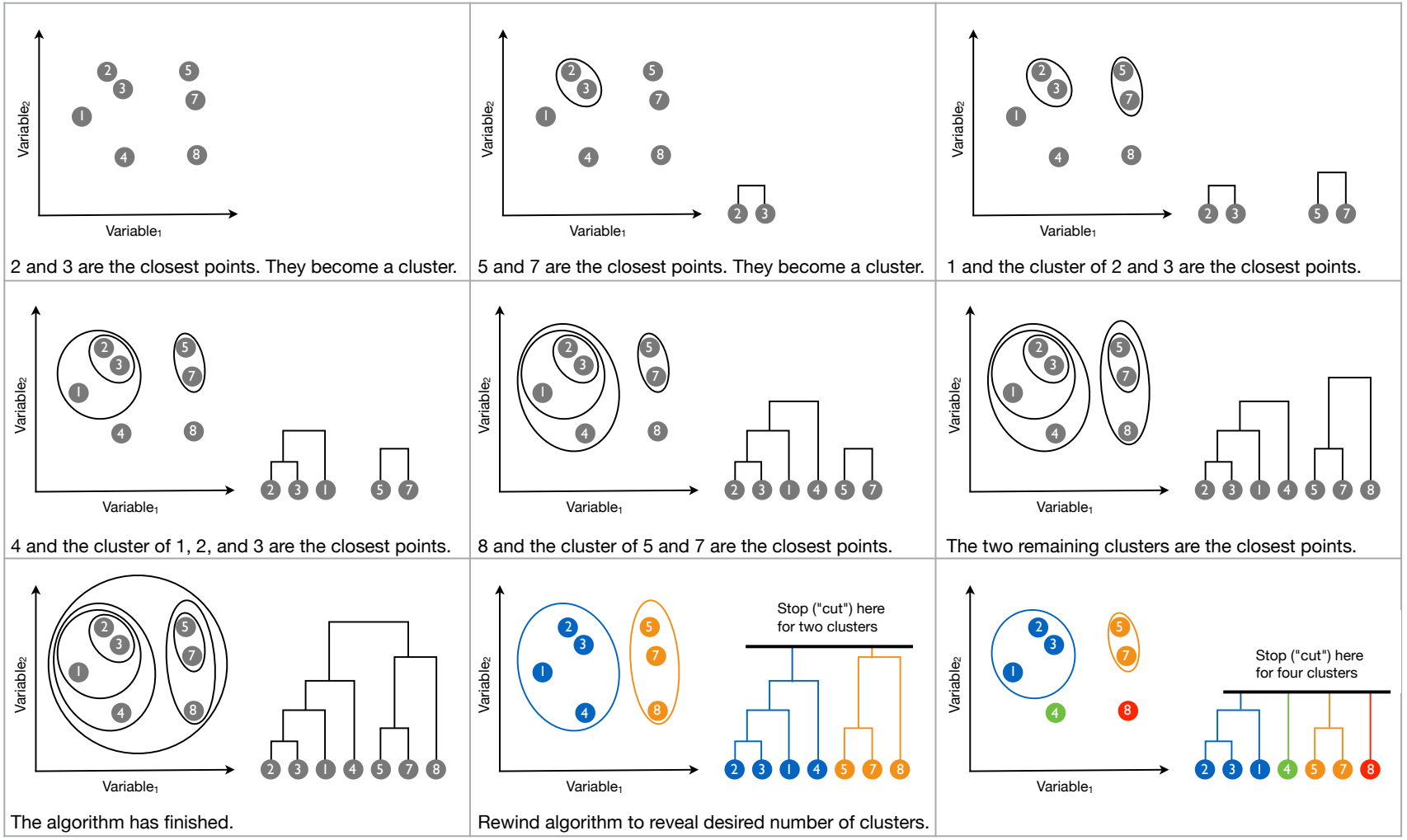
  determine the next closest pair

    include the existing clusters in the comparisons

  continue grouping until all observations have been included

  result is a dendrogram

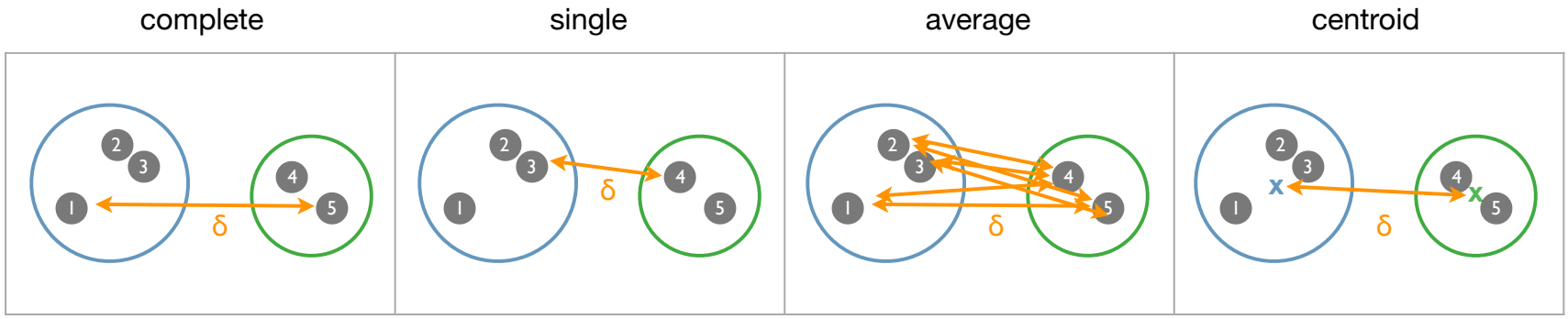
    a hierarchical tree structure



hierarchical clustering algorithm

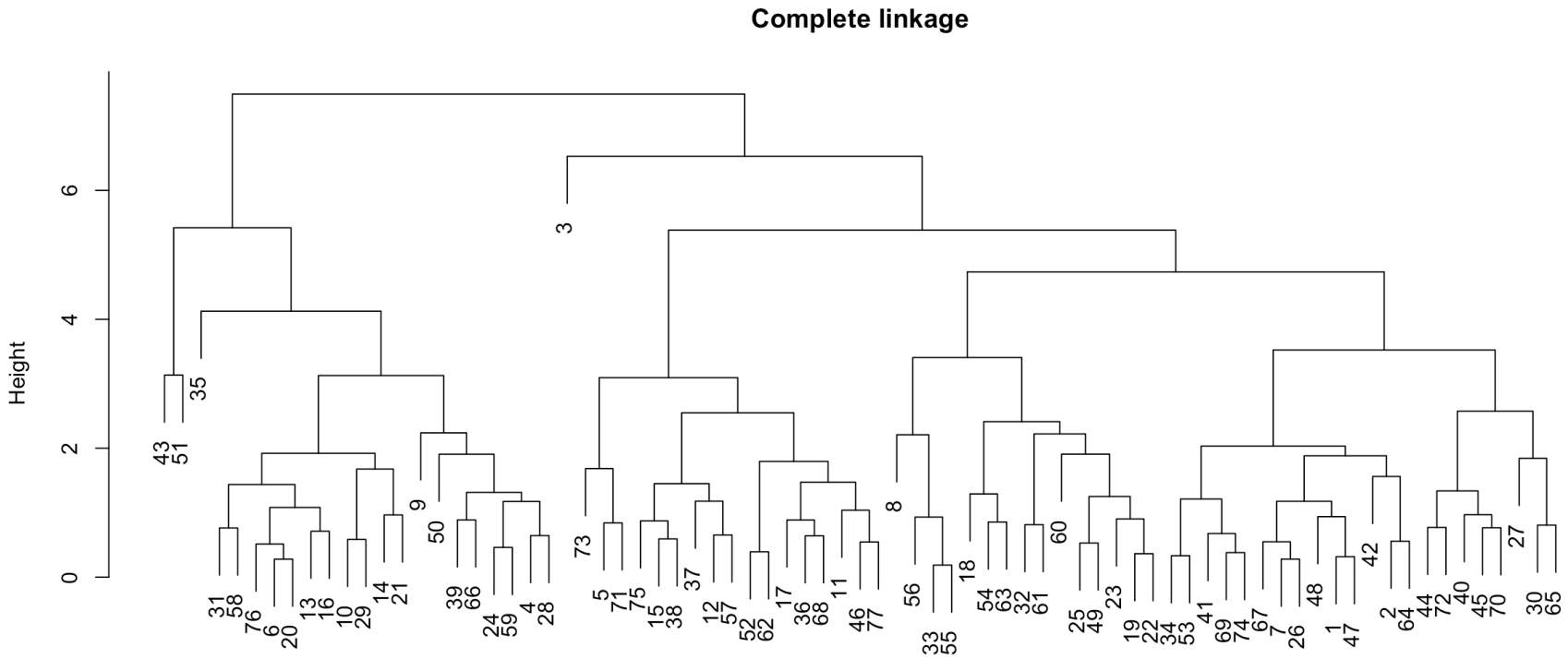
Source: Grolemund and Wickham (2016)

- Practical issues
  - measure of similarity (dissimilarity) between clusters = linkage
    - complete
      - compact clusters
    - single
      - elongated clusters, singletons
    - average
    - centroid
    - others …
  - how many clusters
    - where to cut the tree

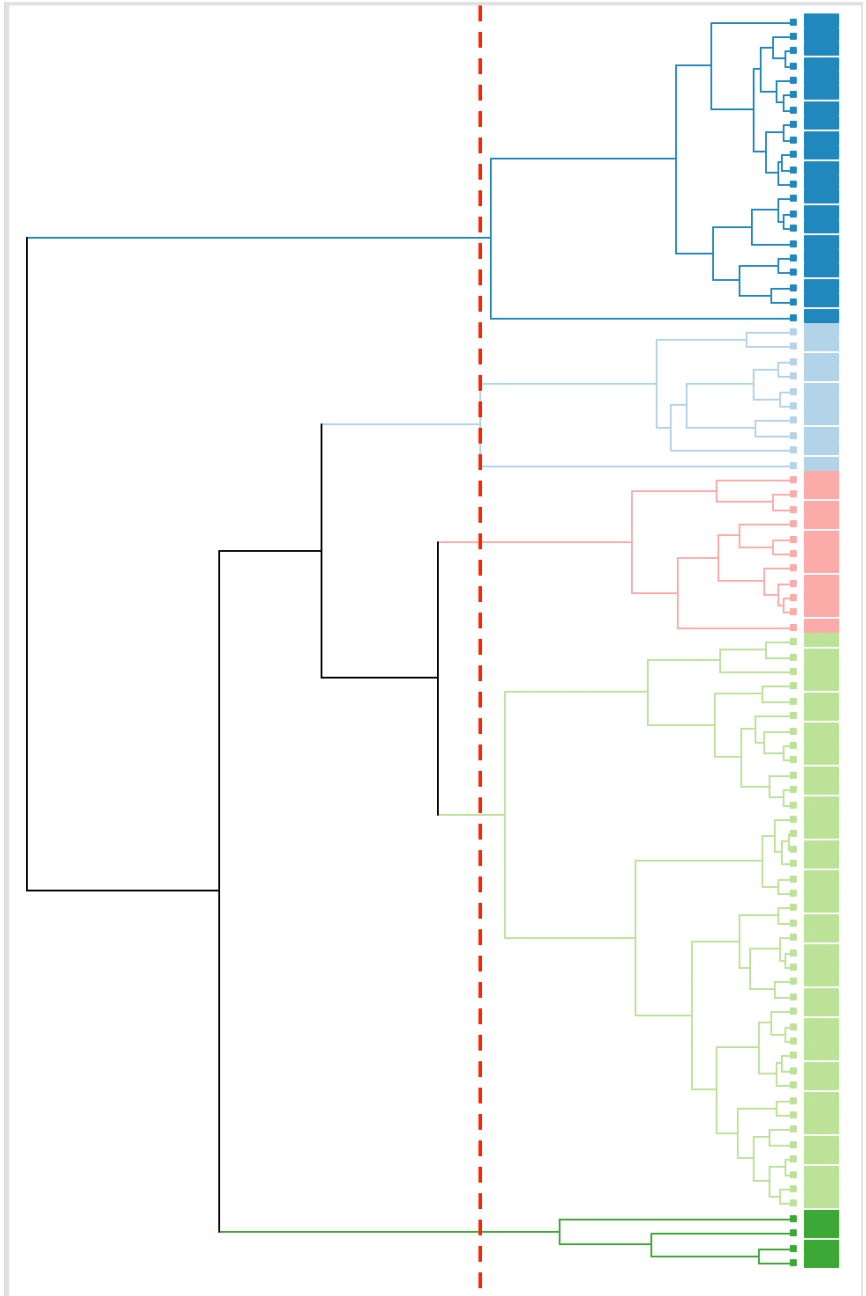


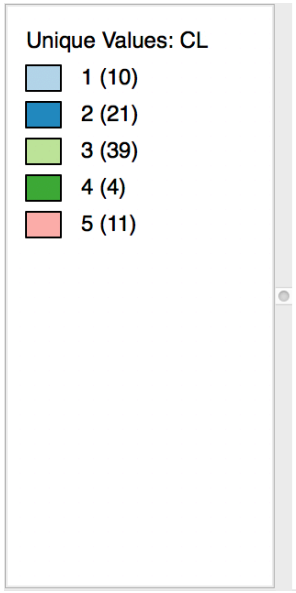
types of linkages

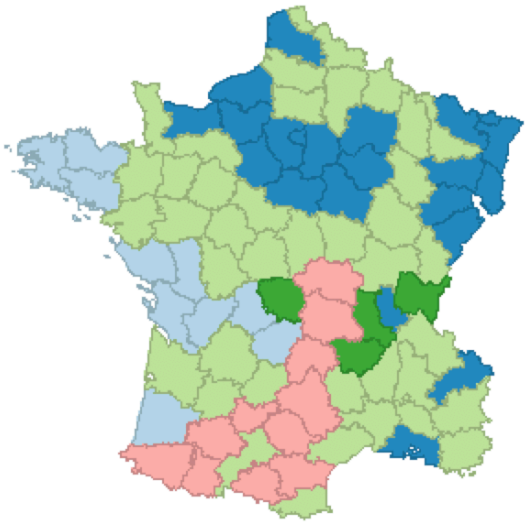
Source: Grolemund and Wickham (2016)



example: complete linkage dendrogram





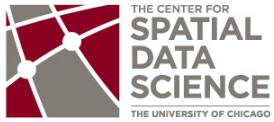


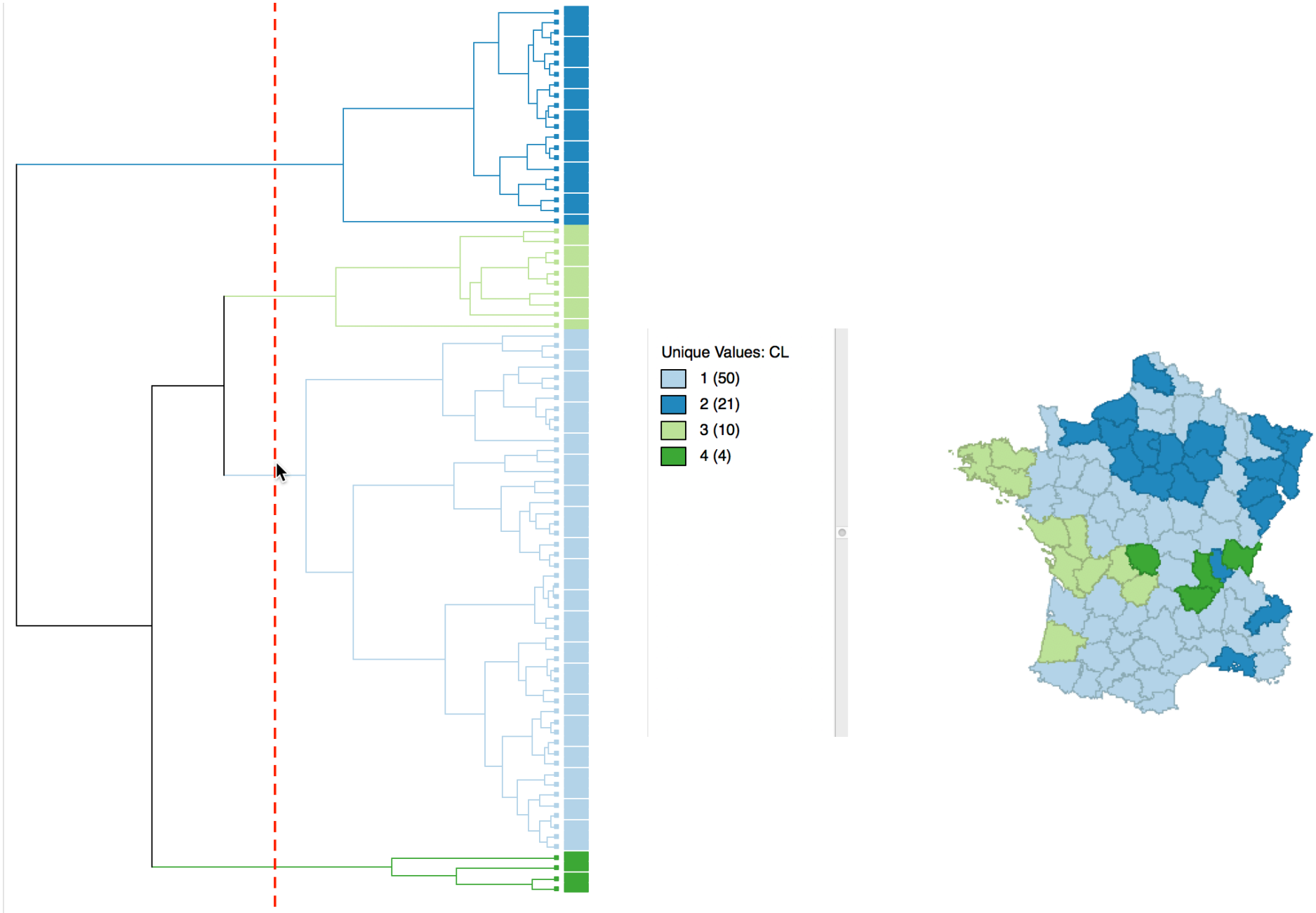
hierarchical clustering: dendrogram and map
complete linkage, k = 5

- Characteristics of Clusters
  - total sum of squares (total SS)
    - relative to overall mean
  - within sum of squares (within SS)
    - for each cluster, relative to mean of each cluster
  - between sum of squares (between SS)
    - total SS - sum of within SS
  - measure of cluster quality
    - ratio of between SS / total SS
    - higher is better

```
Number of cluster:  5
Transformation: Standardize
Method: Complete-linkage
Distance function:  Euclidean
Cluster centers:
|   |Crm_prs   |Crm_prp   |Litercy    |Donatns   |Infants    |Suicids   |
|---|----------|----------|-----------|----------|-----------|----------|
|C1|0.195823  |0.0440721|-0.856997  |1.65638   |1.04385    |0.0653781|
|C2|-0.457972|-0.915078|0.969721   |0.0629438|-0.702665  |-0.609467|
|C3|0.27488   |0.0540026|-0.0845751|-0.310747|0.00451298|-0.322695|
|C4|1.01727   |2.89538   |-0.667716  |-0.23679  |0.908793   |1.5985    |
|C5|-0.648204|0.462572  |-0.529535  |-0.438119|0.0460246  |1.66692   |

The total sum of squares:   504
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|45.2181             |
|C2|56.2374             |
|C3|136.787             |
|C4|21.5168             |
|C5|30.9984             |

The total within-cluster sum of squares:    290.757
The between-cluster sum of squares: 213.243
The ratio of between to total sum of squares:   0.423101
```

cluster characteristics with k=5

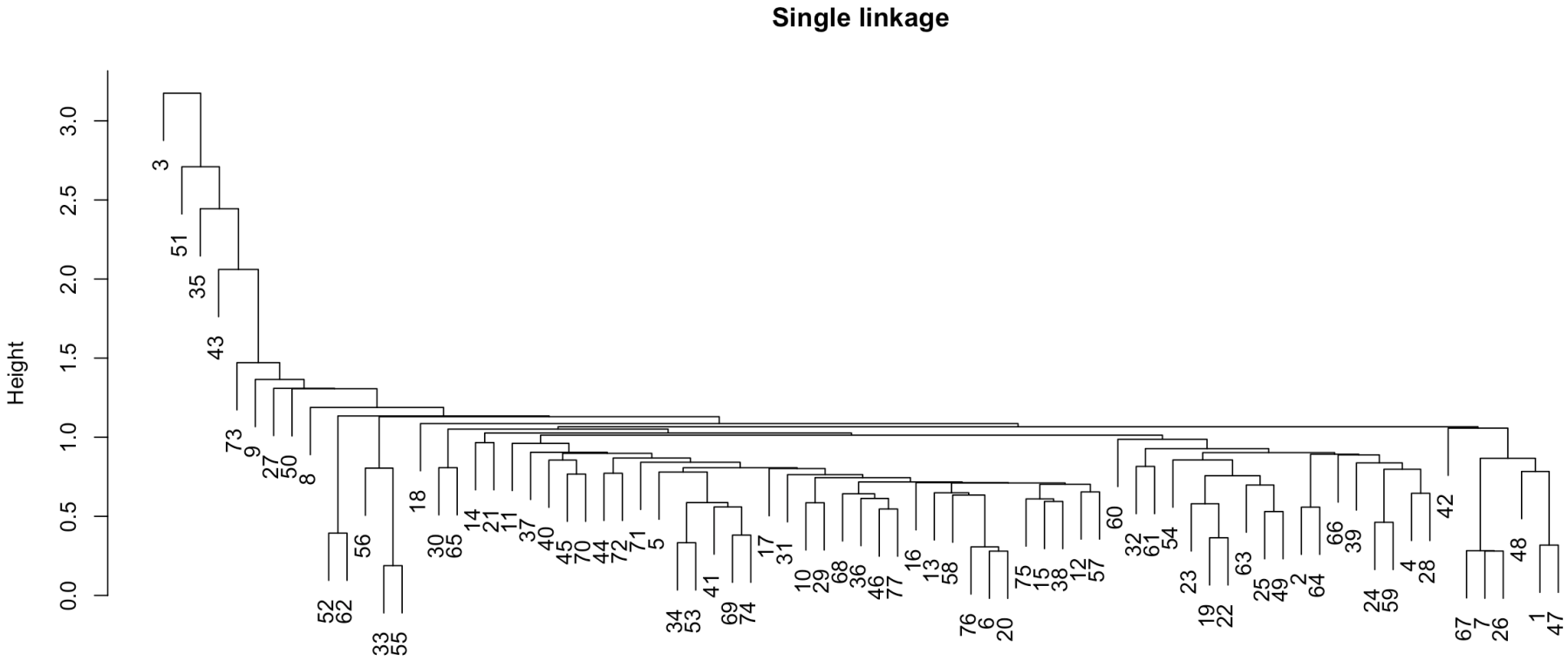




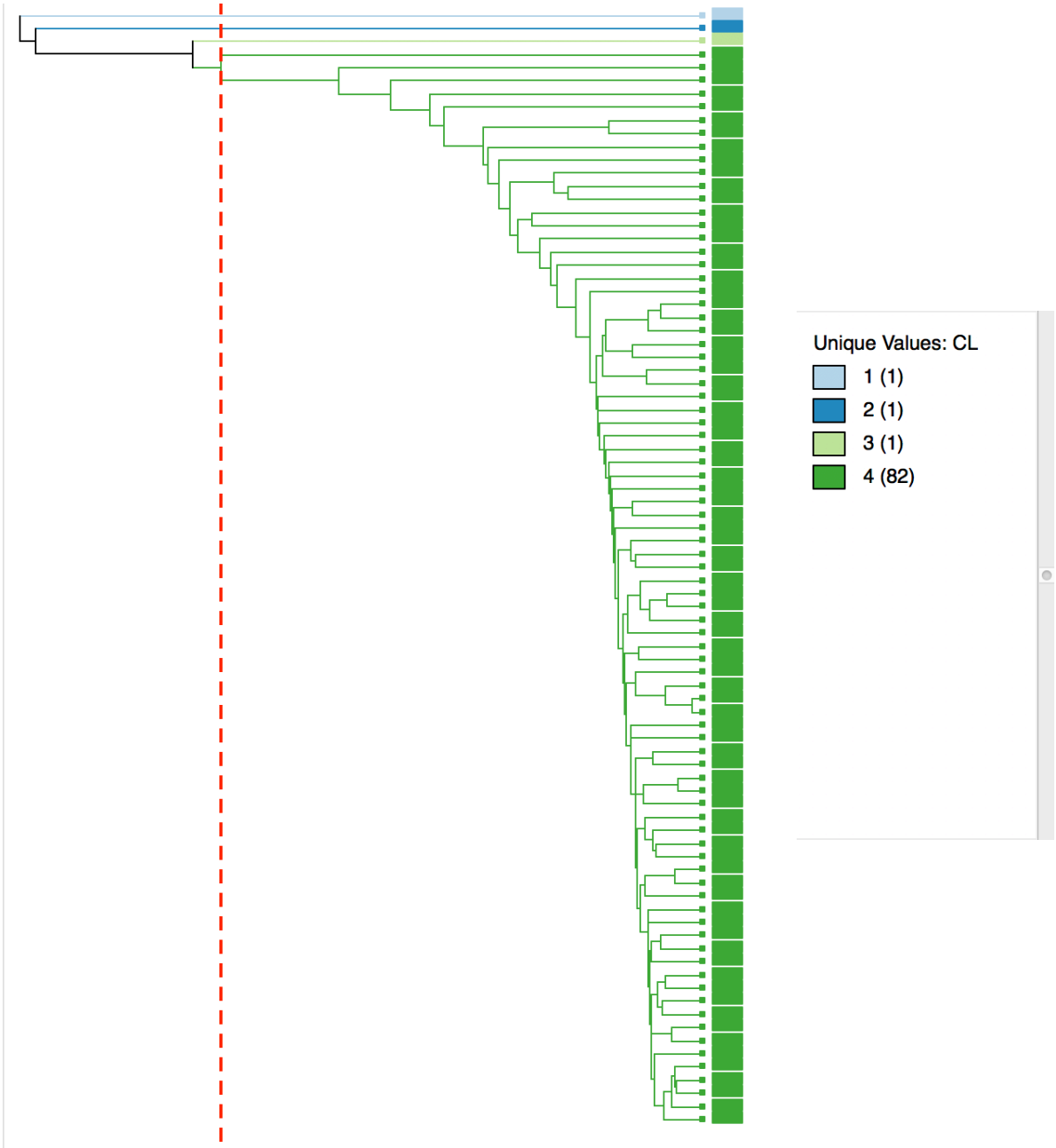
hierarchical clustering: dendrogram and map
complete linkage, k = 4

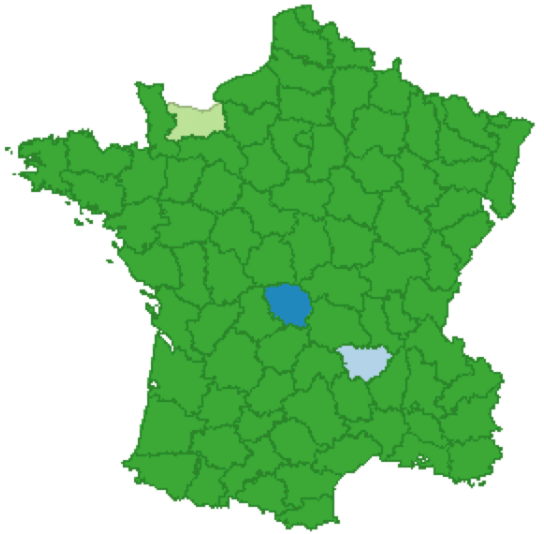
```
Number of cluster:  4
Transformation: Standardize
Method: Complete-linkage
Distance function:  Euclidean
Cluster centers:
|   |Crm_prs   |Crm_prp   |Litercy   |Donatns   |Infants   |Suicids   |
|---|----------|----------|----------|----------|----------|----------|
|C1|0.0718018|0.143888  |-0.182466|-0.338769|0.0136455|0.115021  |
|C2|-0.457972|-0.915078|0.969721  |0.0629438|-0.702665|-0.609467|
|C3|0.195823  |0.0440721|-0.856997|1.65638   |1.04385   |0.0653781|
|C4|1.01727   |2.89538   |-0.667716|-0.23679  |0.908793  |1.5985    |

The total sum of squares:   504
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|212.345             |
|C2|56.2374             |
|C3|45.2181             |
|C4|21.5168             |

The total within-cluster sum of squares:    335.318
The between-cluster sum of squares: 168.682
The ratio of between to total sum of squares:   0.334687
```

cluster characteristics with k=4



example: single linkage dendrogram

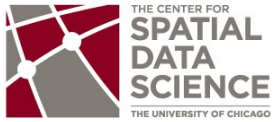




hierarchical clustering: dendrogram and map
single linkage, k = 4

```
Number of cluster:  4
Transformation: Standardize
Method: Single-linkage
Distance function:  Euclidean
Cluster centers:
|   |Crm_prs    |Crm_prp    |Litercy   |Donatns    |Infants    |Suicids    |
|---|-----------|-----------|----------|-----------|-----------|-----------|
|C1|-0.519361  |3.3332     |-1.04054  |-0.817832  |1.3597     |4.02321    |
|C2|2.33628    |4.05222    |-0.925827|0.878772   |-0.521993  |1.31138    |
|C3|-0.326601  |-1.10914   |0.737557  |4.34004    |-1.12986   |-0.149525  |
|C4|-0.0181746|-0.0765401|0.0149855|-0.0536705|0.00356294|-0.0632324|

The total sum of squares:   504
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|0                  |
|C2|0                  |
|C3|0                  |
|C4|424.226            |

The total within-cluster sum of squares:    424.226
The between-cluster sum of squares: 79.7735
The ratio of between to total sum of squares:   0.158281
```

cluster characteristics with k=4, single linkage



k-Means Clustering

- Algorithm

  randomly assign n observations to k groups

  compute group centroid (or other representative point)

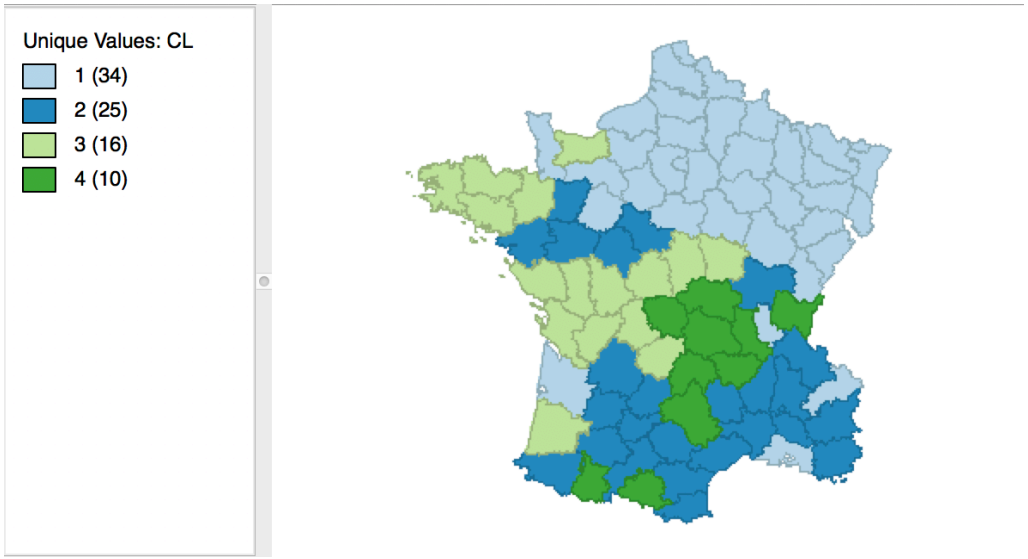
  assign observations to closest centroid

  iterate until convergence

k-means clustering algorithm

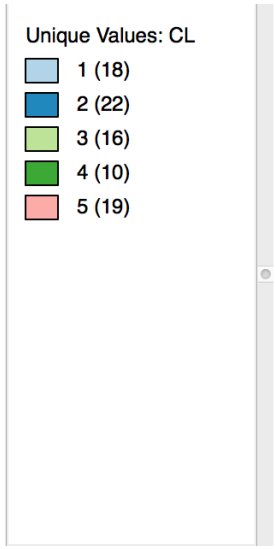
Source: Grolemund and Wickham (2016)

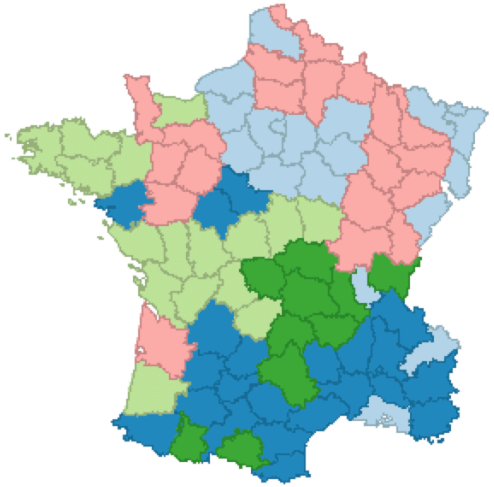
- Practical issues
  - which k to select ?
    - compare solutions on within-group and between-group similarities (sum of squares)
  - sensitivity to starting point
    - use several random assignments and pick the best
  - avoid local optima
    - sensitivity analysis
  - replicability
    - set random seed




```
Number of cluster:  4
Transformation: Standardize
Initialization method:  KMeans++
Initialization re-runs: 50
Maximal iterations: 1000
Method: Arithmetic Mean
Distance function:  Euclidean
Cluster centers:
|   |Crm_prs   |Crm_prp    |Litercy   |Donatns   |Infants   |Suicids    |
|---|----------|-----------|----------|----------|----------|-----------|
|C1|0.174468  |-0.570962  |0.989585  |-0.205957|-0.576296|-0.570768  |
|C2|-0.365388|0.194705   |-0.486817|-0.567548|0.047341  |-0.038075  |
|C3|0.245556  |-0.0788772|-0.83262  |1.49052   |0.895418  |-0.0511492|
|C4|-0.018624|1.50414    |-0.667716|-0.229542|0.34602   |2.06437    |

The total sum of squares:   504
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|81.2921             |
|C2|64.1209             |
|C3|77.7591             |
|C4|62.55               |

The total within-cluster sum of squares:    285.722
The between-cluster sum of squares: 218.278
The ratio of between to total sum of squares:   0.433091
```

k-means, k=4





```
Number of cluster:  5
Transformation: Standardize
Initialization method:  KMeans++
Initialization re-runs: 50
Maximal iterations: 1000
Method: Arithmetic Mean
Distance function:  Euclidean
Cluster centers:
|   |Crm_prs   |Crm_prp    |Litercy   |Donatns   |Infants   |Suicids    |
|---|----------|-----------|----------|----------|----------|-----------|
|C1|-0.683923|0.152655   |-0.438283|-0.546343|0.0942548|0.00249999|
|C2|1.09295   |0.0691862  |0.656049  |-0.318068|-0.394359|-0.381065  |
|C3|-0.525686|-1.02513   |0.954244  |-0.193889|-0.687093|-0.702225  |
|C4|0.245556  |-0.0788772|-0.83262  |1.49052   |0.895418  |-0.0511492|
|C5|-0.018624|1.50414    |-0.667716|-0.229542|0.34602   |2.06437    |

The total sum of squares:   504
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|46.3786            |
|C2|36.232             |
|C3|30.3571            |
|C4|77.7591            |
|C5|62.55              |

The total within-cluster sum of squares:    253.277
The between-cluster sum of squares: 250.723
The ratio of between to total sum of squares:   0.497467
```

k-means, k=5

k=4	Total SS	Within SS	Between SS	Ratio B/T
complete linkage	504	335.3	168.7	0.335
single linkage	504	424.2	79.8	0.158
k-means	504	285.7	218.3	0.433

summary - cluster characteristics