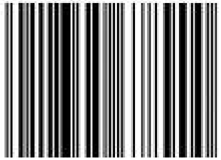
Social Geography: A study in TDA

Andrew Banman

TDA: topological data analysis

1. Collect data
2. Build structure
3. Calculate homology
4. Interpret results

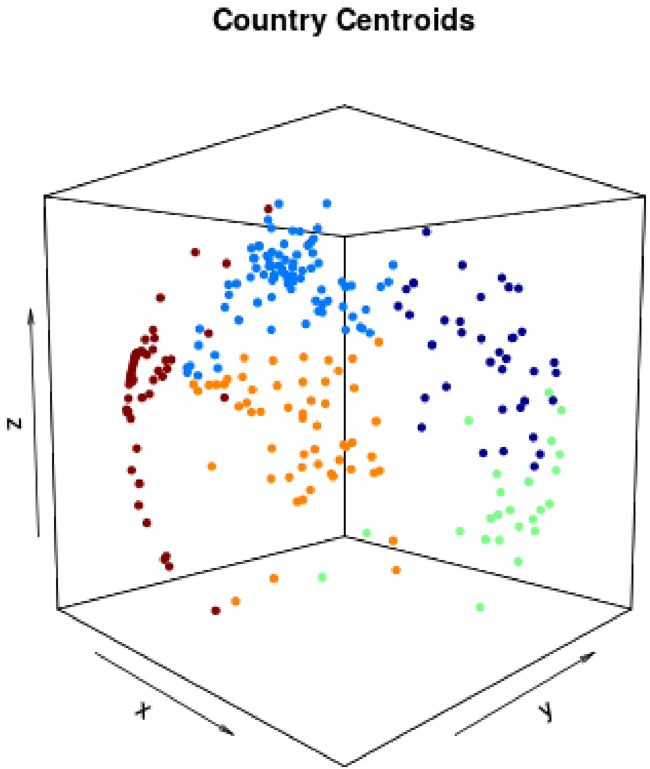


Mind the gap

- Children per woman (total fertility)
- CO2 emissions (tonnes per person)
- Income per person (GDP/capita, PPP$ inflation-adjusted)
- Child mortality (0-5 year-olds dying per 1,000 born)
- Life expectancy (years)
- Aid given (2007 US$)
- Aid given per person (2007 US$)

Geography

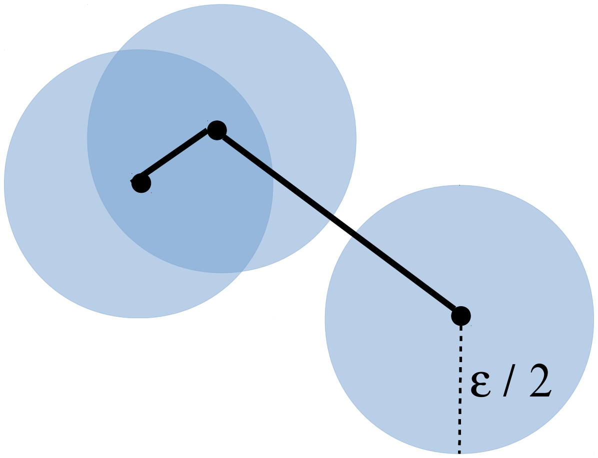
Point cloud of data



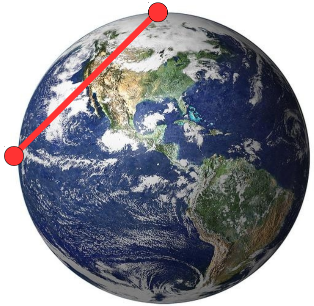
How do we impose structure?

Connecting the dots

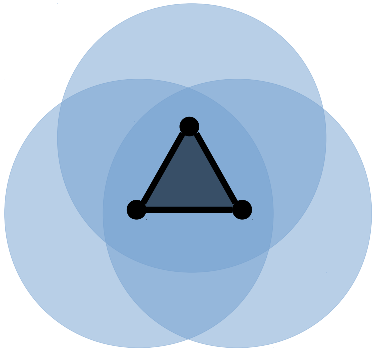
proximity parameter ε



Euclidean distance:
~~as the crow flies~~
as the mole burrows



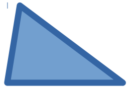
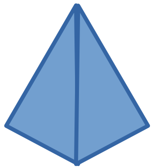
allow higher
dimensions



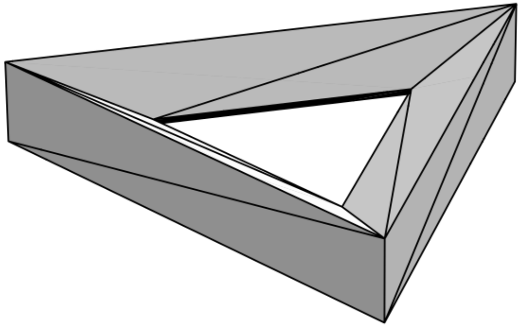
Simplicial Complexes

Oxymoron?

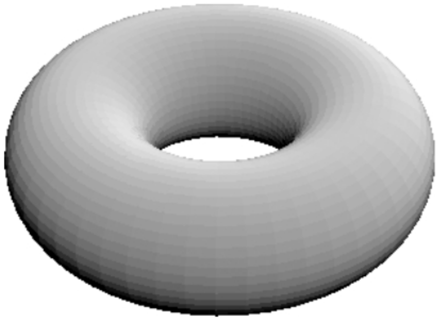
simplices

k=0 k=1 k=2  k=3  ...

simplicial complex:
if $\sigma < \Sigma$, and $\tau < \sigma$, then $\tau < \Sigma$



=



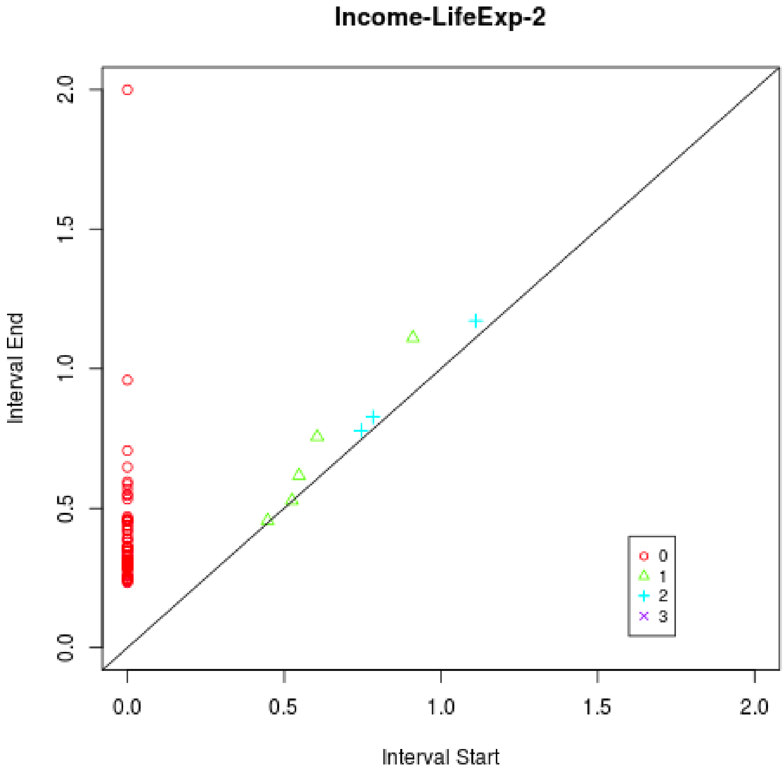
Persistent Homology

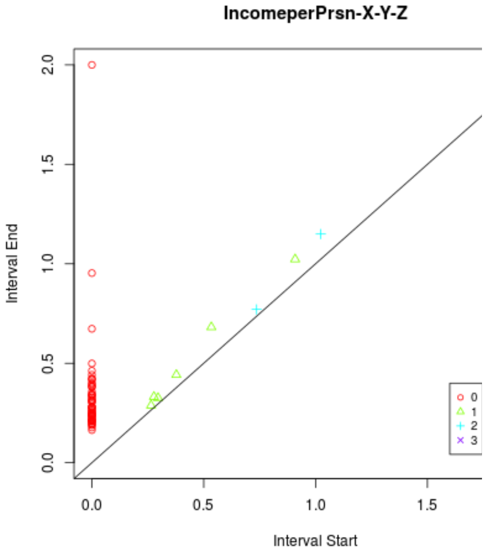
What is the right choice for ε ?

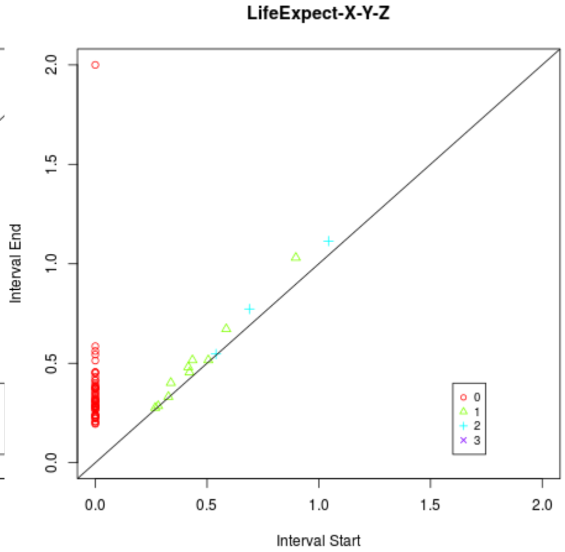
Barcodes

Torus example

Adding social dimensions







Income per person “pulls” the countries into two distinct geographic groups. Life expectancy is not strong enough to pull them back together.

What is topology?

- *“...a topologist is someone who cannot tell the*
- *difference between a tea cup and a doughnut.”* -Crossley

- Notions of equivalence
  - ex) x = y , = , =

- Study of continuous functions.
  - ex) continuous integer-valued function on the real line must be constant. What matters is the topology of **R** and **Z.**

Outline

- Gapminder does geography matter? Looks like it.
- Want to impose structure on data → analyze that structure instead (look for holes)
- Start w country centroids. To impose structure we are tempted to start connecting the dots, but how? By proximity
- Enter the simplicial complex → build up our math structure out of simple parts (simplicies)
- How do we analyze? We look for holes (of n-dimension), these holes tell us different things about the data → connected components, cycles, etc. (hard part)
- Calculate homology of simplicial complex. But which simplicial complex? Enter persistent homology.
- Look for holes that persist over a “significant” parameter range.

Homeomorphisms

Definition: Two topological spaces S and T are said to be **homeomorphic** if there are continuous maps $f : S \to T$ and $g : T \to S$ such that

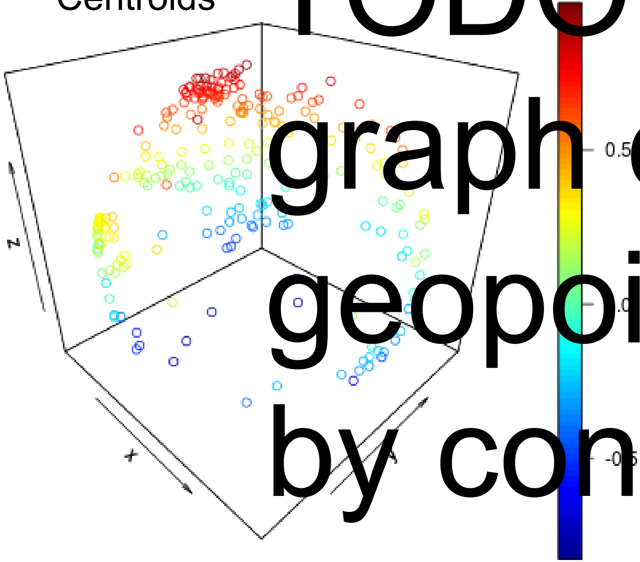
$$(f \circ g) = id_T \qquad \text{and} \qquad (g \circ f) = id_S,$$

then the maps f and g are **homeomorphisms**. The maps f and g are inverse to each other, so we may write f^{-1} in place of g and g^{-1} in place of f.

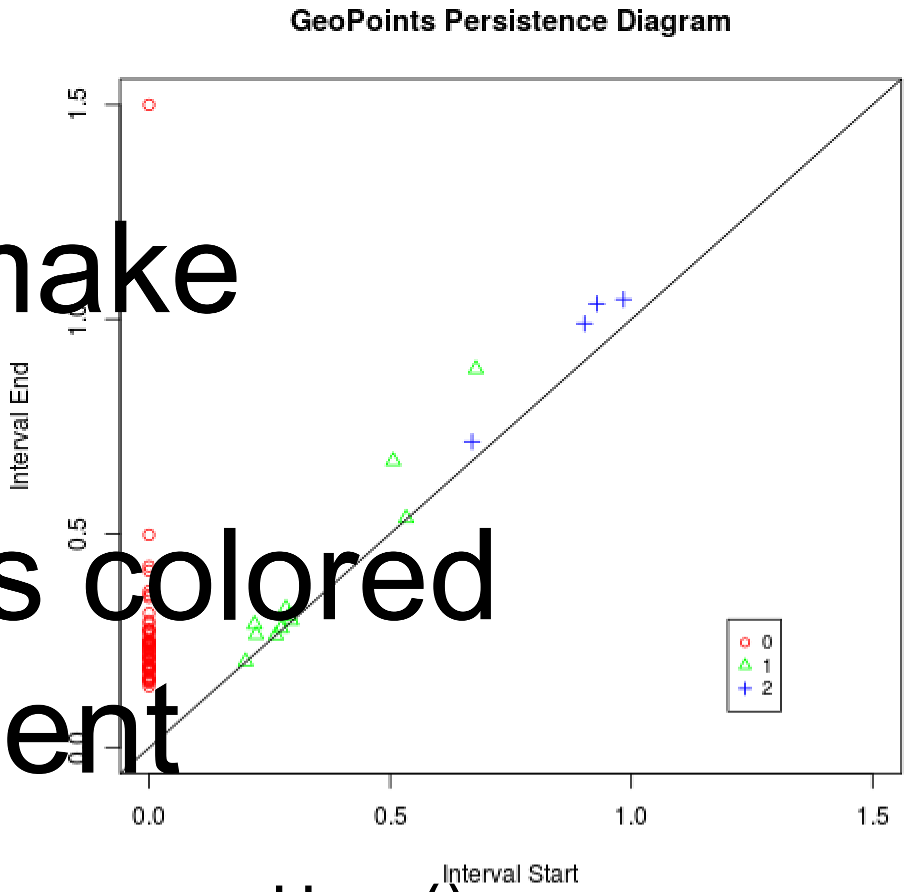
- TODO: Add coffee donut animation!!!

Geography

Country
Centroids



TODO: make graph of geopoints colored by continent



scatter3D() {plot3D}

pHom() {phom}

Limitations

- Coordinate space not theoretically justified
- Statistical significance (examine difference in means)
- Slow as Canadian molasses
- Ask a sociologist

Acknowledgment & References

- Gapminder
- http://www.statmethods.net/advstats/cluster.html (clustering)
- http://earthobservatory.nasa.gov/IOTD/view.php?id=885 (Earth Image)
- Wikipedia
- WolframMathWorld
- Ghrist
- Carlson
- Topology textbook (Crossley)

- Thank you...
- Lori Zeigelmeir & topology class
- MSCS
- y'all

Homotopy

- Two functions (loops or paths) are *homotopic* if there is a *continuous deformation* from one to the other.
- The **homotopy** is the function that "does the deforming."
- Group functions into homotopy *equivalence classes* → can count the number of “holes.”

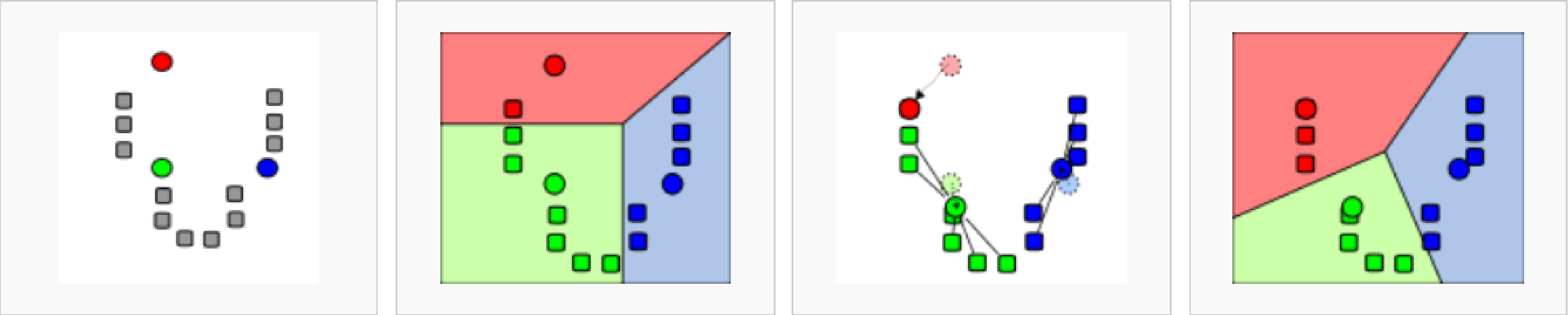
Understanding B_0 with clustering (cluster slides to be replaced w/ improved bettir 0 anlysis)

Can we use persistent homology as a clustering algorithm?

- Slow
- Sensitive to outliers
- Bridges collapse clusters
- Preprocessing algorithms required

In the meantime we'll use k-mean.

A k-means to an end

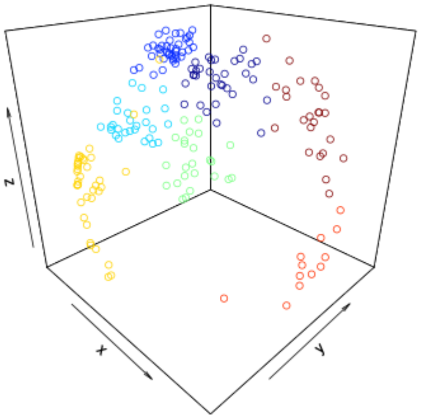


1. Initialize cluster centers.
2. Generate Voronoi Diagram for each center.
3. Let the centroid of each region be the new center.

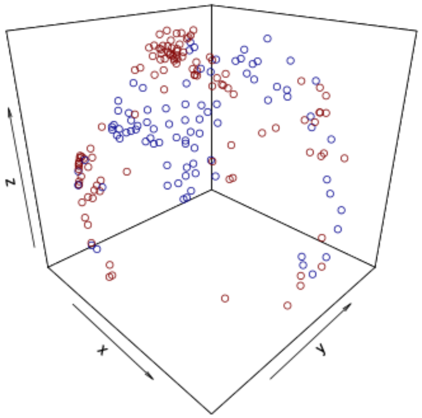
- Requires choice of *k*
- Fast
- Global solution NP-hard
- Heuristic Algorithm

Voronoi Diagram *The partitioning of a plane with points into convex polygons such that each polygon contains exactly one generating point and every point in a given polygon is closer to its generating point than to any other* -WolframMathWolrd

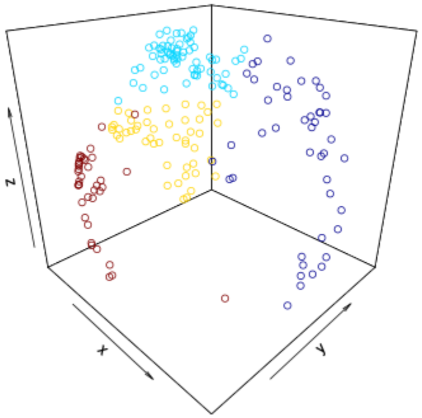
Geographic and social clusters



+



=



k = 7

Clusters based on geographic data only.

k = 2

Clusters based on Income per person and Life Expectancy

k = 4

Combined Geographic, Life expectancy, Income

0.45, -0.74 Europe/Eurasia
0.60. -0.58 Asia/South Pacific
-0.38, -0.94 Africa
0.53, -0.78 Americas

pamk() {fpc package}