Geographic Data Science

The need to group data

Carmen Cabrera-Arnau and Elisabetta Pietrostefani

The need to group data

Everything should be made as simple as possible, but not simpler

Albert Einstein

The need to group data

- The world is **complex** and **multidimensional**
- **Univariate** analysis focuses on **only one** dimension
- Sometimes, world issues are best understood as **multivariate**. E.g.
  - Percentage of foreign-born Vs. *What is a neighborhood?*
  - Years of schooling Vs. *Human development*
  - Monthly income Vs. *Deprivation*

Grouping as simplifying

- Define a given number of categories based on **many characteristics** (multi-dimensional)
- Find the **category** where each observation *fits best*
- **Reduce complexity**, keep all the **relevant information**
- Produce easier-to-understand outputs

Types of grouping

- Non-spatial clustering
- Regionalisation

Geographic Data Science

Non–spatial clustering

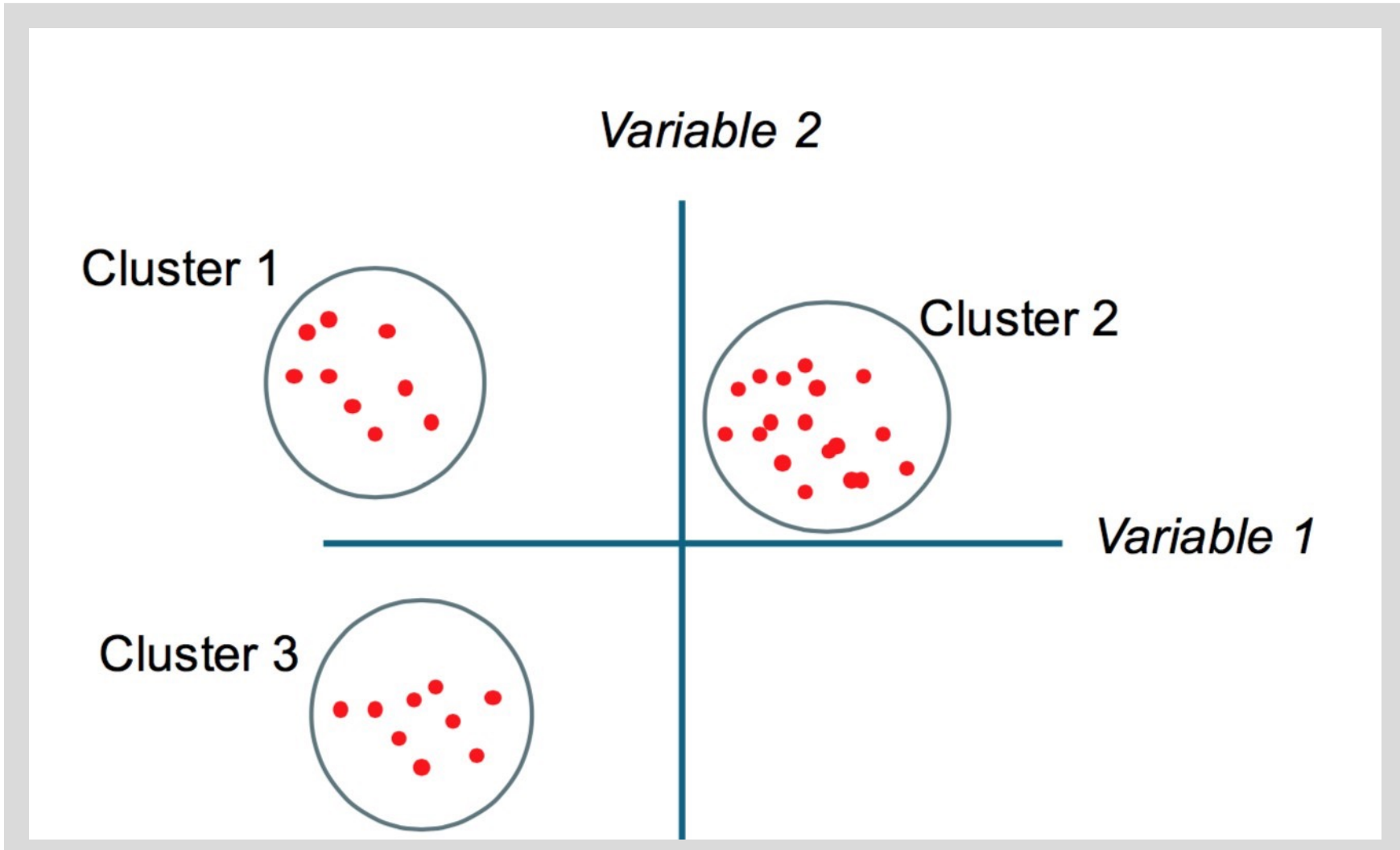
Non-spatial clustering

Split a dataset into **groups** of observations that are **similar within** the group and **dissimilar between** groups, based on a series of **attributes**

Machine learning

Unsupervised

Intuition



K-means

- Most popular clustering algorithm
- Good, but not perfect

More clustering…

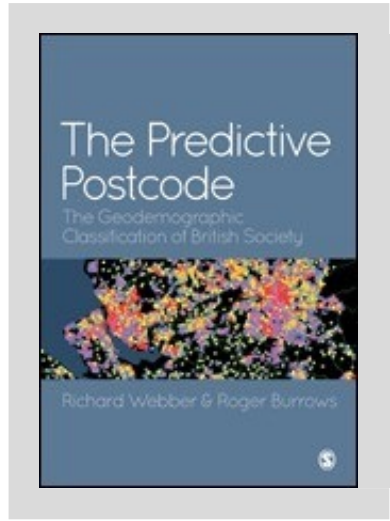
- Hierarchical clustering
- Agglomerative clustering
- Spectral clustering
- Neural networks (e.g. Self-Organizing Maps)
- DBSCAN
- …

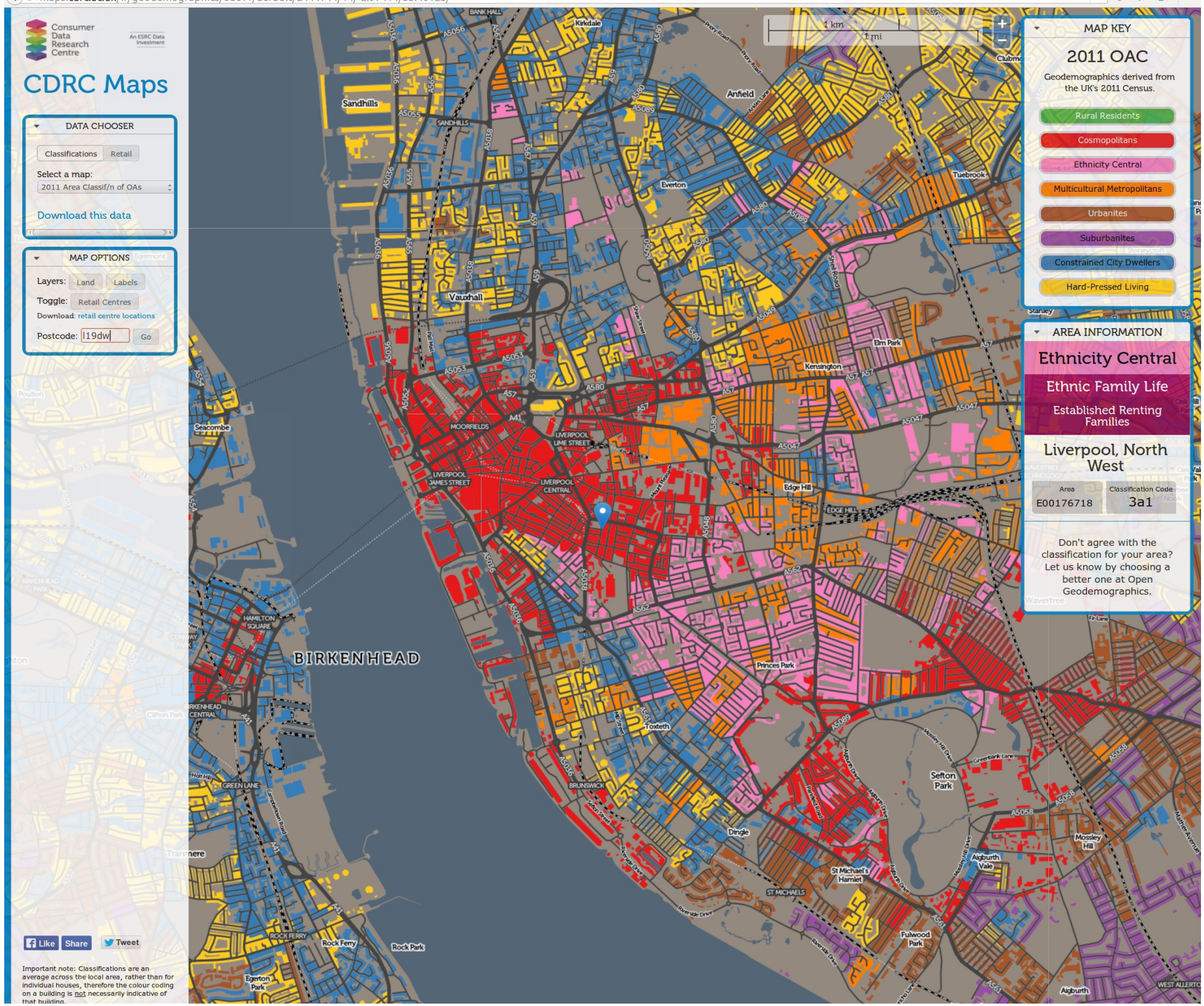
Different properties, different best usecases

Geodemographic analysis

Geodemographic analysis

- 1970's, Richard Webber
- **Identify similar neighborhoods**
  → Target urban deprivation funding
- **Public** Sector (policy) → **Private** sector (marketing and business intelligence)





[Internet user classification - CDRC]

Regionalization

Unsupervised **Spatial** Machine Learning

Aggregating basic spatial units (areas) into larger units (regions)

Regionalization

Split a dataset into **groups** of observations that are **similar within** the group and **dissimilar between** groups, based on a series of **attributes**…

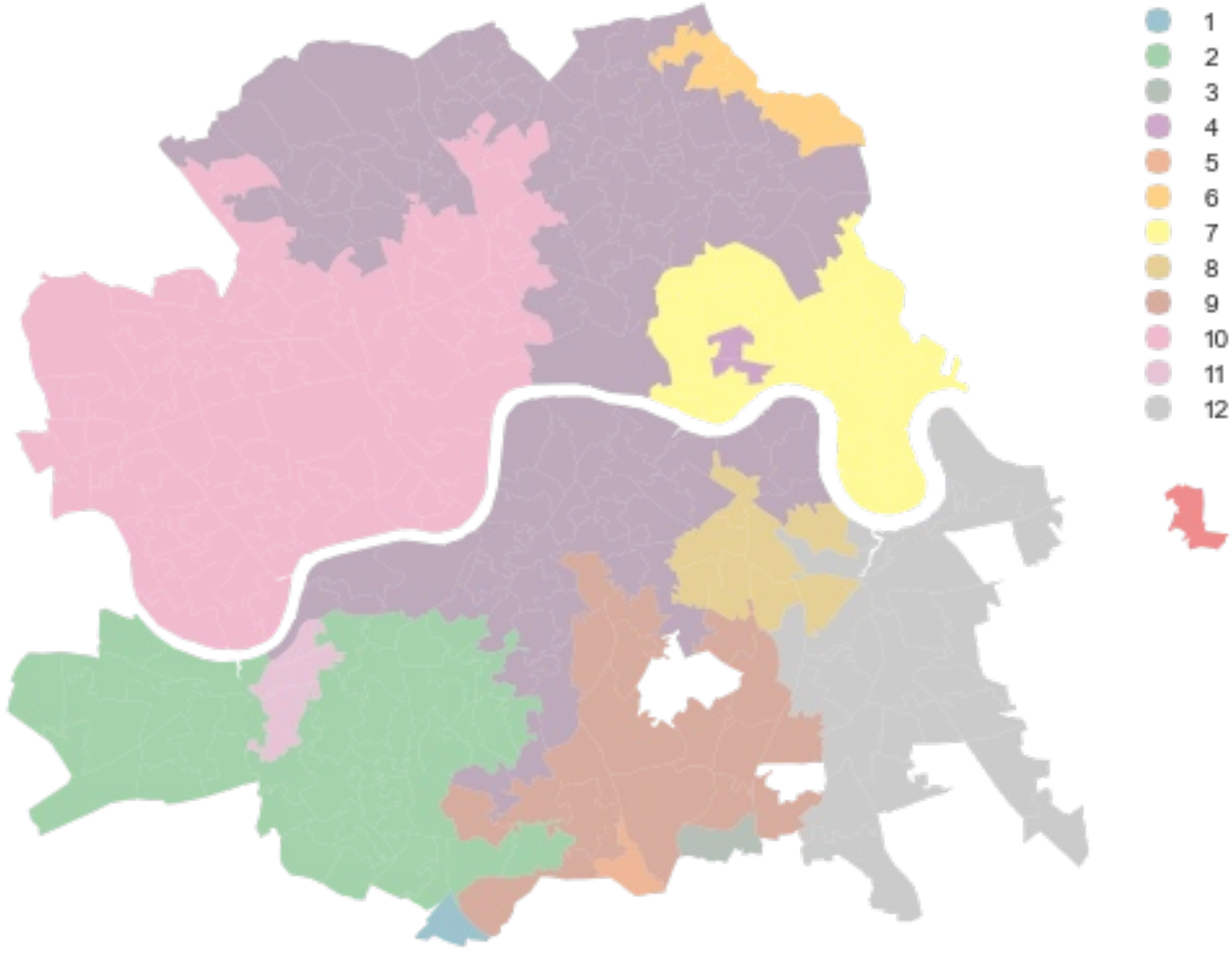
…with the additional constraint observations need to be **spatial neighbors**

Regionalization

Regionalization

- All the methods aggregate geographical areas into a predefined number of regions, while optimizing a particular aggregation criterion;
- The areas within a region must be geographically connected (the spatial contiguity constraint);
- The number of regions must be smaller than or equal to the number of areas; Each area must be assigned to one and only one region;
- Each region must contain at least one area.

Duque et al. (2007)


-1
0
1
2
3
4
5
6
7
8
9
10
11
12

Algorithms

- Automated Zoning Procedure (AZP)
- Max-P
- …

See Duque et al. (2007) for an excellent, though advanced, overview

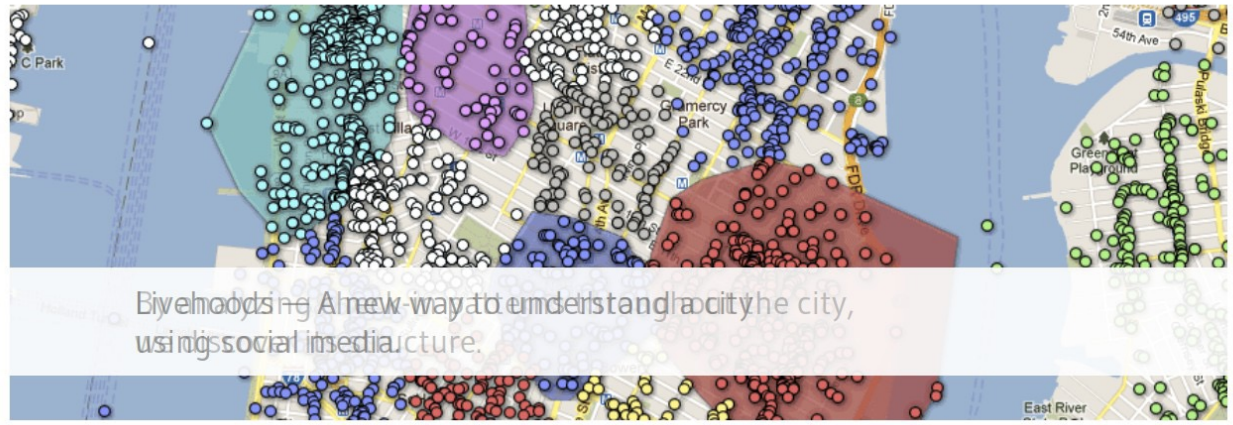
Examples

Census geographies

Environment and Planning A 1995, volume 27, pages 425–446

Algorithms for reengineering 1991 Census geography

S Openshaw, L Rao¶
School of Geography, University of Leeds, Leeds LS2 9JT, England
Received 22 April 1994; in revised form 6 October 1994



Re-Imagining the City in the Age of Social Media

Livehoods offer a new way to conceptualize the dynamics, structure, and character of a city by analyzing the social media its residents generate. By looking at people's checkin patterns at places across the city, we create a mapping of the different dynamic areas that comprise it. Each Livehood tells a different story of the people and places that shape it.

> MORE

Using Machine-Learning to Study Cities

Our research hypothesis is that the character of an urban area is defined not just by the the types of places found there, but also by the people that make it part of their daily life. To explore this idea, we use data from approximately 18 million check-ins collected from the location-based social network foursquare, and apply clustering algorithms to discover the different areas of the city.

> MORE

Current Maps

> New York City

> San Francisco

> Pittsburgh

> More Maps

News and Press

Livehood at ICWSM
Our work with Livehoods won the best paper award at ICWSM in Dublin this June! **Watch the video from our presentation.**

Livehoods on CBC Radio
Justin was on the CBC Radio program Spark talking with host Nora Young about the Livehoods Project. **Listen to the full interview.**

Livehoods in the Atlantic
Livehoods appeared as the Map of the Day on the Atlantic's Cities blog. **See their post about us.**

Wired Insider
Wired's Insider blog says Livehoods is "taking a big swing" at minining insights into "cultural habits and how societies flow." **Read the full post.**

> MORE

Recent Tweets

@tiffehr
Best map/location mashup I've seen in quite some time: **http://livehoods.org/maps/nyc#** (Via **http://roomthily.tumblr.com**)

@Werner
Livehoods is a cool CMU research project to visualize cities through the use of social media (**@foursquare** in this case) **http://wv.ly/IJZ3We**

@tomcoates
The 'Related' tab on **http://livehoods.org** is the best. See which neighboring places people travel too. Algorithmic divination of commuting!

@brainpicker
Forget neighborhoods, it's about Livehoods — Carnegie Mellon maps the dynamic character of cities through social media **http://j.mp/HzmkoN**

@kellan
clearly i live on the wrong side of the bqe - **http://livehoods.org/maps/nyc**

> MORE

Subscribe to our newsletter

Find out more about Livehoods and get updates on future developments by subscribing to our mailing list.

EMAIL*

NAME

SUBSCRIBE