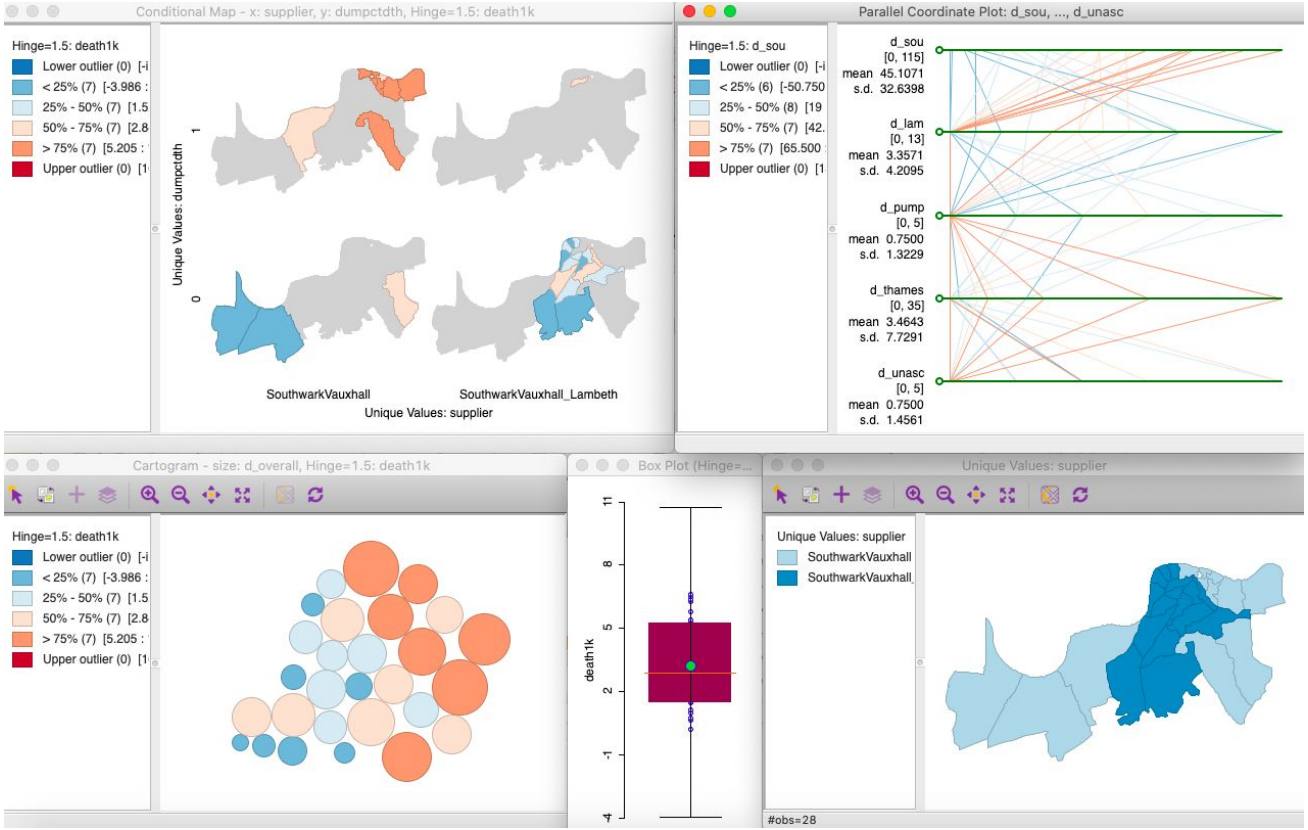
EDA and ESDA with GeoDa

John Snow & the 19th Century Cholera Epidemic

Julia Koschinsky
Marcos Falcone
spatial@uchicago.edu

September 2020

THE UNIVERSITY OF CHICAGO | THE CENTER FOR SPATIAL DATA SCIENCE



Resource Links

Download Data + Documentation

- https://geodacenter.github.io/data-and-lab//snow/

Download GeoDa

- https://geodacenter.github.io/

See GeoDa Snow Scripts in Context

- Storymap: https://bit.ly/3mSGZiS
- Video: https://bit.ly/365giRY

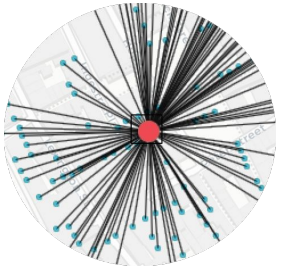
Examples and Spatial Data Files for Use in GeoDa

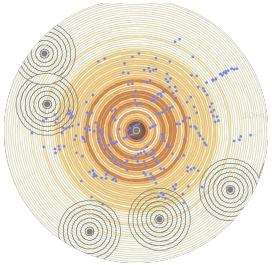
Broad St Pump

578 individual cholera deaths
Dataset 1

Cholera deaths in 40 housing blocks
Dataset 3

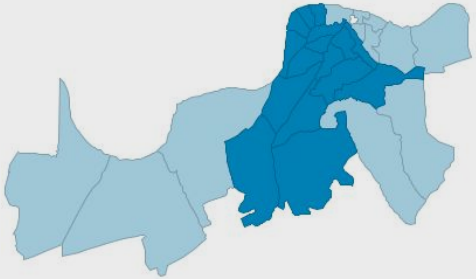
250 cholera deaths by building
Datasets 2 + 4





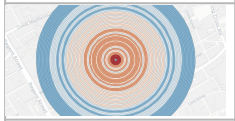
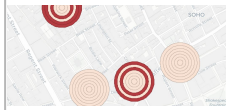
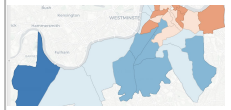
Cholera deaths around Broad St pump
Datasets 4, 5 + 6

S. London Experiment



Results for 28 subdistricts
Dataset 7

Overview of 7 Spatial Data Files: John Snow and the Cholera Epidemic

Screenshot	File # and name	Description	Case	Type	N	Var.	Contemporary Source	Original Source	License
	1. deaths	Individual deaths	Broad St Pump	Point	578	4	Tobler 1994, Arribas-Bel et al. 2017	Snow 1855 (Map 1)	GPL
	2. deaths_by_bldg	Deaths aggregated to buildings	Broad St Pump	Point	250	8	Wilson 2011, Arribas-Bel et al. 2017	Snow 1855 (Map 1)	Unknown
	3. deaths_by_block	Deaths aggregated to blocks	Broad St Pump	Polygon	40	3	Wilson 2011, Arribas-Bel et al. 2017. **Added workhouse by CSDS**	Snow 1855 (Map 1)	Unknown
	4. pumps	6 pumps in the Broad St area	Broad St Pump	Point	6	4	Wilson 2011, Arribas-Bel et al. 2017	Snow 1855 (Map 1)	Unknown
	5. deaths_by_bsrings	Deaths aggregated to 5m rings around Broad St pump	Broad St Pump	Polygon	60	6	Tobler 1994, Wilson 2011, Arribas-Bel et al. 2017. **Rings + calculations by CSDS**	Snow 1855 (Map 1)	GPL
	6. deaths_by_otherrings	Deaths aggregated to 10m rings around other pumps	Broad St Pump	Polygon	35	6	Tobler 1994, Wilson 2011, Arribas-Bel et al. 2017. **Rings + calculations by CSDS**	Snow 1855 (Map 1)	GPL
	7. subdistricts	London subdistricts as of 1855 with data	South London Natural Experiment	Polygon	28	28	Data by Coleman 2019. Original boundaries by Koch and Denike 2006 (no data). **Modified boundaries by CSDS.**	Snow 1855 (Map 2)	BSD 2

Overview of GeoDa Scripts: **Broad St Pump** & **South London Natural Experiment**

MORE CHOLERA DEATHS NEAR **BROAD STREET PUMP**

Identifying Clusters and Spatial Concentrations:

Connect deaths with nearby pumps:
Exploring the Relationship Between Two Point Layers

Explore deaths near the closest pumps:
K-Means Clustering and Heat Maps

View concentrations of deaths near Broad St pump:
Identifying Distance Decay

Find hotspots near the pump -- with a spatial outlier:
Local Moral Cluster Mapping

Comparing Distributions Across Groups:

Compare deaths near & far from pump:
Conditional Box Plots

SOUTH LONDON NATURAL EXPERIMENT: MORE DEATHS FOR SOME WATER SUPPLIERS

Comparing Trends:

Compare trends of deaths by water supply area:
Using the Time Editor and the Averages Chart

Exploring a Question with Multiple EDA and ESDA Tools:

Explore deaths, causes and water suppliers:
Scatter Plots, Box Plots, Parallel Coordinate Plots, Conditional Box Plots/Maps, Maps, and Cartograms

THE BROAD ST PUMP CASE

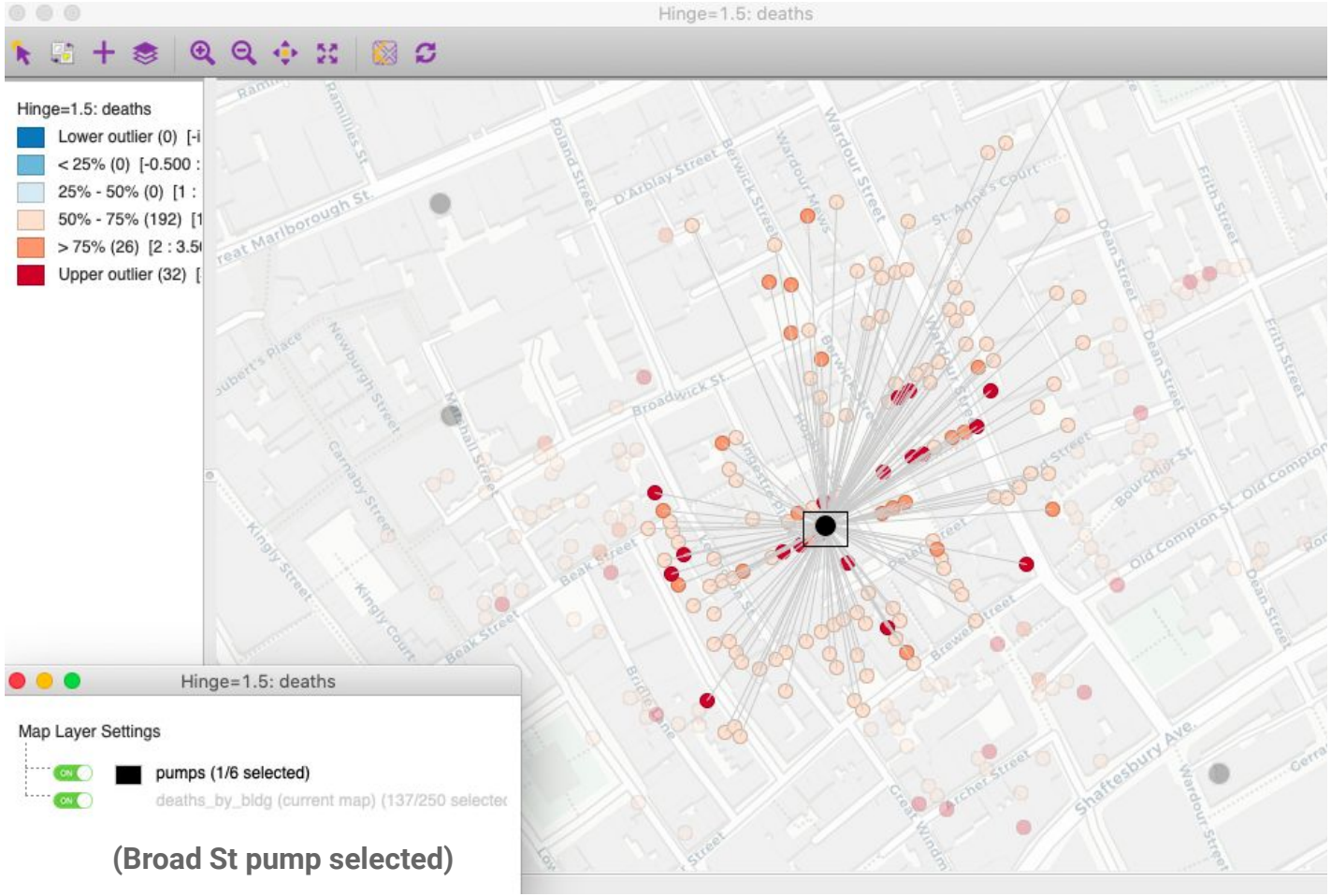
IDENTIFYING CLUSTERS AND SPATIAL CONCENTRATIONS

STEP-BY-STEP EXAMPLE 1: EXPLORING THE RELATIONSHIP BETWEEN TWO POINT LAYERS

Identifying clusters and spatial concentrations:
Connect cholera deaths with nearby pumps

Resource Links

Select a pump to see which cholera deaths are closest to that pump



(Broad St pump selected)

GeoDa Implementation

DATA - 2 shapefiles (shp, shx, dbf):

- deaths_by_bldg
- pumps

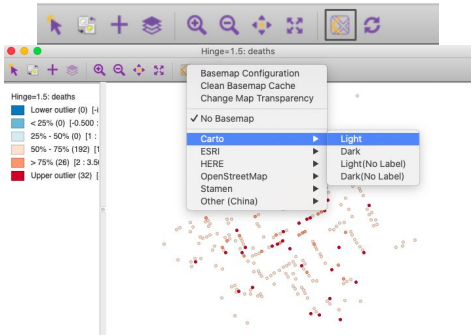
VARIABLES

- deaths_by_bldg: **deaths**
- deaths_by_bldg: **pumpID**

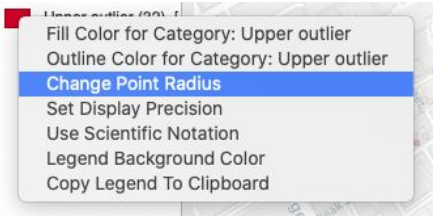
STEPS

1. **Map-Box Map** (deaths)
2. **Add basemap** (Carto Light)
3. Change point radius to 5 (right-click on legend, e.g. on red box)
4. **Add layer to boxmap**: + pumps and move to top
   then right-click pumps:
   a. Change fill color of pumps to black
   b. Change point radius to 8
   c. Set Highlight Association for pumps to link ID of 6 pumps to pumpID of cholera deaths (deaths, pumpID, ID)
5. Linking and brushing: select pump(s)
6. Close map

2. Add basemap

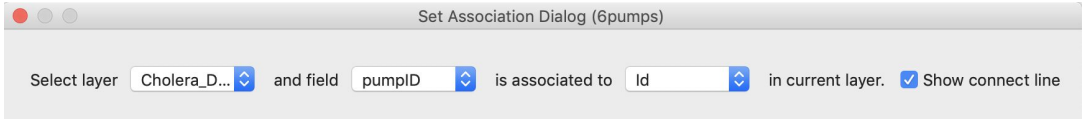


3. Change point radius



4. Change settings

4c. Set highlight association

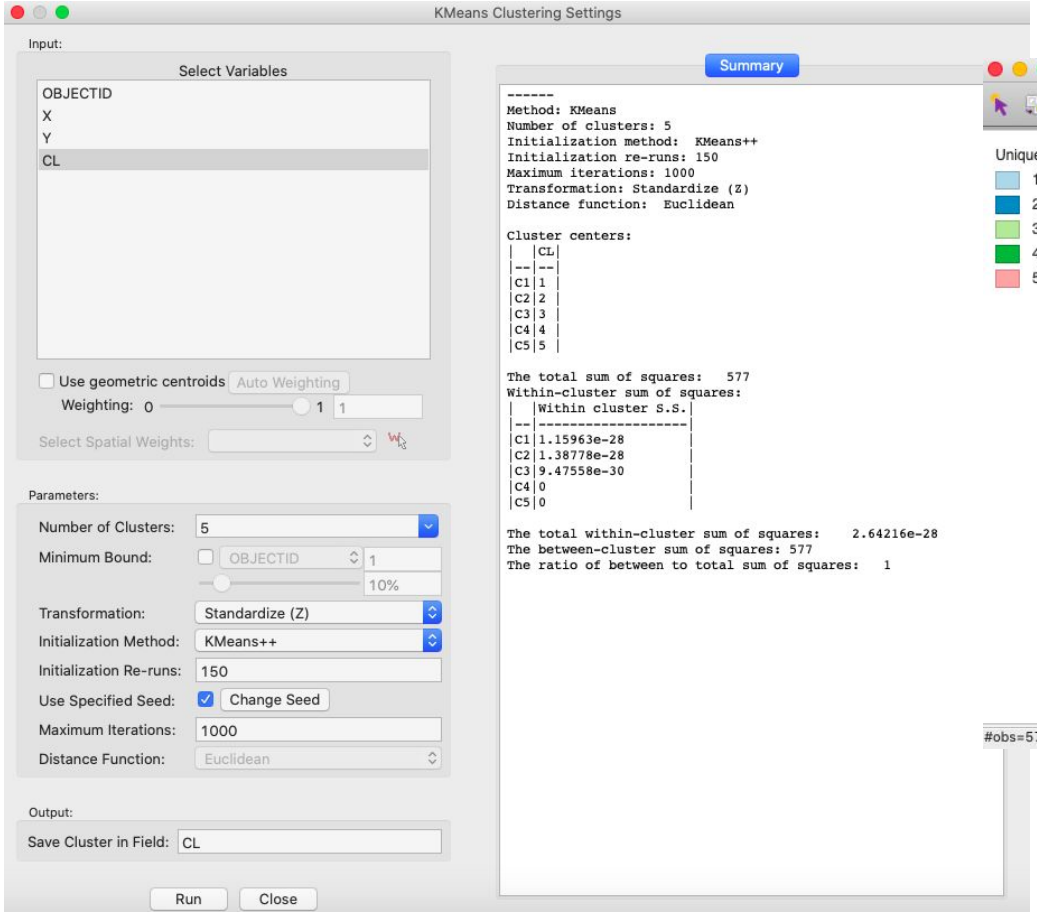


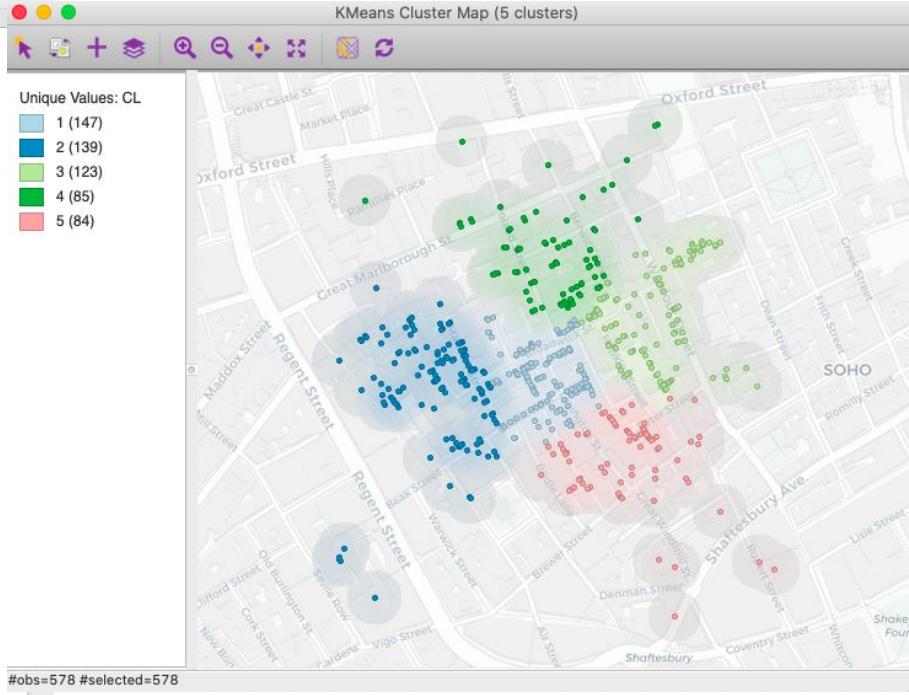
STEP-BY-STEP EXAMPLE 2: K-MEANS CLUSTERING AND HEAT MAPS

Identifying clusters and spatial concentrations:
Explore deaths near the closest pumps

Resource Links

Cluster deaths by proximity to nearest pump (K-Means Clustering)





Run Heat Maps on Clusters

GeoDa Implementation

DATA - 1 shapefile (shp, shx, dbf):

- deaths

VARIABLE

- CL

STEPS

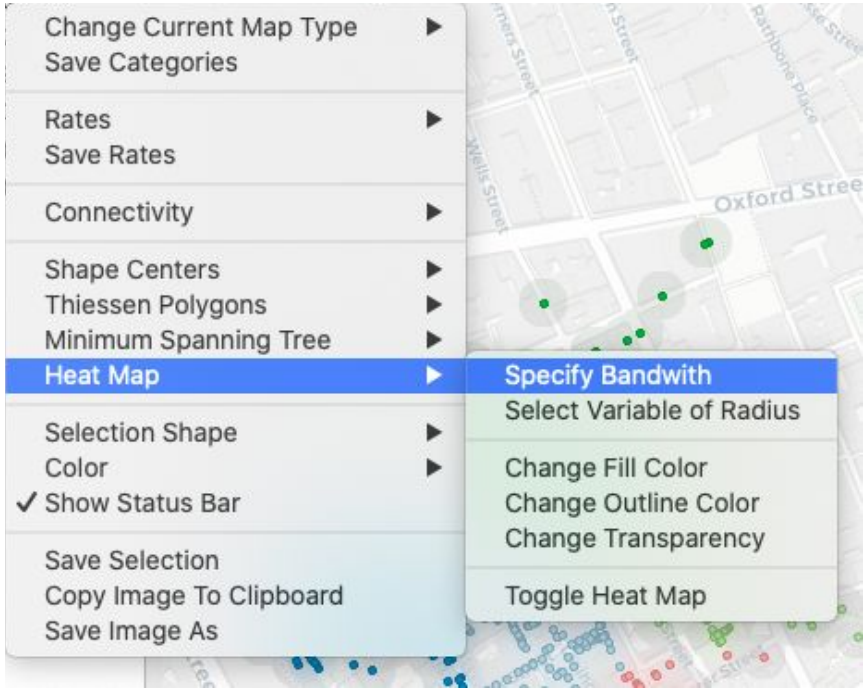
Run a K-Means Clustering Analysis

1. **Clusters-K Means**
2. **Select** "CL" as variable
3. **Set** the number of clusters as 5
4. **Save** Cluster in Field "CL"

Create a Heat Map

5. **Right click** on resulting map
6. **Heat Map-Specify Bandwidth**
7. **Select** desired bandwidth

6 - Heat Map - Specify Bandwidth

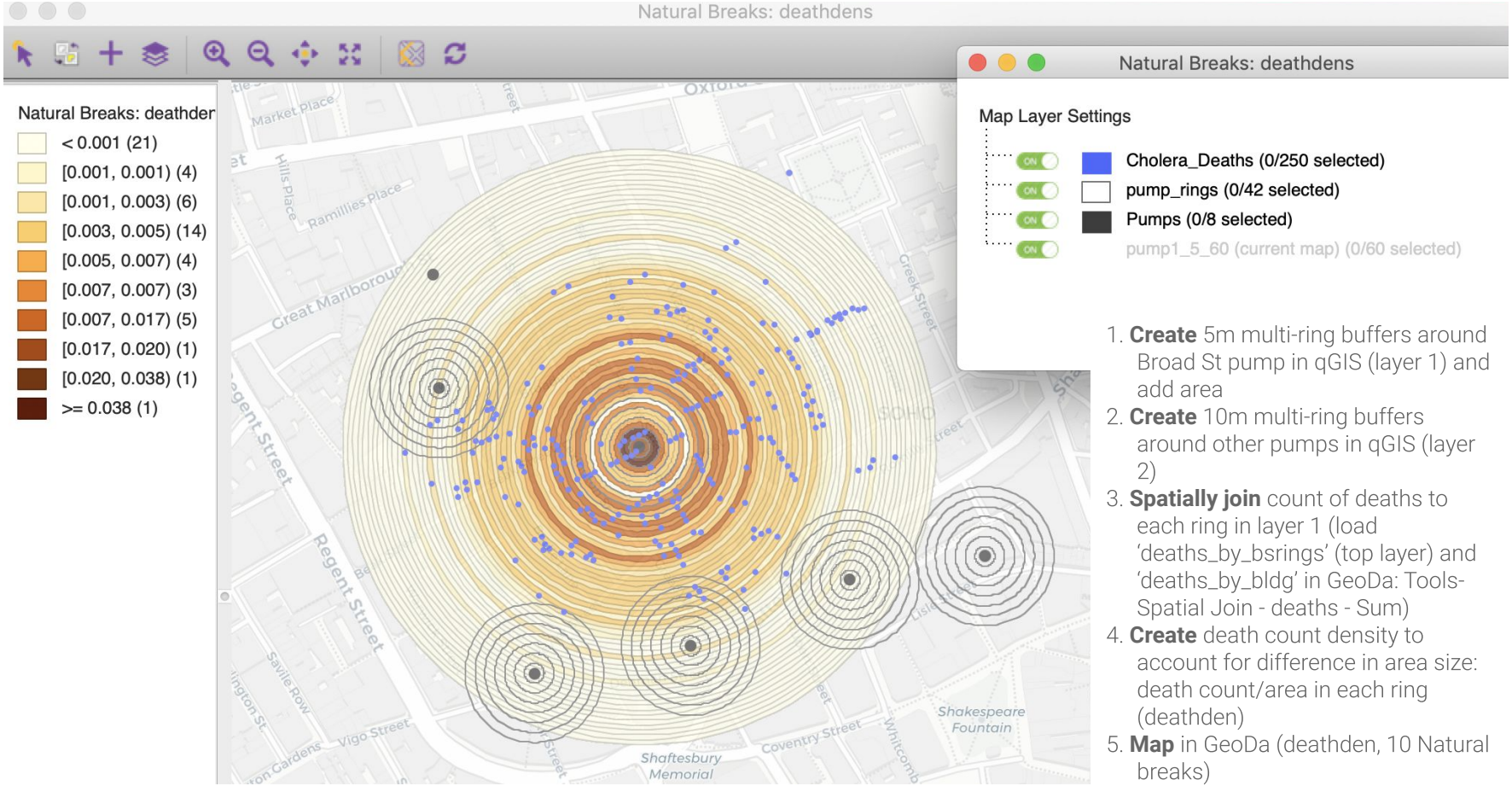


STEP-BY-STEP EXAMPLE 3: IDENTIFYING DISTANCE DECAY

Identifying clusters and spatial concentrations:
View concentrations of deaths near Broad St pump

Resource Links

More Deaths Near Broad St Pump: Distance Decay Demonstration



1. **Create** 5m multi-ring buffers around Broad St pump in qGIS (layer 1) and add area
2. **Create** 10m multi-ring buffers around other pumps in qGIS (layer 2)
3. **Spatially join** count of deaths to each ring in layer 1 (load 'deaths_by_bsrings' (top layer) and 'deaths_by_bldg' in GeoDa: Tools-Spatial Join - deaths - Sum)
4. **Create** death count density to account for difference in area size: death count/area in each ring (deathden)
5. **Map** in GeoDa (deathden, 10 Natural breaks)

GeoDa Implementation

DATA - 2 shapefiles (shp, shx, dbf):
- deaths_by_bldg
- deaths_by_bsrings

VARIABLES
- deaths_by_bldg: deaths
- deaths_by_bsrings: area

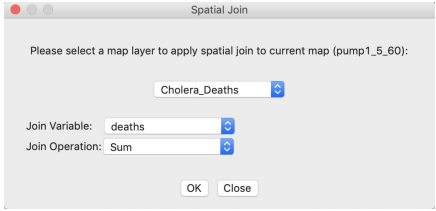
STEPS

Spatially join count of deaths to each ring around Broad St pump:
1. **Load** deaths_by_bsrings first (base layer to join points to)
2. **Load** deaths_by_bldg (move to top to see points)
3. **Tools-Spatial Join** (**Map Layer** = deaths, **Join Variable** = deaths, **Join Operation** = Sum)
4. **Add** new field to deaths_by_rings: deaths
5. **Table-Edit Variable Properties**: Real to integer
6. **Save** (this adds counts of deaths by ring to BroadStPump5mRings)

Calculate death density:
7. **Table-Calculator**
8. **Bivariate-Add Variable**: deathden → deaths DIVIDE area (decimals: 6, display 6)
9. **Save** (this adds deaths/area to table)

Map deathden:
1. Right-click on map- **Change Current Map Type** - Natural Breaks: 10 (deathden)
2. Close project

3. Tools - Spatial Join



7. Table - Calculator

STEP-BY-STEP EXAMPLE 4: LOCAL MORAN CLUSTER MAP

Identifying clusters and spatial concentrations:
Find hotspots near the pump -- with a spatial outlier

Resource Links

GeoDa Implementation

DATA - 2 shapefiles (shp, shx, dbf):

- deaths_by_block
- pumps

VARIABLE

- deaths_by_block: **deaths**

STEPS

1. **Tools-Weights Manager-Create**
2. **Select ID** variable (ID)
3. **Distance Weight-Specify Bandwidth**: 150 meters.
4. **Space-Univariate Local Moran's I**
5. **Select variable** ("deaths"), then "Cluster Map"
6. **Add layer to boxmap**: pumps and move to top
   then right-click pumps:
   a. Change fill color of 6pumps to black
   b. Change point radius to 5
7. Close map

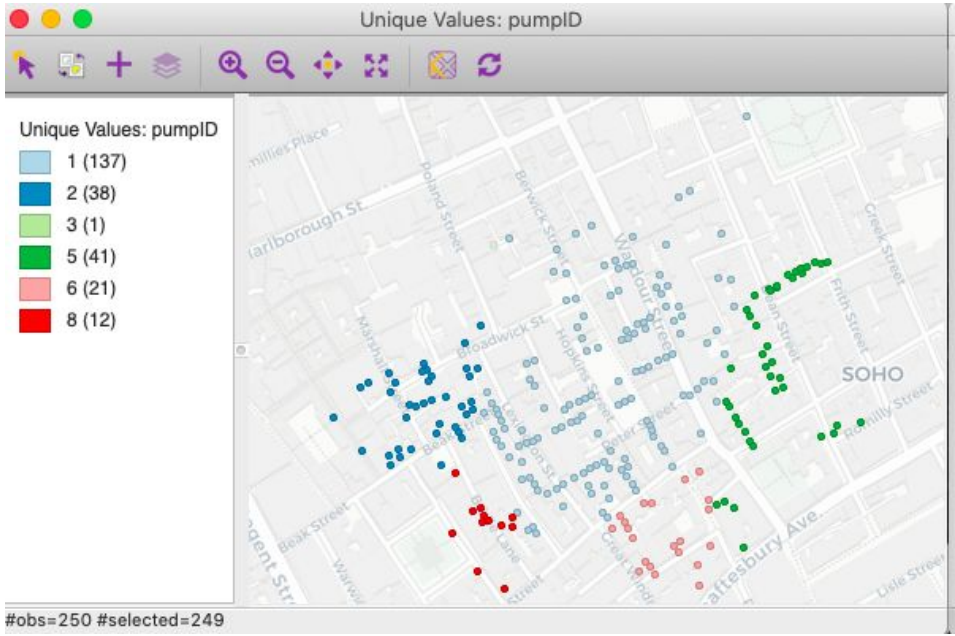
COMPARING DISTRIBUTIONS ACROSS GROUPS

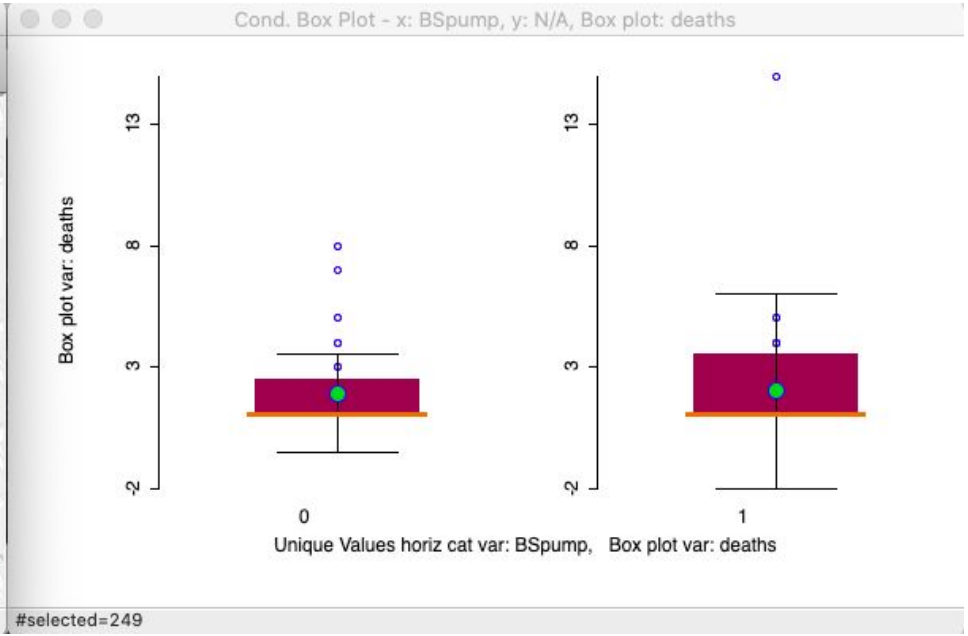
STEP-BY-STEP EXAMPLE 5: CONDITIONAL BOX PLOTS

Comparing distributions across groups:
Compare deaths near & further from pump

Resource Links

Closer Proximity to Broad St Pump Associated with More Cholera Deaths





Buildings with deaths, colored by which pump the building is closest to.
If Broad St pump is closest then BSpump = 1, all others = 0

closest pump = other

closest pump = Broad St

Conditional Boxplot: Number of deaths, broken out by whether Broad St pump is the closest pump or not.

Caveats: There is no information in this dataset whether individuals drank water from the Broad St pump or not. Also, people who did not die are not included.

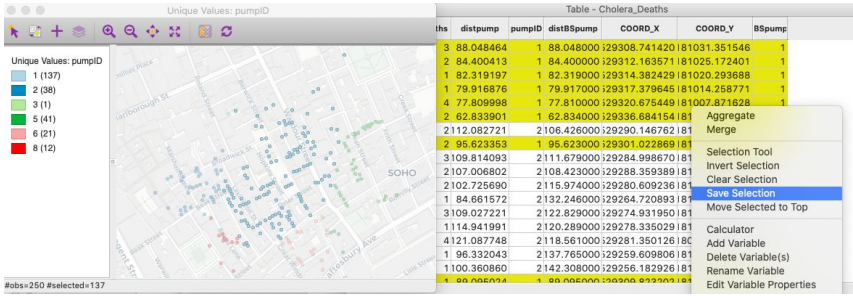
GeoDa Implementation

DATA - 1 shapefile (shp, shx, dbf):
- deaths_by_bldg

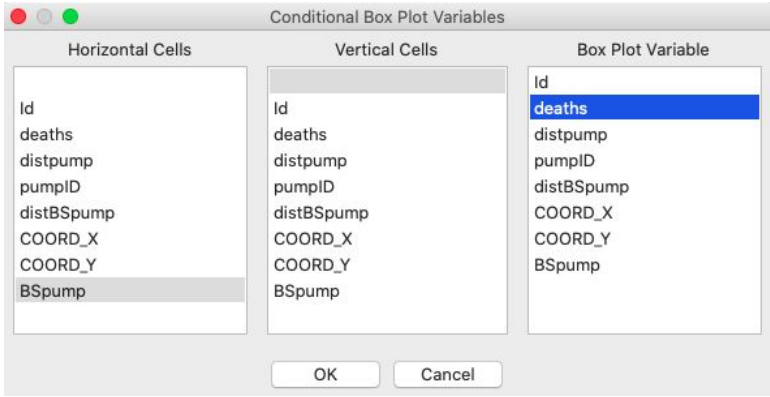
VARIABLES
- deaths
- pumpID

STEPS
1. **Map-Unique Values Map** - Select “pumpID”.
2. **Add Basemap** (Carto Light)
3. **Select category 1** in unique values map legend (pumpID = 1)
4. **Table** - **Save selection** as new variable (**BSpump**): buildings with deaths where Broad St pump is closest (1) or other pump is closest (0)
5. **Explore-Conditional boxplot** with horizontal = BSpump, vertical = blank, and map theme = deaths (1 row, 2 columns)
    a. Right-click: **Change horizontal bin breaks to unique values** for categorical representation of 0-1

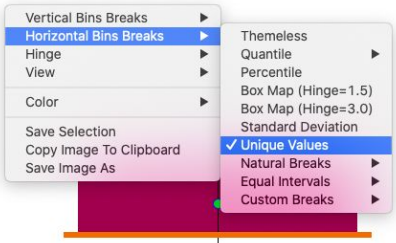
3. Select category 1
4. Right-click to save selection



5. Select variables



5.a. Modify horizontal bin breaks



THE SOUTH LONDON NATURAL EXPERIMENT

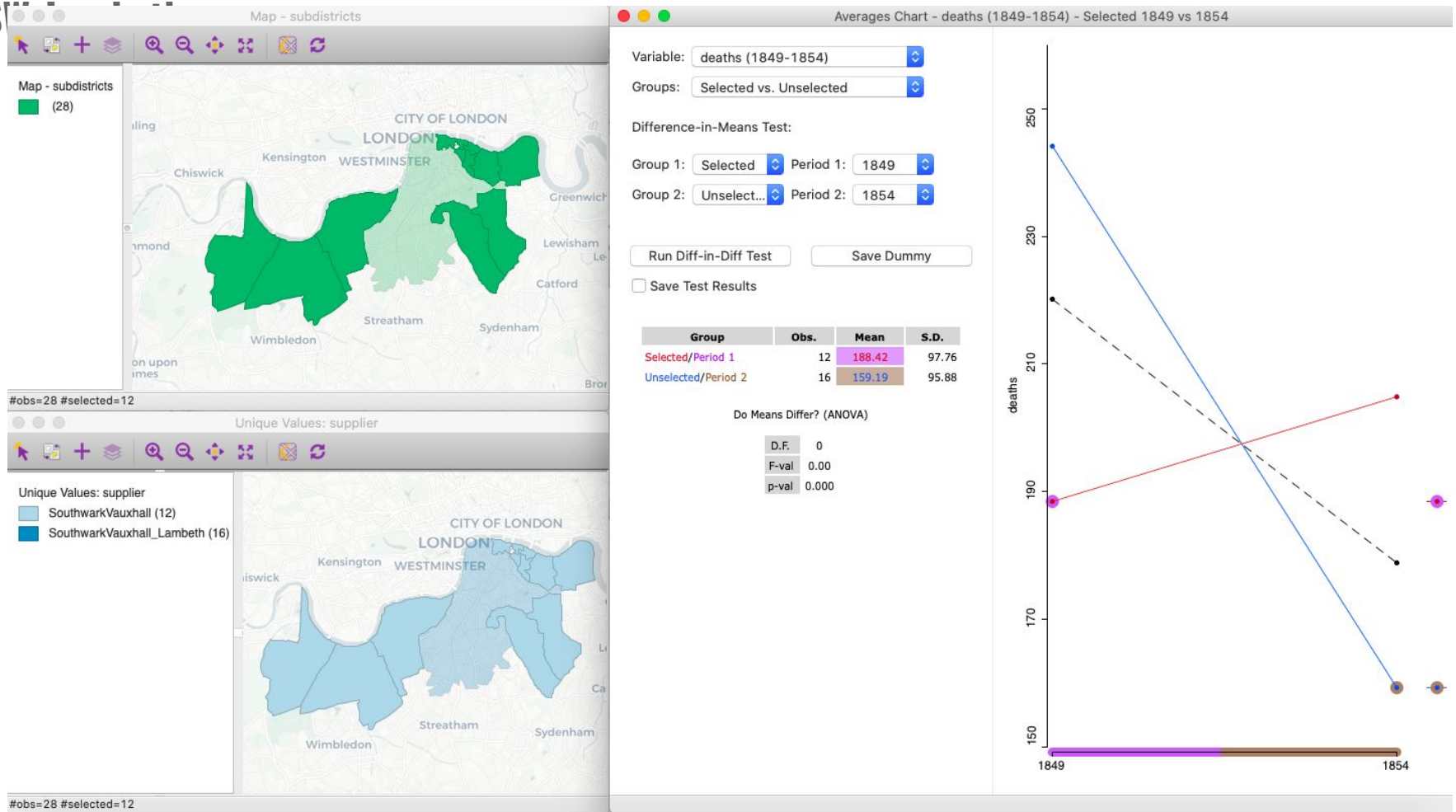
COMPARING TRENDS

STEP-BY-STEP EXAMPLE 6:
USING THE TIME EDITOR AND THE AVERAGES CHART

Comparing trends:
Compare trends of deaths by water supply area

Resource Links

SOUTH LONDON natura EXP.: SW Water Supplier Has Worse Cholera Death Trend Than S



GeoDa Implementation

DATA - 1 shapefile (shp, shx, dbf):
- subdistricts

VARIABLES
- deaths1849
- deaths1854

STEPS

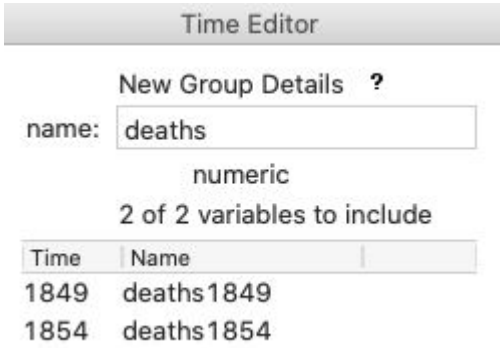
Creating a time variable:

1. **Time - Time Editor:** Select “deaths1849” and “deaths1854” and click on right arrow to move them from left to center
2. **Rename** new variable as “deaths”
3. **Double click** on “Time” and replace the two values with “1849” and “1854” respectively
4. Click on right arrow to group variables and move them from center to right

Comparing distributions across time and space:

5. **Explore-Averages Chart:** Select “deaths(1849-1854)” as variable, change Group 2-Period 2 to “1854”
6. **Map-Unique Values Map:** Select “supplier”
7. **Select** only “Southwark&Vauxhall” observations on the “supplier” unique values map.

1-3. Time Editor



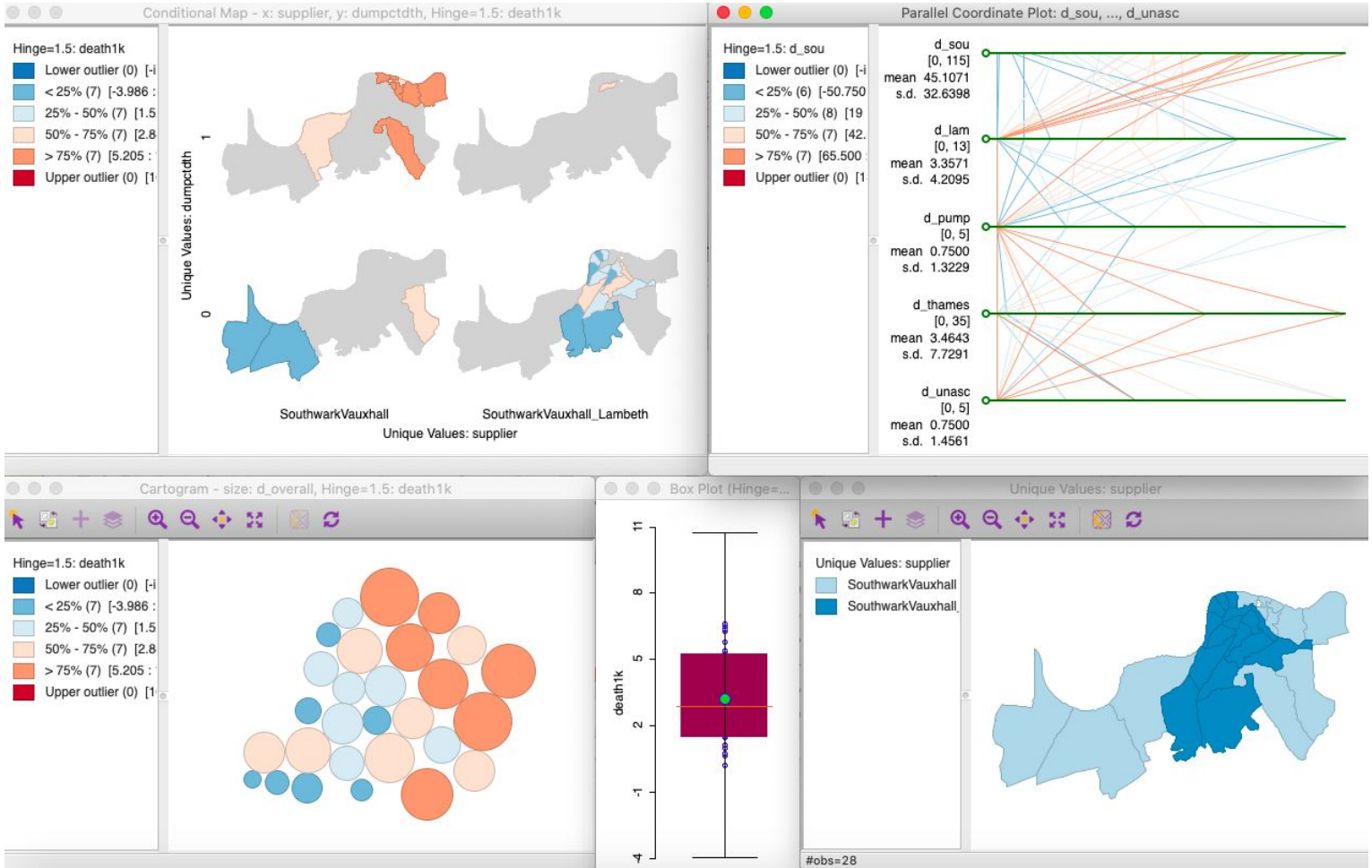
EXPLORING A QUESTION WITH MULTIPLE EDA + ESDA TOOLS

STEP-BY-STEP EXAMPLE 7:
SCATTER PLOTS, BOX PLOTS, PARALLEL COORDINATE PLOTS, CONDITIONAL BOX PLOTS/MAPS, MAPS, AND CARTOGRAMS

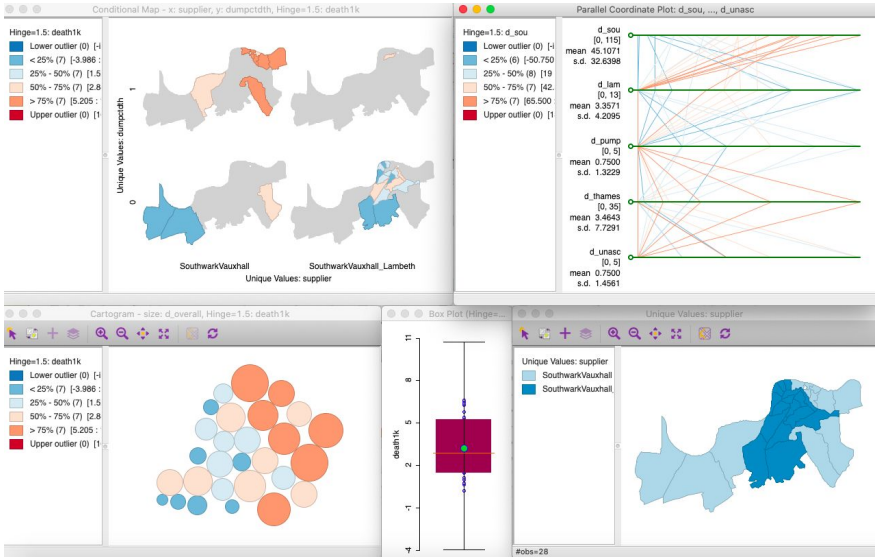
Exploring a question with multiple EDA and ESDA tools:
Explore deaths, causes and water suppliers

Resource Links

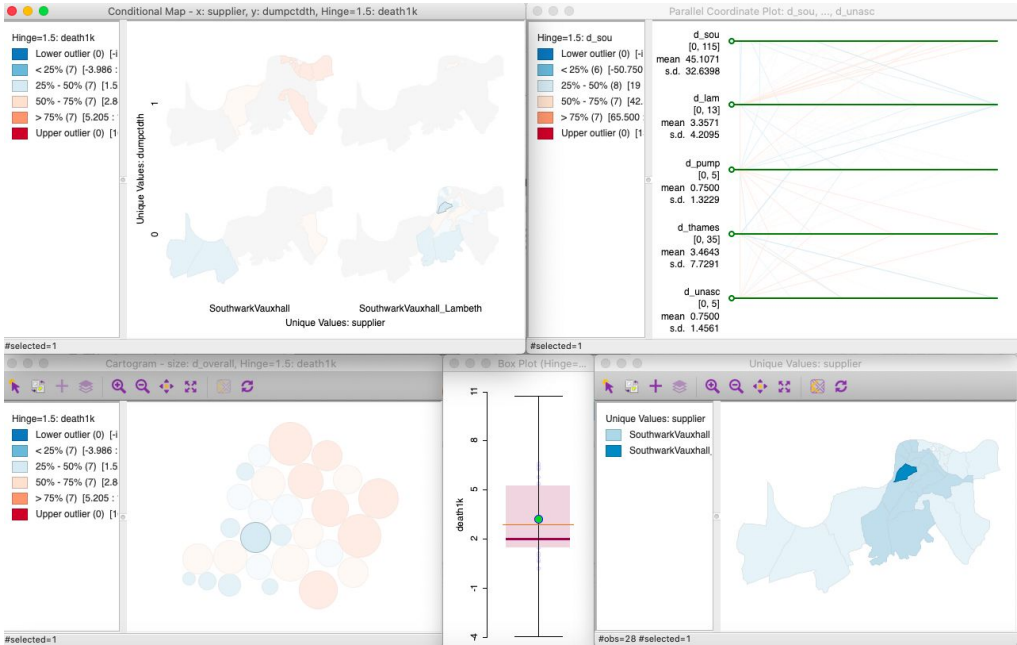
SOUTH LONDON EXP.: ESDA - Multiple Views of Deaths, Death Causes and Water Suppliers



SOUTH LONDON EXPERIMENT: Linking and Brushing to Drill Into Unusual Observations



Selecting one observation in one view will also select it in the other views



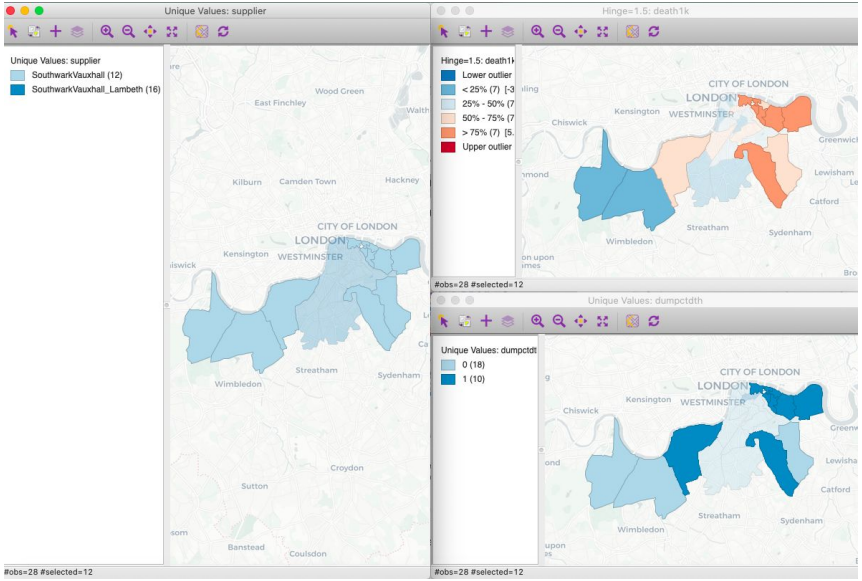
SOUTH LONDON EXPERIMENT

Subdistricts with Southwark&Vauxhall as Water Supplier Seem to Have Higher Share of Cholera Deaths

Maps of Conditional Boxplot Variables

Unique Values Map:
water **supplier**

Boxmap**: death1k**



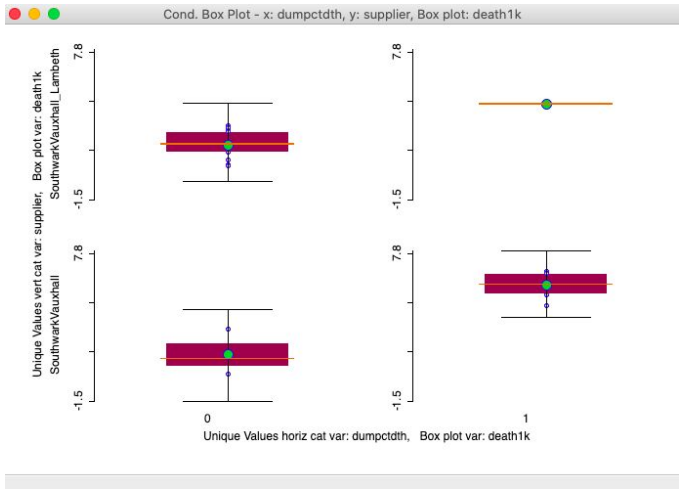
low-high death1k category
(**dumpctdth**: 0 = 0-3 deaths/1k, 1 = 4-14)

by water
supplier

Conditional Boxplot

%death broken out by supplier and low/high %death

death1k

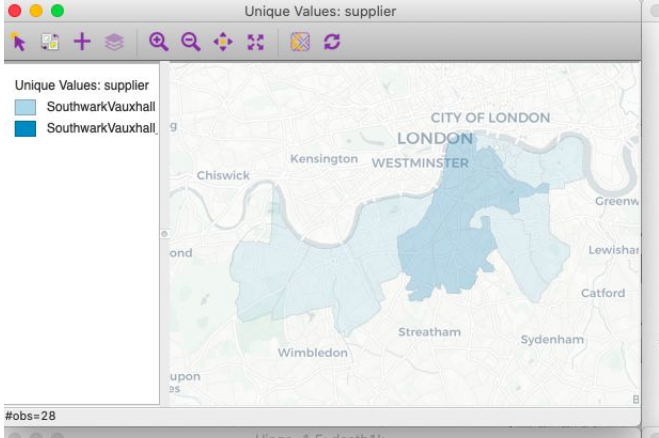


by low-high death1k category
(**dumpctdth**: 0 = 0-3 deaths/1k, 1 = 4-14)

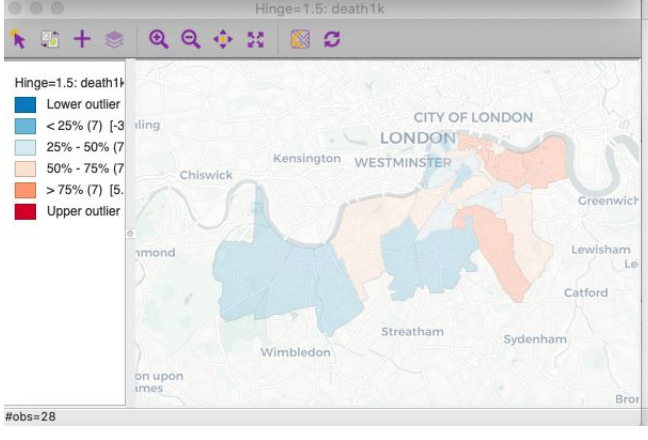
SOUTH LONDON EXPERIMENT

Higher Share of Deaths in Subdistricts Associated with Southwark Water Company

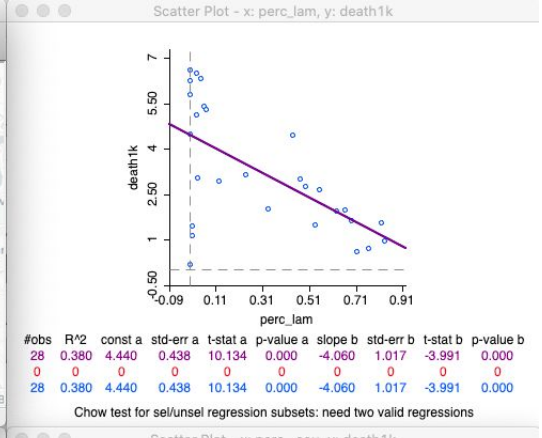
Unique value map supplier: Water supplier



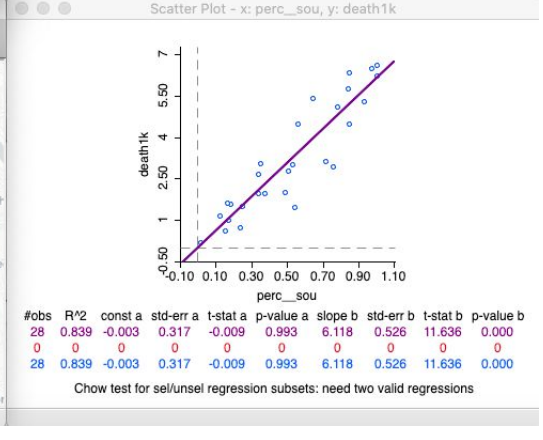
Boxmap death1k: Cholera deaths per 1000 people



Scatterplot | death1k: Cholera deaths per 1000 people



perc_lam: % population served by Lambeth company



perc_south: % population served by Southwark & Vauxhall company

GeoDa Implementation

DATA - 1 shapefile (shp, shx, dbf):

- subdistricts

VARIABLES

- **death1k** (deaths per 1,000 people; see below)
- **dumpctdth** (creates a 0-1 indicator variable for death1k: 0 is 0-3 deaths/1k people, 1 is 4-14 deaths per 1k people; see below)
- **supplier**

STEPS

Calculate death1k:

- **Table-Calculator-Bivariate-Add Variable**: death1k → 'd_overall' DIVIDE 'population' (decimals: 6, display 6)
- **Table-Calculator-Bivariate-Add Variable**: death1k → 'd_overall' MULTIPLY 1000 (decimals: 6, display 6)
- **Save** (this adds death1k to table)

Calculate dumpctdth:

- **Table- Sort** death1k highest to lowest
- **Select** observations equal to 4 or more: **Save Selection**
- **Write** 'dumpctdth' as variable name-Leave rest of the settings-**Save** (this adds dumpctdth to table)

1. **Map-Box Plot** (death1k), add Carto Dark basemap
2. **Map-Unique Values Map** (supplier), add Carto Dark basemap
3. **Map-Unique Values Map** (dumpctdth), add Carto Light basemap
4. **Explore-Conditional Box Plot** with horizontal = **dumpctdth**, vertical = **supplier**, and map theme = **death1k** (2 rows, 2 columns)
    a. Right-click: **Change horizontal bin breaks to unique values** for categorical representation of 0-1

SOUTH LONDON EXPERIMENT: Scatter Plots

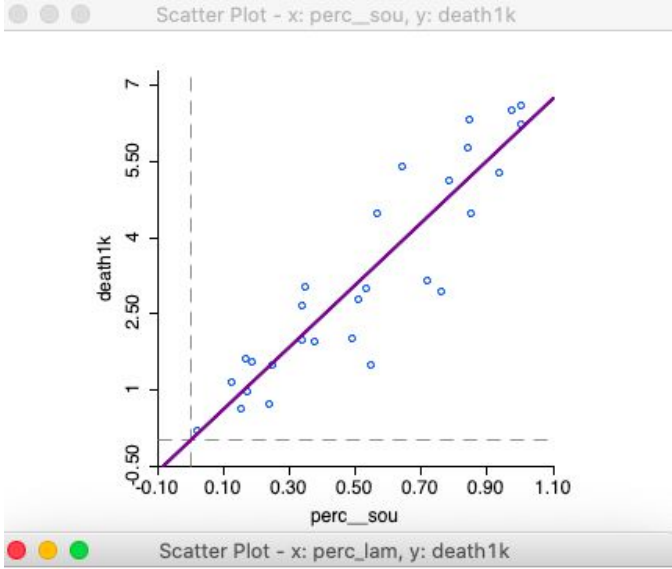
Close conditional boxplot and unique values map (dumpctdth)
Leave other two maps open (death1k and supplier)

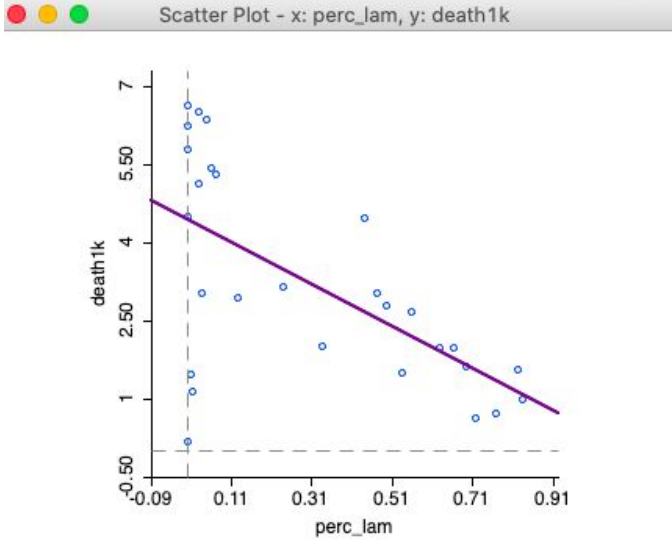
Variables:

- **death1k**
- **pct_lam:** % population served by Lambeth company
- **pct_south:** % population served by Southwark & Vauxhall company

Functionality:

1. Open scatterplot (**X: pct_lam**, **Y: death1k**)
2. Open scatterplot (**X: pct_south**, **Y: death1k**)





SOUTH LONDON EXPERIMENT: Parallel Coordinate Plot

DATA - 1 shapefile (shp, shx, dbf):

- subdistricts

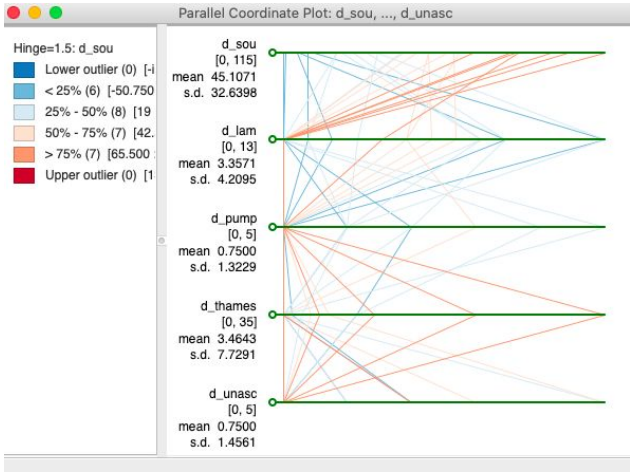
VARIABLES

Deaths attributed to ...

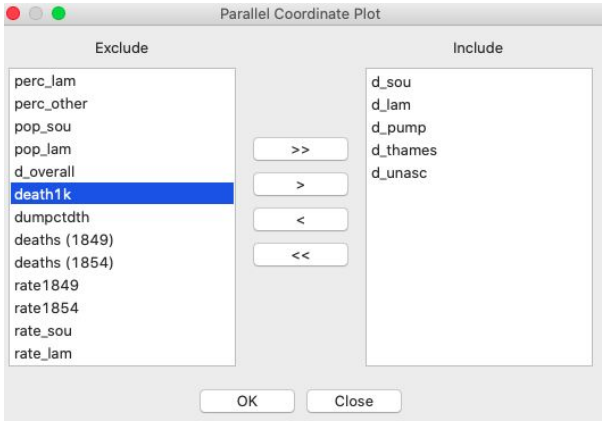
- **d_sou:** … the Southwark company
- **d_lam:** … the Lambeth company
- **d_pump:** … pumps or wells
- **d_thames:** … Thames water
- **d_unasc** … an unknown source

STEPS

1. **Parallel coordinate plot:** 
   a. **Double-click** on all 'd_x' variables: d_sou, d_lam, d_pump, d_thames, d_unasc
   b. **Right-click on plot: Classification Theme - Boxplot Theme - Hinge = 1.5**
   c. **Move axes** (by grabbing green circle at left start of axes) from top to bottom: **d_sou, d_unknw, d_pump, d_lam, d_thames**



1. Parallel coordinate plot variables



SOUTH LONDON EXPERIMENT: Conditional Map and Cartogram

DATA - 1 shapefile (shp, shx, dbf):

- subdistricts

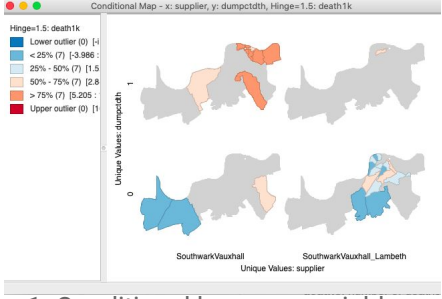
VARIABLES

- **death1k:** Cholera deaths per 1000 people
- **supplier:** Water supply companies
- **dumpctdth:** low-high death1k category (dummy variable): 0 = 0-3 deaths/1k, 1 = 4-14)
- **deaths**: number of deaths

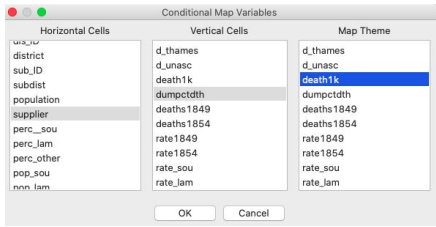
STEPS

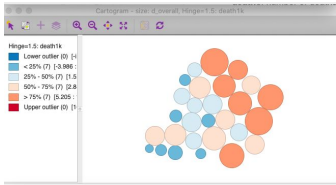
1. **Explore-Conditional Plot-Boxplot**  with horizontal = **supplier**, vertical = **dumpctdth**, and map theme = **death1k** (2 rows, 2 columns)
   a. **Right-click: Change vertical bin breaks to unique values** for categorical representation of 0-1

2. **Cartogram** 
   Circle size = deaths (i.e. number of deaths)
   Circle color - death1k (i.e. deaths per 1k)

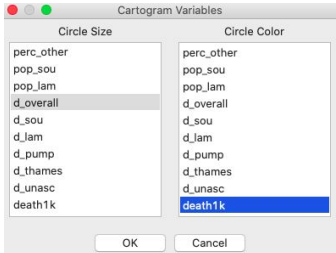


1. Conditional boxmap: variables





2. Cartogram variables



SOUTH LONDON EXPERIMENT: Unique Values Map and Boxplot

1 shapefile (shp, shx, dbf):

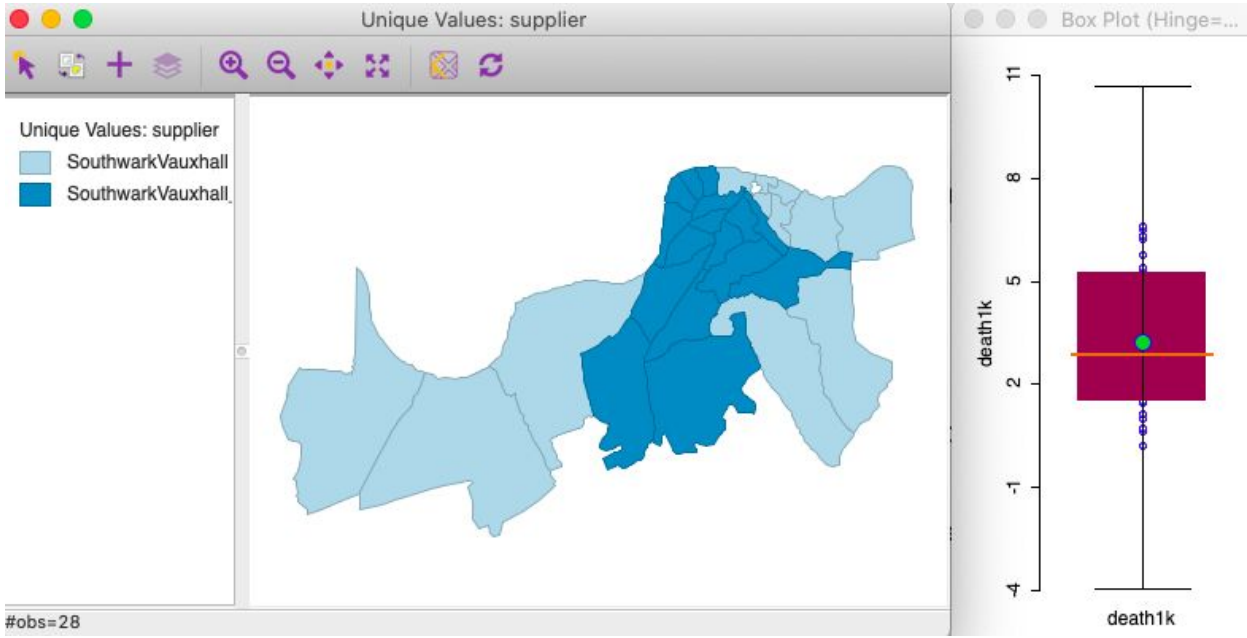
- subdistricts

Variables:

- **supplier**
- **death1k**

Functionality:

1. **Map-Unique Values Map** for ‘supplier’
2. **Explore-Box Plot** for ‘death1k’



REFERENCES

Arribas-Bel, D., de Graaff, T., & Rey, S. J. (2017). Looking at John Snow's Cholera map from the twenty first century: A practical primer on reproducibility and open science. In *Regional Research Frontiers*-Vol. 2 (pp. 283-306). Springer, Cham. Data can be downloaded from Dani Arribas-Bel's 'reproducible john snow' BitBucket repository at https://bitbucket.org/darribas/reproducible_john_snow/src/master/

Chave, S. P. W. (1958). *Henry Whitehead and Cholera in Broad Street*, Medical History, Volume 2, Number 2, pp. 92-108.

Coleman, T. (2019). *Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference*. Working paper. Available on SSRN at https://papers.ssrn.com/abstract=3262234. Data can be downloaded from https://github.com/tscoleman/SnowCholera (last accessed September 2, 2020).

Coleman, T. (2020). John Snow, Cholera, and South London Reconsidered. Working paper. Available on SSRN at https://papers.ssrn.com/abstract=3696028 Data can be downloaded from https://github.com/tscoleman/SnowCholera (last accessed September 2, 2020).

Snow, J. (1855). *On the Mode of Communication of Cholera*, London, second edition, Map 1, available at https://www.bl.uk/learning/images/makeanimpact/publichealth/large12735.html

Snow, J. (1855). *On the Mode of Communication of Cholera*, London, second edition, Map 2, reprinted in Jefferson, Tom (2007), Cattive acque. John Snow e la vera storia del colera a Londra, Rome, Il Pensiero Scientifico Editore.

Tobler, W. (1994). *Snow's Cholera Map*, http://www.ncgia.ucsb.edu/pubs/snow/snow.html. Data files were obtained from the HistData CRAN R package.

Vinten-Johansen, P. (Ed.). (2020). *Investigating Cholera in Broad Street: A History in Documents*. Broadview Press.

Wilson, R (2011). *John Snow's Cholera data in more formats*, http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/. Reprojected data can also be downloaded from Dani Arribas-Bel's 'reproducible john snow' BitBucket repository at https://bitbucket.org/darribas/reproducible_john_snow/src/master/