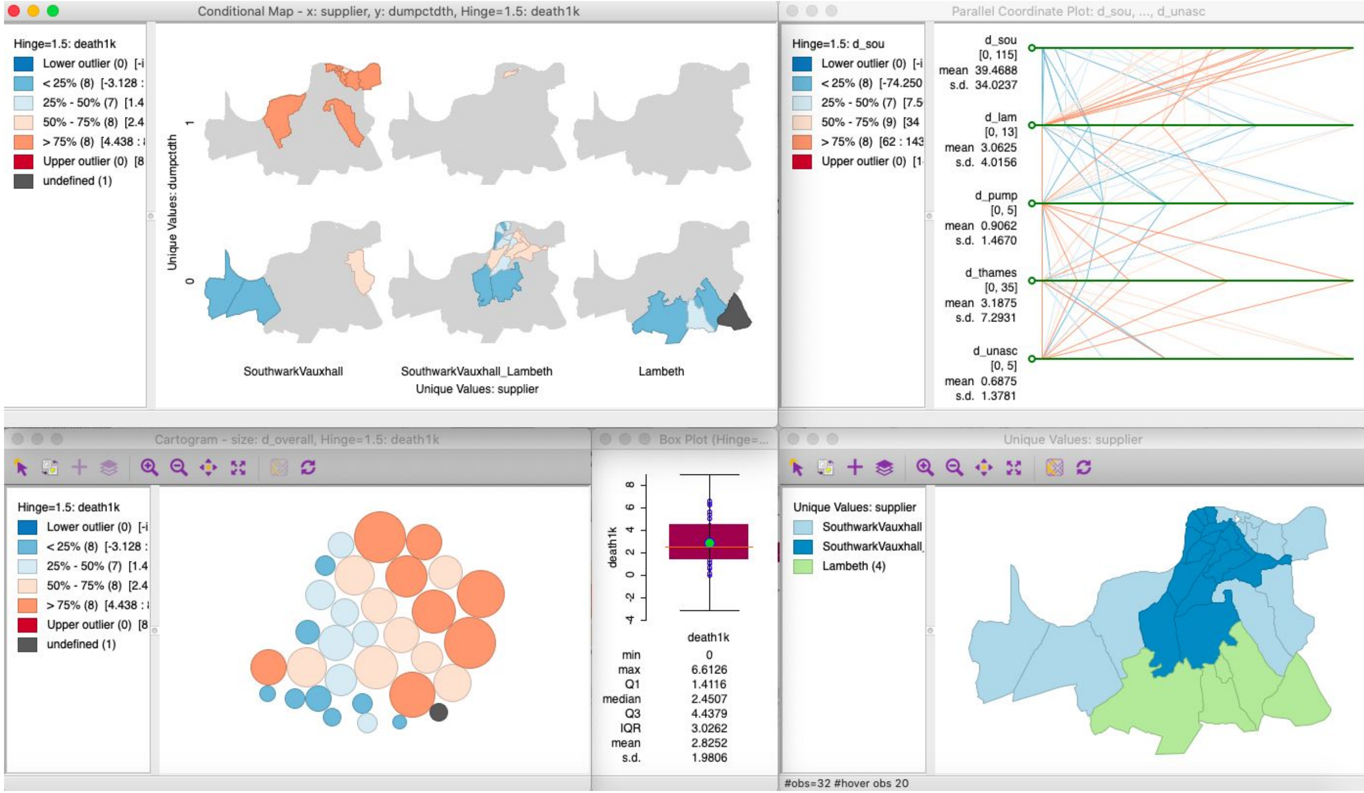
# EDA and ESDA with GeoDa

## 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



```
min       0
max       6.6126
Q1        1.4116
median    2.4507
Q3        4.4379
IQR       3.0262
mean      2.8252
s.d.      1.9806
```

#obs=32 #hover obs 20

# 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

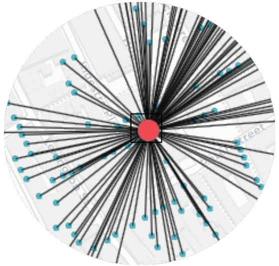
## Broad St Pump

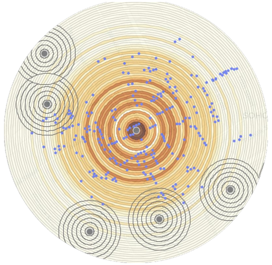
578 individual cholera deaths
Dataset 1

D'Arblay St
Broadwick St
cant (34)

Cholera deaths in 40 housing blocks
Dataset 3

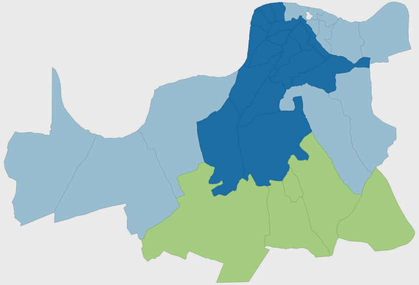
250 cholera deaths by building
Datasets 2 + 4





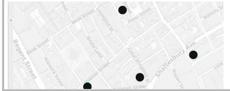
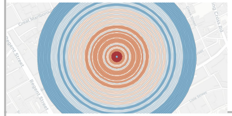
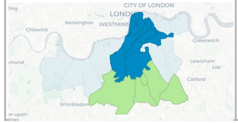
Cholera deaths around Broad St pump
Datasets 4, 5 + 6

## S. London Experiment



Results for 32 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.<br>deaths | Individual deaths | Broad St Pump | Point | 578 | 4 | Tobler 1994, Arribas-Bel et al. 2017 | Snow 1855 (Map 1) | GPL |
| | 2.<br>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.<br>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.<br>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.<br>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.<br>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.<br>subdistricts | London subdistricts as of 1855 with data | South London Natural Experiment | Polygon | 32 | 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

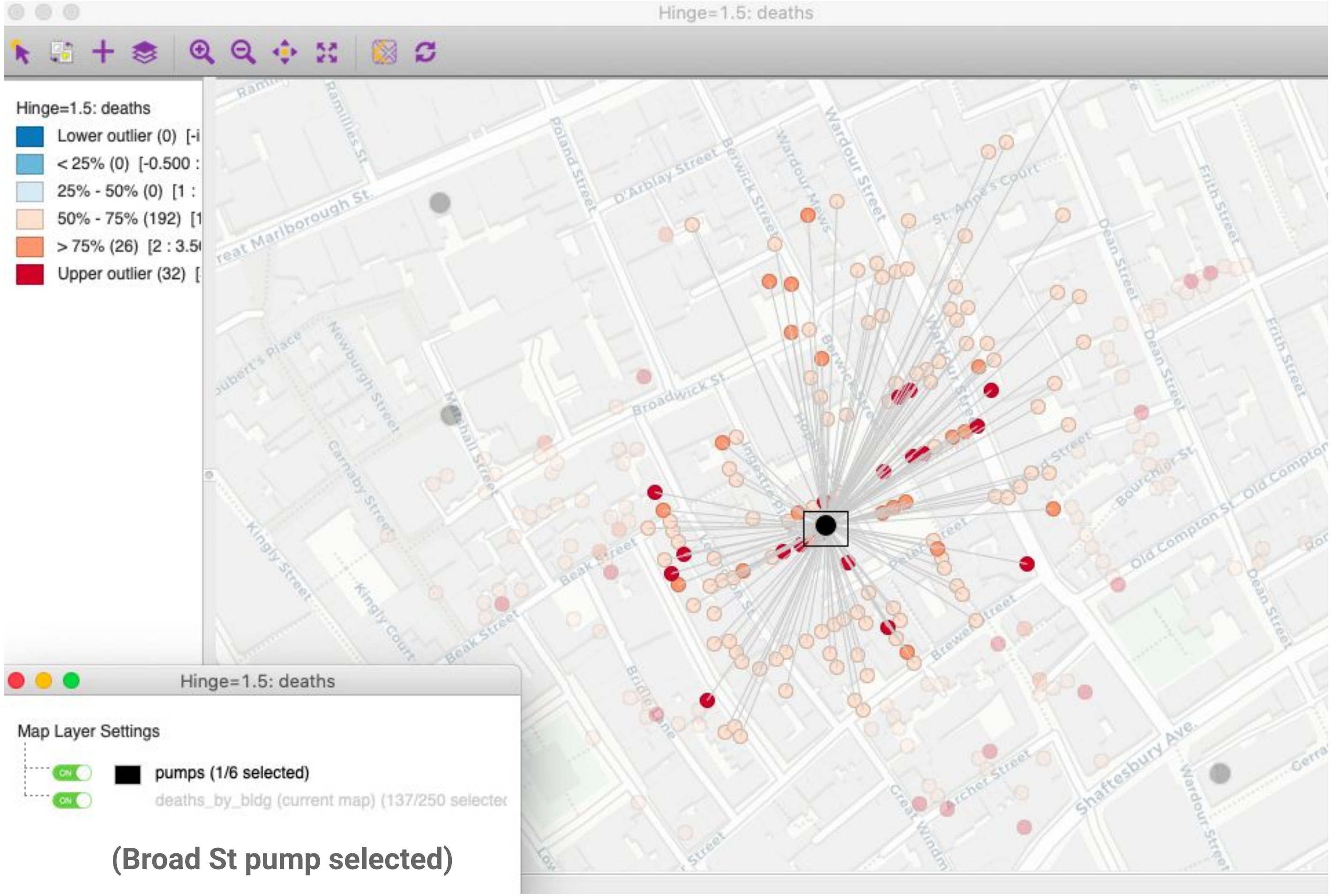
# IDENTIFYING CLUSTERS AND SPATIAL CONCENTRATIONS

# 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



Hinge=1.5: deaths

Hinge=1.5: deaths

- 🟦 Lower outlier (0) [-i
- 🟦 < 25% (0) [-0.500 :
- ⬜ 25% - 50% (0) [1 :
- ⬜ 50% - 75% (192) [1
- 🟧 > 75% (26) [2 : 3.5
- 🟥 Upper outlier (32) [

Hinge=1.5: deaths

Map Layer Settings

ON ■ pumps (1/6 selected)

ON deaths_by_bldg (current map) (137/250 selectec

(Broad St pump selected)

# GeoDa Implementation

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

- deaths_by_bldg
- pumps

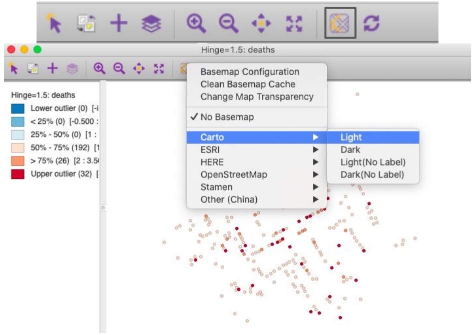
## VARIABLES

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

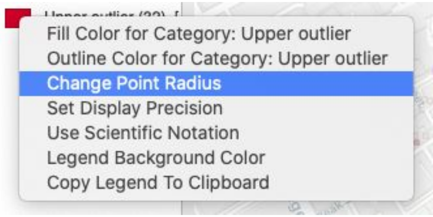
## STEPS

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

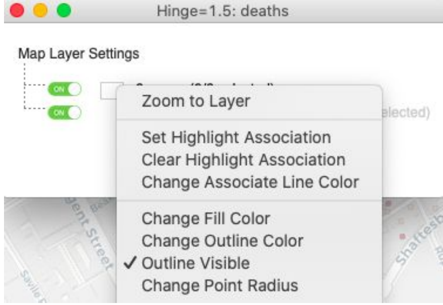
2. Add basemap



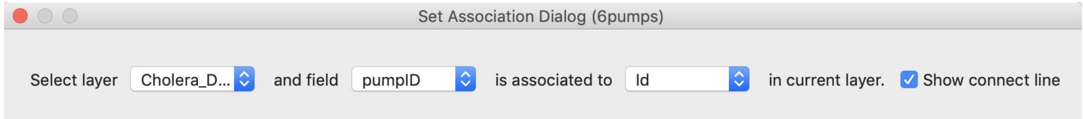
3. Change point radius



4. Change settings



4c. Set highlight association


Set Association Dialog (6pumps)

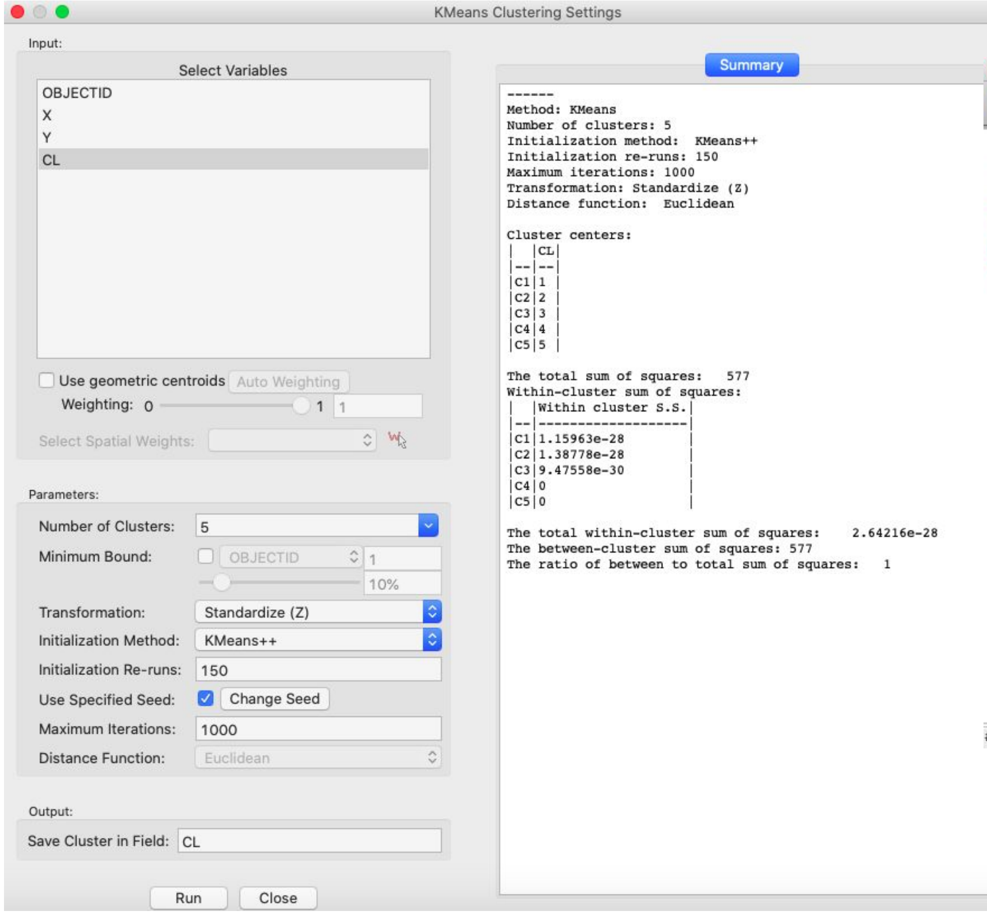
Select layer Cholera_D... and field pumpID is associated to Id in current layer. ☑ Show connect line

# 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)



KMeans Clustering Settings

Input:

Select Variables

- OBJECTID
- X
- Y
- CL

☐ Use geometric centroids  Auto Weighting

Weighting: 0 1 1

Select Spatial Weights:

Parameters:

Number of Clusters: 5

Minimum Bound: ☐ OBJECTID 1 10%

Transformation: Standardize (Z)

Initialization Method: KMeans++

Initialization Re-runs: 150

Use Specified Seed: ☑ Change Seed

Maximum Iterations: 1000

Distance Function: Euclidean

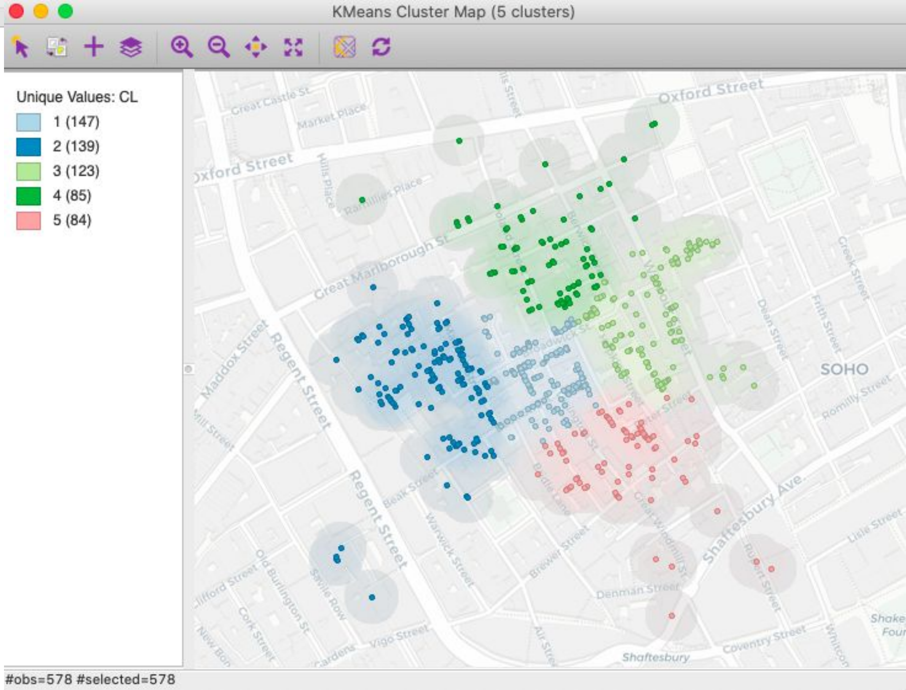
Output:

Save Cluster in Field: CL

Run  Close

Summary

```
------
Method: KMeans
Number of clusters: 5
Initialization method:  KMeans++
Initialization re-runs: 150
Maximum iterations: 1000
Transformation: Standardize (Z)
Distance function:  Euclidean

Cluster centers:
|  |CL|
|--|--|
|C1|1 |
|C2|2 |
|C3|3 |
|C4|4 |
|C5|5 |

The total sum of squares:   577
Within-cluster sum of squares:
|  |Within cluster S.S.|
|--|-------------------|
|C1|1.15963e-28        |
|C2|1.38778e-28        |
|C3|9.47558e-30        |
|C4|0                  |
|C5|0                  |

The total within-cluster sum of squares:    2.64216e-28
The between-cluster sum of squares: 577
The ratio of between to total sum of squares:   1
```



KMeans Cluster Map (5 clusters)

```
#obs=578 #selected=578
```

## Run Heat Maps on Clusters

# GeoDa Implementation

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

- deaths

**VARIABLE**

- CL

## STEPS

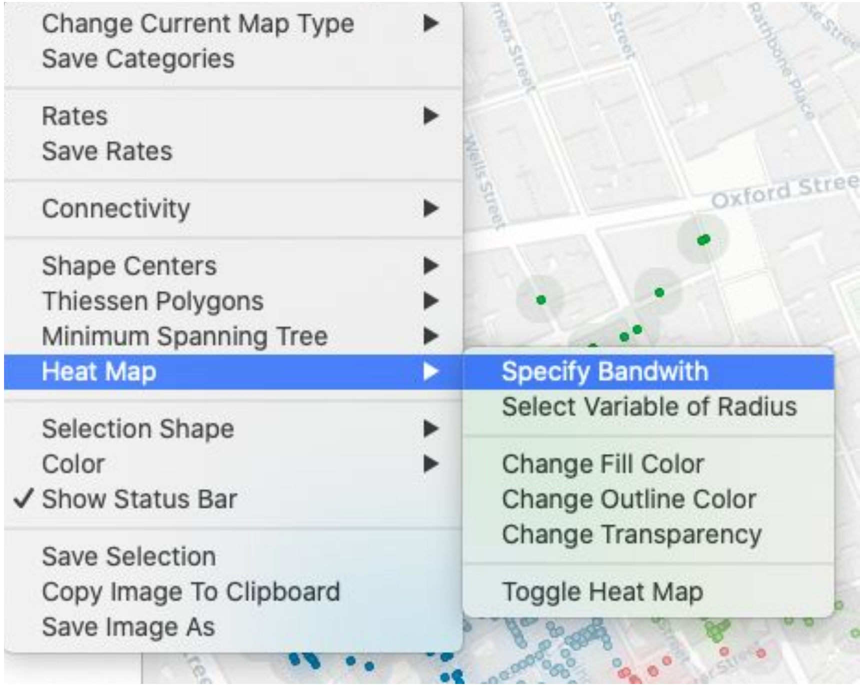
**Run a K-Means Clustering Analysis**

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

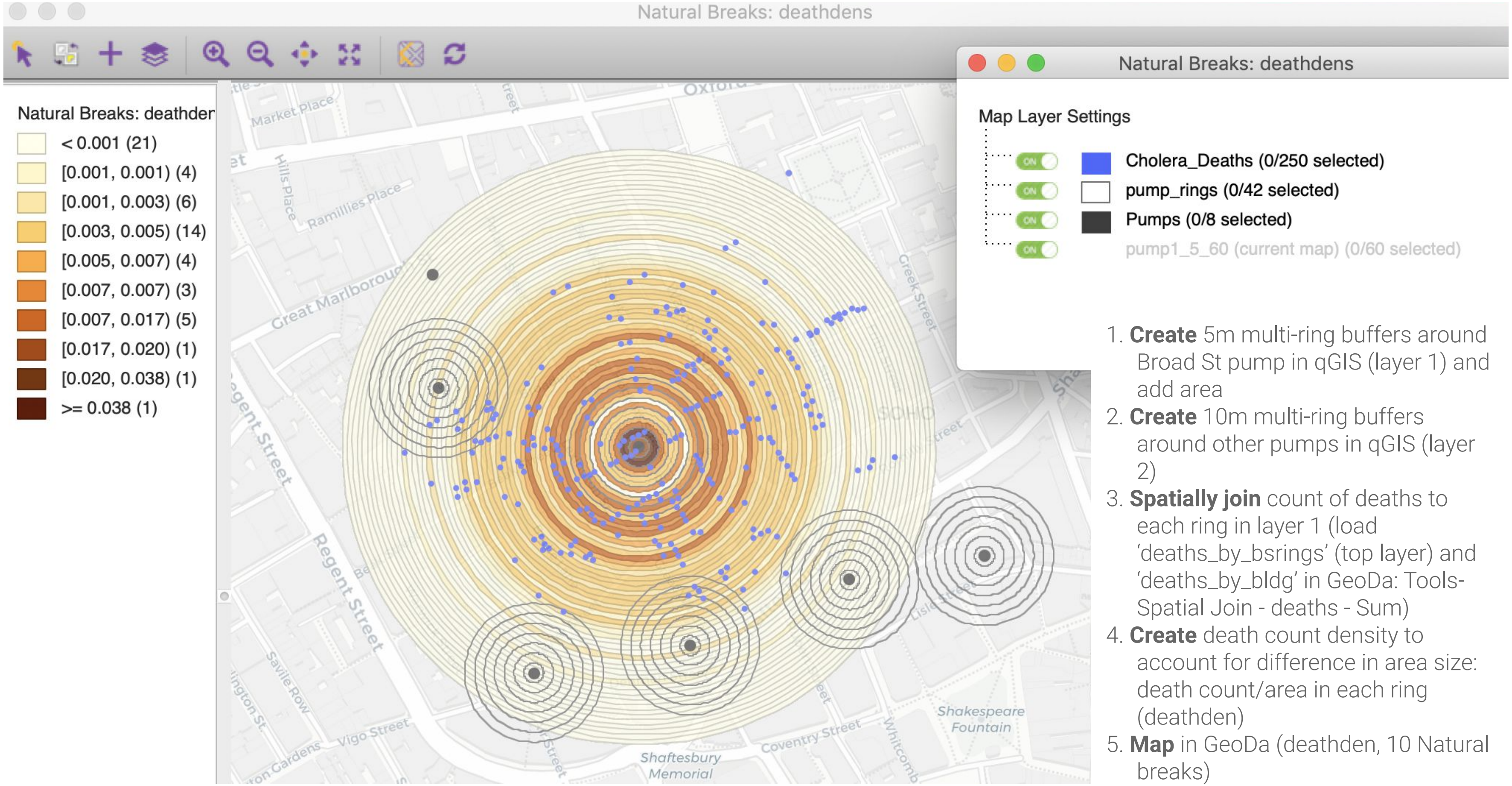


# STEP-BY-STEP EXAMPLE 3: IDENTIFYING DISTANCE DECAY

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

Resource Links

# More Deaths Near Broad St Pump: Distance Decay Demonstration



Natural Breaks: deathder

| | |
|---|---|
| □ | < 0.001 (21) |
| □ | [0.001, 0.001) (4) |
| 🟨 | [0.001, 0.003) (6) |
| 🟨 | [0.003, 0.005) (14) |
| 🟧 | [0.005, 0.007) (4) |
| 🟧 | [0.007, 0.007) (3) |
| 🟧 | [0.007, 0.017) (5) |
| 🟫 | [0.017, 0.020) (1) |
| 🟫 | [0.020, 0.038) (1) |
| 🟫 | >= 0.038 (1) |

Natural Breaks: deathdens

Map Layer Settings

| | | |
|---|---|---|
| ON | 🟦 | Cholera_Deaths (0/250 selected) |
| ON | □ | pump_rings (0/42 selected) |
| ON | ⬛ | Pumps (0/8 selected) |
| ON | | pump1_5_60 (current map) (0/60 selected) |

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)

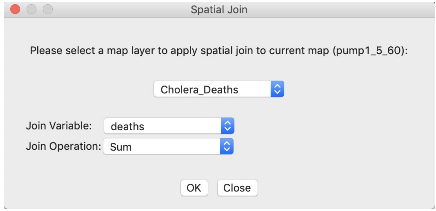
**Calculate death density:**

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



Spatial Join

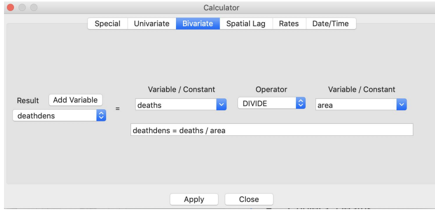
Please select a map layer to apply spatial join to current map (pump1_5_60):

Cholera_Deaths

Join Variable: deaths

Join Operation: Sum

7. Table - Calculator



Calculator

Result: deathdens

Variable / Constant: deaths

Operator: DIVIDE

Variable / Constant: area

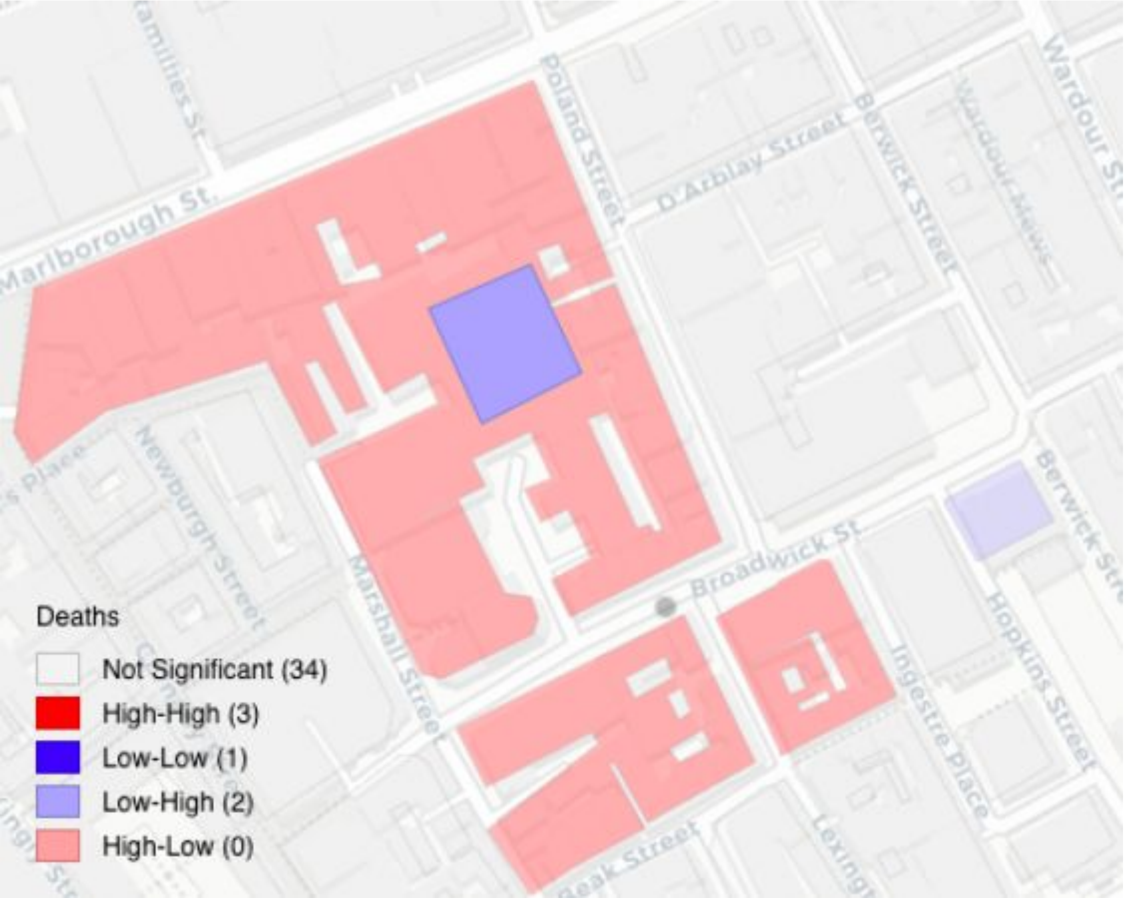
deathdens = deaths / area

# 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

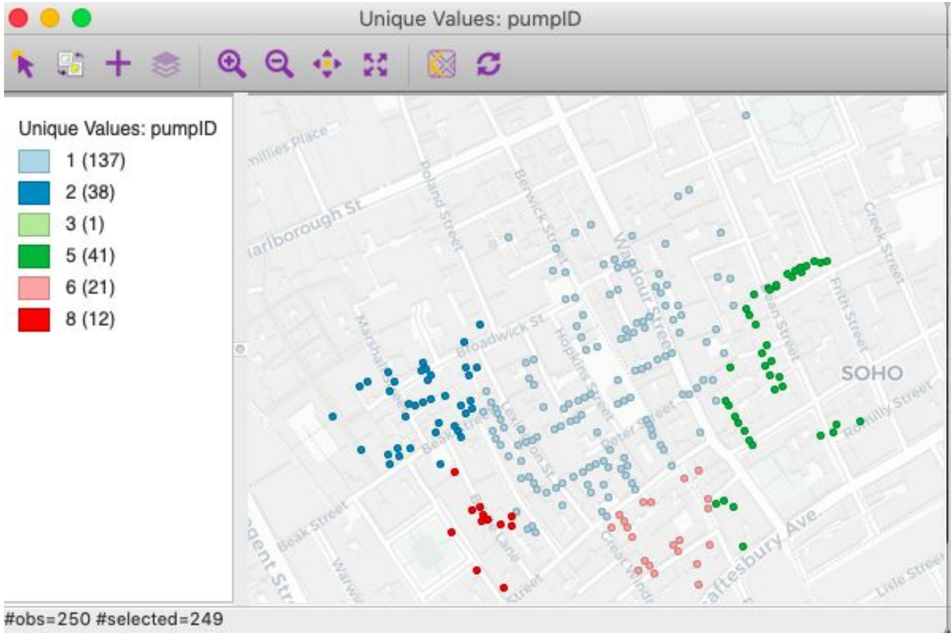
# COMPARING DISTRIBUTIONS ACROSS GROUPS

# STEP-BY-STEP EXAMPLE 5: CONDITIONAL BOX PLOTS

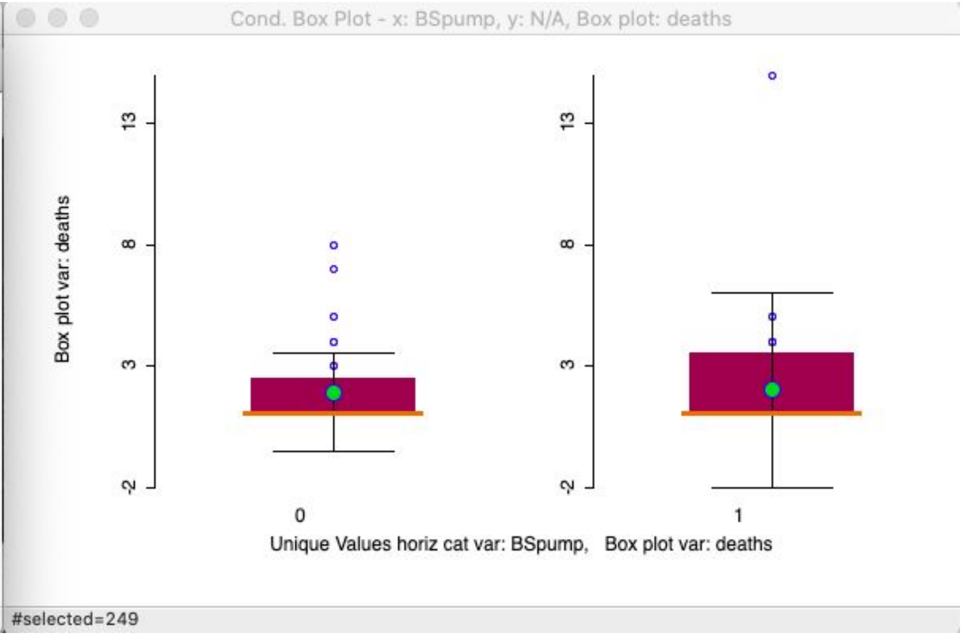
Comparing distributions across groups:
Compare deaths near & further from pump

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

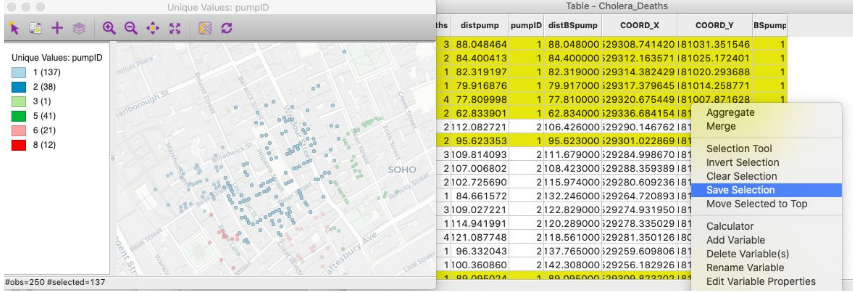
**VARIABLES**

- deaths
- pumpID

**STEPS**

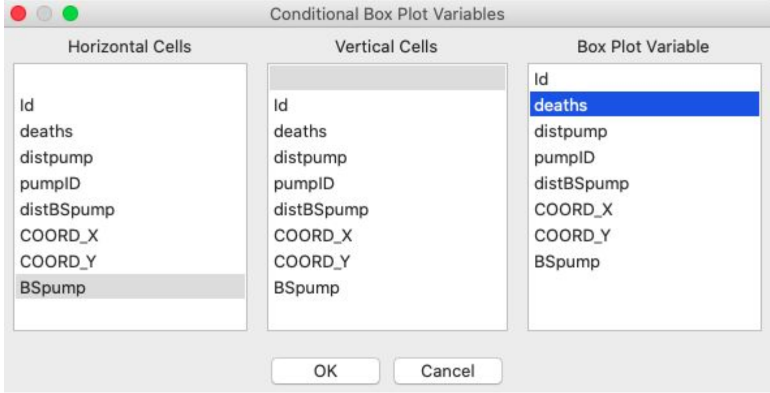
1. **Map-Unique Values Map** - Select "pumpID".
2. **Add Basemap** (Carto Light)
3. **Select category 1** in unique values map legend (pumpID = 1)
4. **Table** - **Save selection** as new variable (**BSpump**): buildings with deaths where Broad St pump is closest (1) or other pump is closest (0)
5. **Explore-Conditional 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



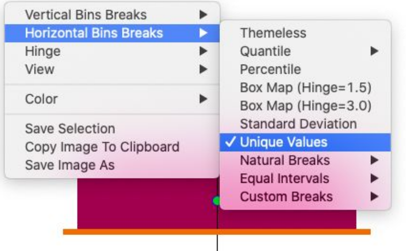
| ths | distpump | pumpID | distBSpump | COORD_X | COORD_Y | BSpump |
|---|---|---|---|---|---|---|
| 3 | 88.048464 | 1 | 88.048000 | 529308.741420 | 181031.351546 | 1 |
| 2 | 84.400413 | 1 | 84.400000 | 529312.163571 | 181025.172401 | 1 |
| 1 | 82.319197 | 1 | 82.319000 | 529314.382429 | 181020.293688 | 1 |
| 1 | 79.916876 | 1 | 79.917000 | 529317.379645 | 181014.258771 | 1 |
| 4 | 77.809998 | 1 | 77.810000 | 529320.675449 | 181007.871628 | 1 |
| 2 | 62.833901 | 1 | 62.834000 | 529336.684154 | 181 | |
| 2 | 112.082721 | 2 | 106.426000 | 529290.146762 | 181 | |
| 2 | 95.623353 | 1 | 95.623000 | 529301.022869 | 181 | |
| 3 | 109.814093 | 2 | 111.679000 | 529284.998670 | 181 | |
| 2 | 107.006802 | 2 | 108.423000 | 529288.359389 | 181 | |
| 2 | 102.725690 | 2 | 115.974000 | 529280.609236 | 181 | |
| 1 | 84.661572 | 2 | 132.246000 | 529264.720893 | 181 | |
| 3 | 109.027221 | 2 | 122.829000 | 529274.931950 | 181 | |
| 1 | 114.941991 | 2 | 120.289000 | 529278.335029 | 181 | |
| 4 | 121.087748 | 2 | 118.561000 | 529281.350126 | 180 | |
| 1 | 96.332043 | 2 | 137.765000 | 529259.609806 | 181 | |
| 1 | 100.360860 | 2 | 142.308000 | 529256.182926 | 181 | |
| 1 | 89.095024 | 1 | 89.095000 | 529309.823202 | 181 | |

5. Select variables



| Horizontal Cells | Vertical Cells | Box Plot Variable |
|---|---|---|
| | | Id |
| Id | Id | deaths |
| deaths | deaths | distpump |
| distpump | distpump | pumpID |
| pumpID | pumpID | distBSpump |
| distBSpump | distBSpump | COORD_X |
| COORD_X | COORD_X | COORD_Y |
| COORD_Y | COORD_Y | BSpump |
| BSpump | BSpump | |

5.a. Modify horizontal bin breaks



# THE SOUTH LONDON NATURAL EXPERIMENT

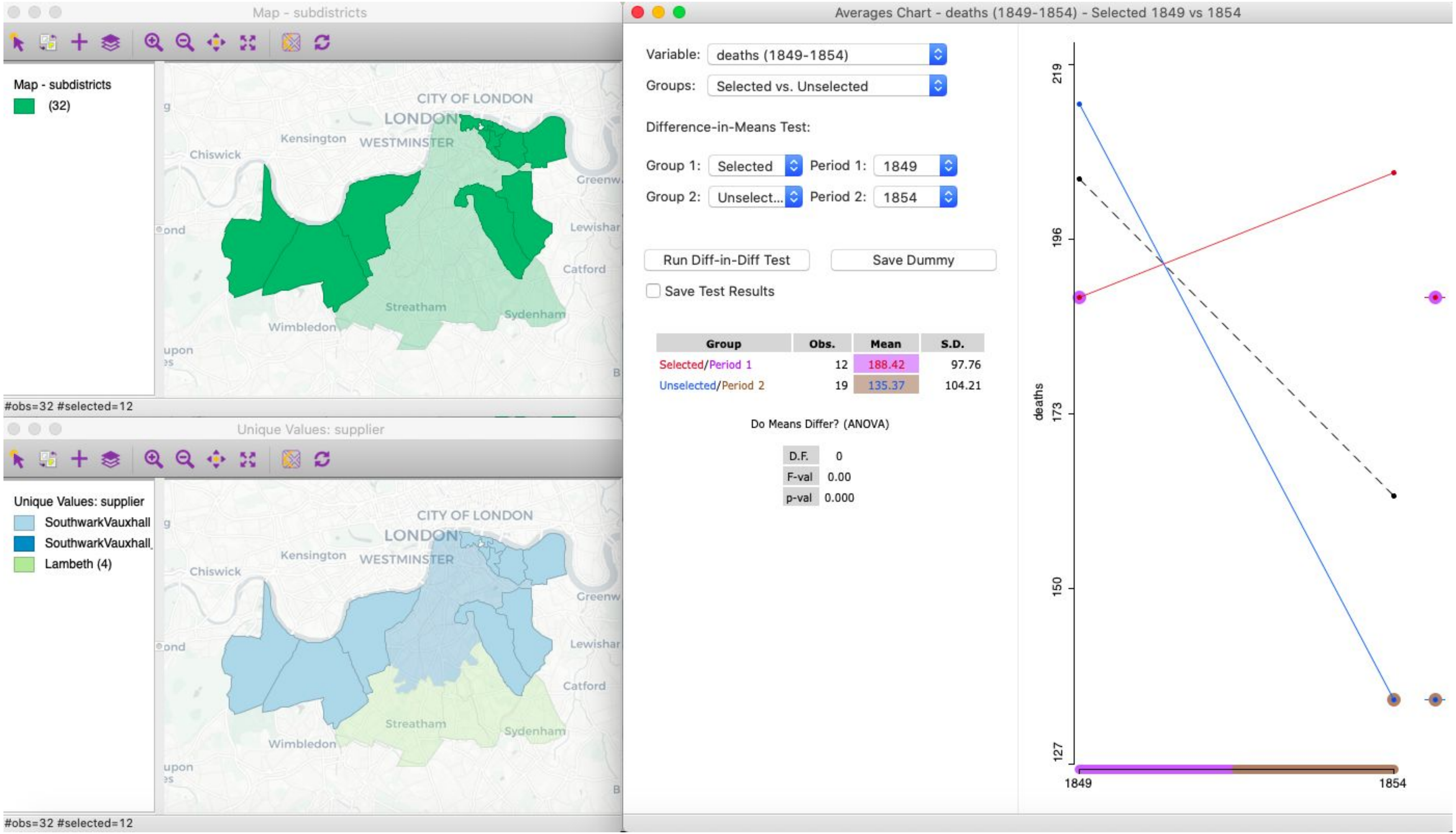
# COMPARING TRENDS

# 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 EXP.: SW Water Supplier Has Worse Cholera Death Trend Than SW-Lambeth



Map - subdistricts

Map - subdistricts

🟩 (32)

#obs=32 #selected=12

Unique Values: supplier

Unique Values: supplier

- SouthwarkVauxhall
- SouthwarkVauxhall_
- Lambeth (4)

#obs=32 #selected=12

Averages Chart - deaths (1849-1854) - Selected 1849 vs 1854

Variable: deaths (1849-1854)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 1849

Group 2: Unselect... Period 2: 1854

Run Diff-in-Diff Test    Save Dummy

☐ Save Test Results

| Group | Obs. | Mean | S.D. |
|---|---|---|---|
| Selected/Period 1 | 12 | 188.42 | 97.76 |
| Unselected/Period 2 | 19 | 135.37 | 104.21 |

Do Means Differ? (ANOVA)

| | |
|---|---|
| D.F. | 0 |
| F-val | 0.00 |
| p-val | 0.000 |

# GeoDa Implementation

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

- subdistricts

**VARIABLES**

- deaths1849
- deaths1854

**STEPS**

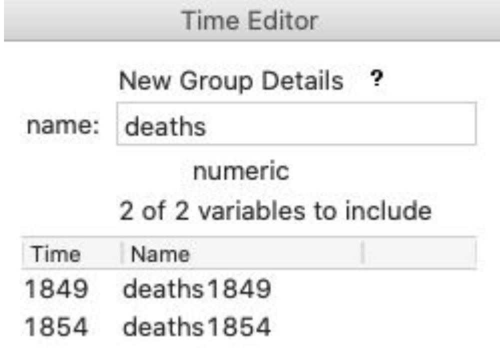
**Creating a time variable:**

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



Time Editor

New Group Details ?

name: deaths

numeric

2 of 2 variables to include

| Time | Name |
|---|---|
| 1849 | deaths1849 |
| 1854 | deaths1854 |

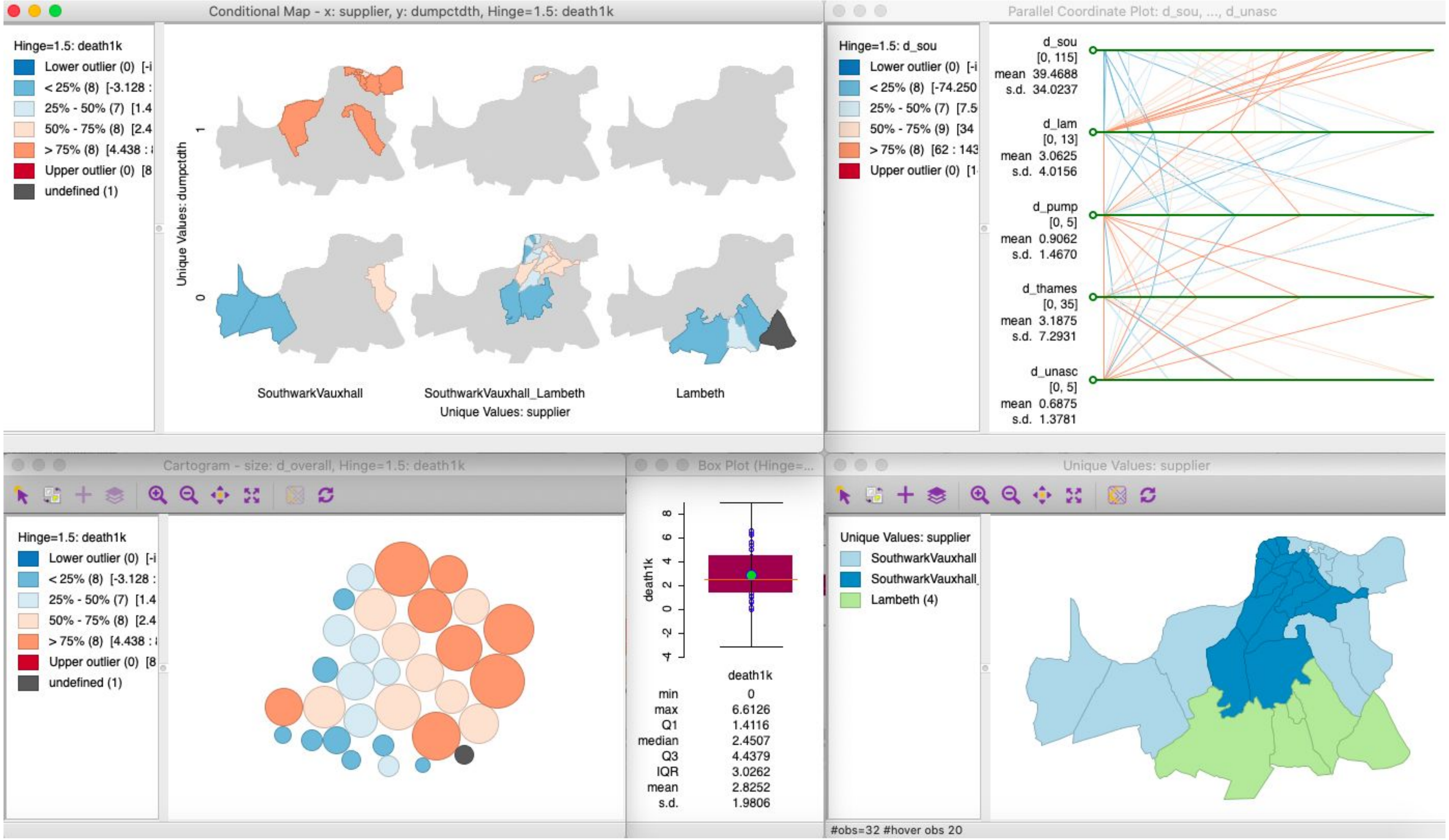
# EXPLORING A QUESTION WITH MULTIPLE EDA + ESDA TOOLS

# 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



Conditional Map - x: supplier, y: dumpctdth, Hinge=1.5: death1k

Hinge=1.5: death1k
- Lower outlier (0) [-i
- < 25% (8) [-3.128 :
- 25% - 50% (7) [1.4
- 50% - 75% (8) [2.4
- > 75% (8) [4.438 :
- Upper outlier (0) [8
- undefined (1)

Parallel Coordinate Plot: d_sou, ..., d_unasc

Hinge=1.5: d_sou
- Lower outlier (0) [-i
- < 25% (8) [-74.250
- 25% - 50% (7) [7.5
- 50% - 75% (9) [34
- > 75% (8) [62 : 143
- Upper outlier (0) [1

```
   d_sou
 [0, 115]
mean 39.4688
s.d. 34.0237

   d_lam
  [0, 13]
mean 3.0625
s.d. 4.0156

  d_pump
   [0, 5]
mean 0.9062
s.d. 1.4670

 d_thames
  [0, 35]
mean 3.1875
s.d. 7.2931

  d_unasc
   [0, 5]
mean 0.6875
s.d. 1.3781
```

Cartogram - size: d_overall, Hinge=1.5: death1k

Hinge=1.5: death1k
- Lower outlier (0) [-i
- < 25% (8) [-3.128 :
- 25% - 50% (7) [1.4
- 50% - 75% (8) [2.4
- > 75% (8) [4.438 :
- Upper outlier (0) [8
- undefined (1)

Box Plot (Hinge=...

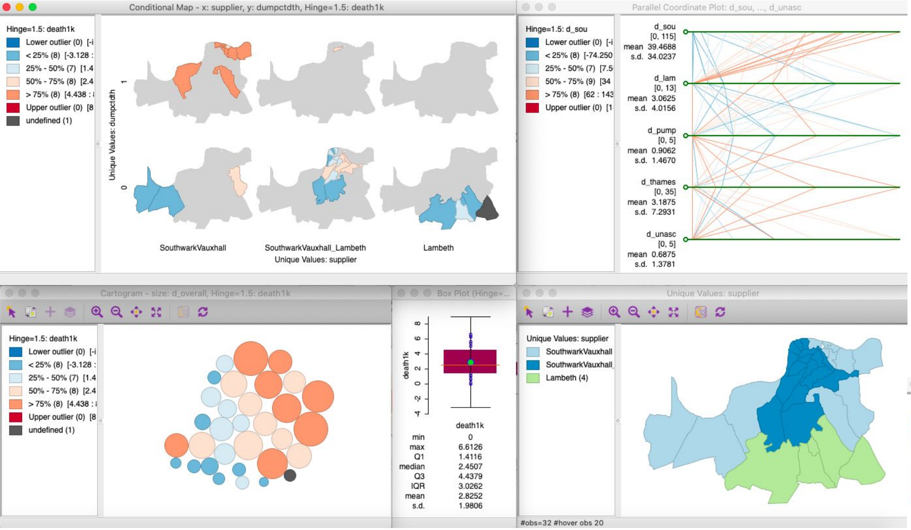
```
        death1k
min     0
max     6.6126
Q1      1.4116
median  2.4507
Q3      4.4379
IQR     3.0262
mean    2.8252
s.d.    1.9806
```

Unique Values: supplier

Unique Values: supplier
- SouthwarkVauxhall
- SouthwarkVauxhall_
- Lambeth (4)

#obs=32 #hover obs 20

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

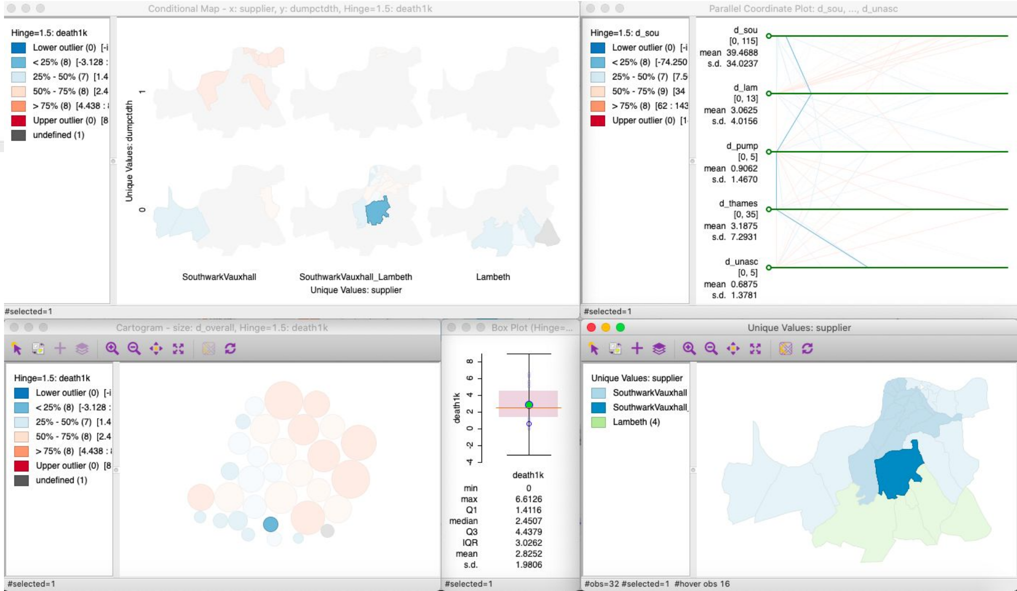


```
d_sou
[0, 115]
mean 39.4688
s.d. 34.0237

d_lam
[0, 13]
mean 3.0625
s.d. 4.0156

d_pump
[0, 5]
mean 0.9062
s.d. 1.4670

d_thames
[0, 35]
mean 3.1875
s.d. 7.2931

d_unasc
[0, 5]
mean 0.6875
s.d. 1.3781
```

```
death1k
min        0
max        6.6126
Q1         1.4116
median     2.4507
Q3         4.4379
IQR        3.0262
mean       2.8252
s.d.       1.9806
```

#obs=32 #hover obs 20

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



#selected=1

```
d_sou
[0, 115]
mean 39.4688
s.d. 34.0237

d_lam
[0, 13]
mean 3.0625
s.d. 4.0156

d_pump
[0, 5]
mean 0.9062
s.d. 1.4670

d_thames
[0, 35]
mean 3.1875
s.d. 7.2931

d_unasc
[0, 5]
mean 0.6875
s.d. 1.3781
```

```
death1k
min        0
max        6.6126
Q1         1.4116
median     2.4507
Q3         4.4379
IQR        3.0262
mean       2.8252
s.d.       1.9806
```

#obs=32 #selected=1 #hover obs 16

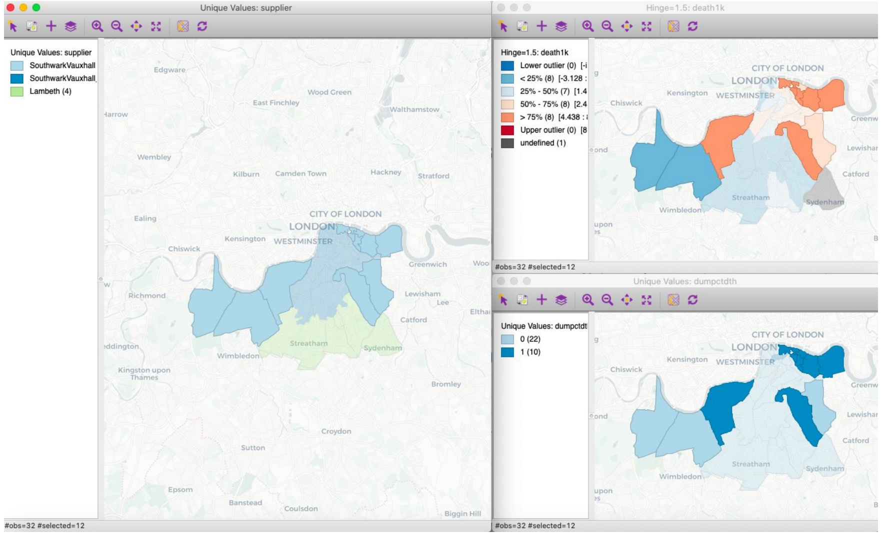
# SOUTH LONDON EXPERIMENT

## 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**



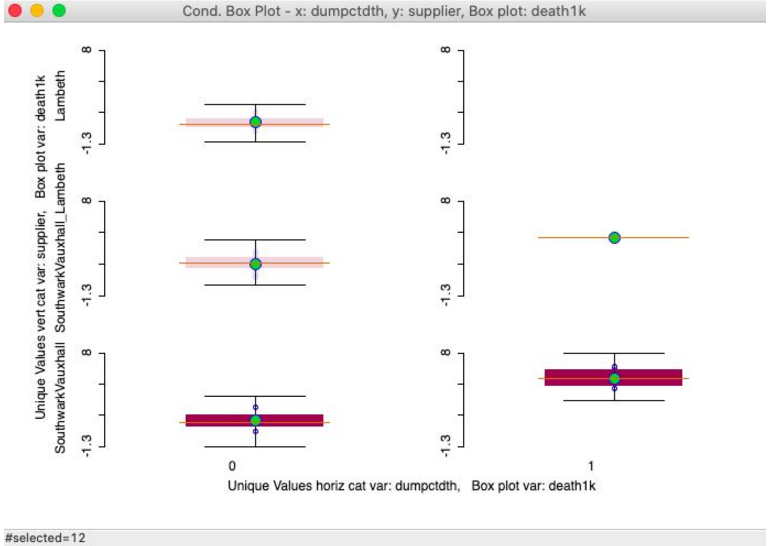
by water
**supplier**

Unique Values Map: dumpctdth
(**dumpctdth**: 0 = 0-3 deaths/1k, 1 = 4-14)

### Conditional Boxplot

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

**death1k**



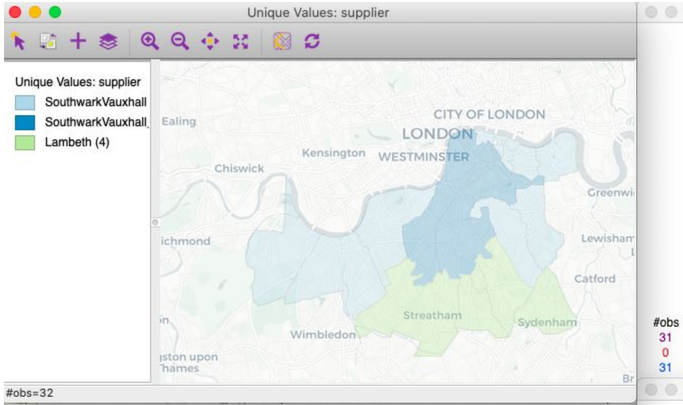
by low-high death1k category
(**dumpctdth**: 0 = 0-3 deaths/1k, 1 = 4-14)

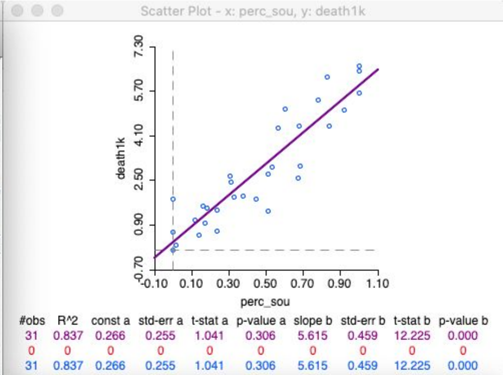
# SOUTH LONDON EXPERIMENT

## Higher Share of Deaths in Subdistricts Associated with Southwark Water Company

**Scatterplot | death1k**: Cholera deaths per 1000 people

Unique Values Map:
Water **supplier**

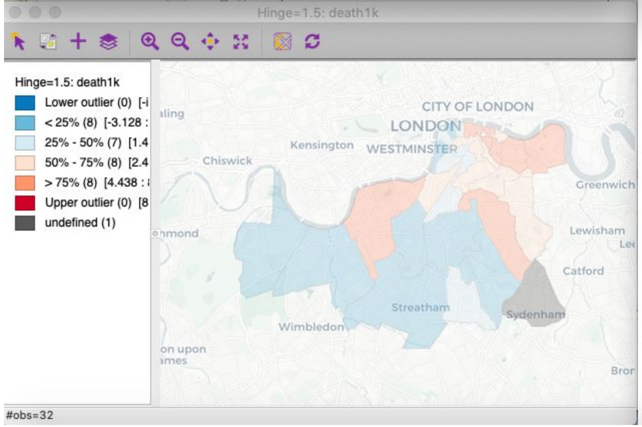


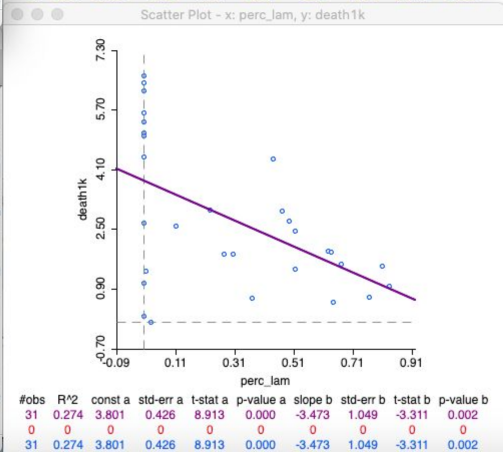


| #obs | R^2 | const a | std-err a | t-stat a | p-value a | slope b | std-err b | t-stat b | p-value b |
|---|---|---|---|---|---|---|---|---|---|
| 31 | 0.837 | 0.266 | 0.255 | 1.041 | 0.306 | 5.615 | 0.459 | 12.225 | 0.000 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | 0.837 | 0.266 | 0.255 | 1.041 | 0.306 | 5.615 | 0.459 | 12.225 | 0.000 |

**perc_sou**: % population served by Southwark & Vauxhall company

Boxmap:
**death1k**
(Cholera deaths per 1000 people)





| #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 |
|---|---|---|---|---|---|---|---|---|---|
| 31 | 0.274 | 3.801 | 0.426 | 8.913 | 0.000 | -3.473 | 1.049 | -3.311 | 0.002 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | 0.274 | 3.801 | 0.426 | 8.913 | 0.000 | -3.473 | 1.049 | -3.311 | 0.002 |

**perc_lam**: % population served by Lambeth 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’ **- Add** (this adds death1k to table)
- **Table-Calculator-Bivariate-death1k**: death1k → ‘d_overall’ DIVIDE ‘pop1854’ (decimals: 6, display 6) - **Apply**
- **Table-Calculator-Bivariate-death1k**: death1k → ‘death1k’ MULTIPLY by 1000 (decimals: 6, display 6) **- Apply**

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-**Apply** (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

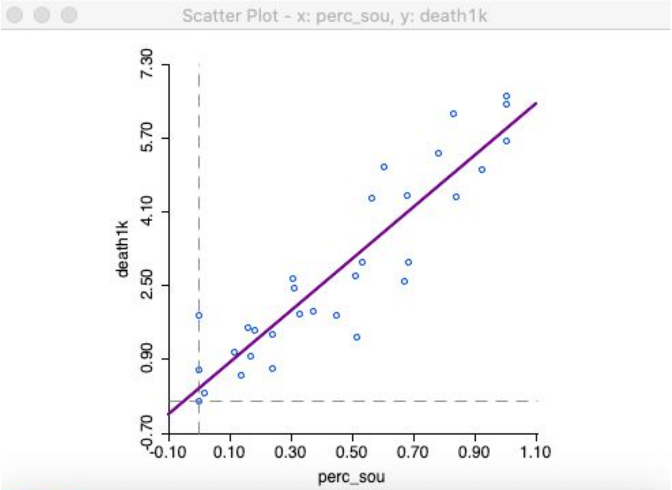
Close conditional boxplot and unique values map (dumpctdth)
Leave other two maps open (death1k and supplier)

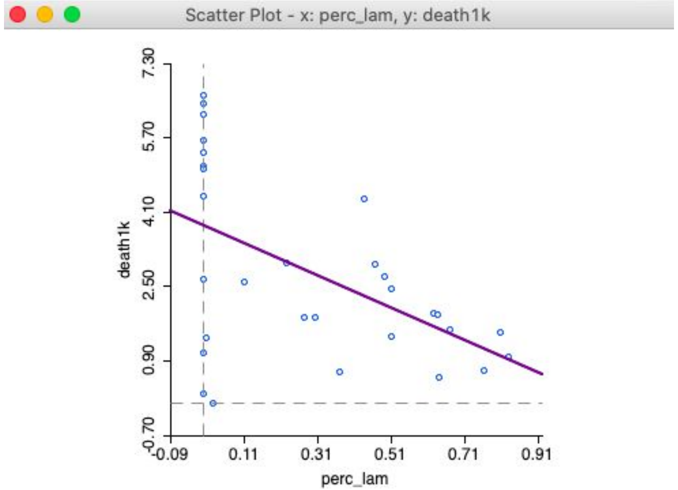
Variables:

- **death1k**
- **perc_lam:** % population served by Lambeth company
- **perc_south:** % population served by Southwark & Vauxhall company

Functionality:

1. Open scatterplot (**X: perc_sou**, **Y: death1k**)
2. Open scatterplot (**X: perc_lam**, **Y: death1k**)





# SOUTH LONDON EXPERIMENT: Parallel Coordinate Plot

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

- subdistricts

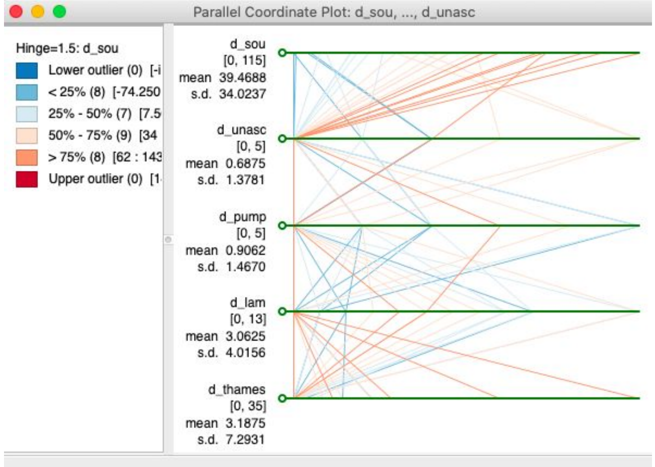
**VARIABLES**

Deaths attributed to ...

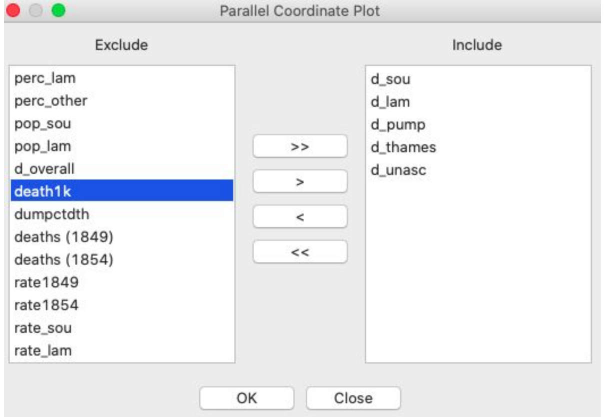
- **d_sou:** … the Southwark company
- **d_lam:** … the Lambeth company
- **d_pump:** … pumps or wells
- **d_thames:** … Thames water
- **d_unasc** … an unknown source

**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_unasc, d_pump, d_lam, d_thames**



1. Parallel coordinate plot variables



Parallel Coordinate Plot

| Exclude | Include |
|---|---|
| perc_lam | d_sou |
| perc_other | d_lam |
| pop_sou | d_pump |
| pop_lam | d_thames |
| d_overall | d_unasc |
| death1k | |
| dumpctdth | |
| deaths (1849) | |
| deaths (1854) | |
| rate1849 | |
| rate1854 | |
| rate_sou | |
| rate_lam | |

# SOUTH LONDON EXPERIMENT: Conditional Map and Cartogram

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

- subdistricts

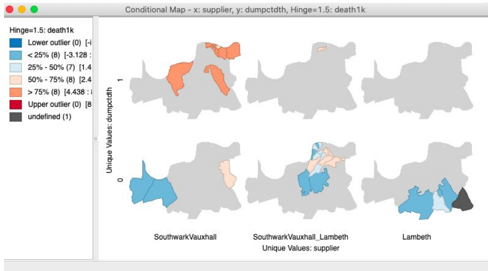
## VARIABLES

- **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

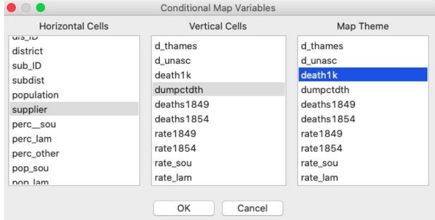
## STEPS

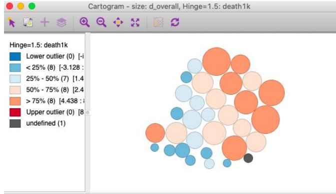
1. **Explore-Conditional Plot-Boxplot** with horizontal = **supplier**, vertical = **dumpctdth**, and map theme = **death1k** (2 rows, 2 columns)
   a. **Right-click: Change vertical bin breaks to unique values** for categorical representation of 0-1

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

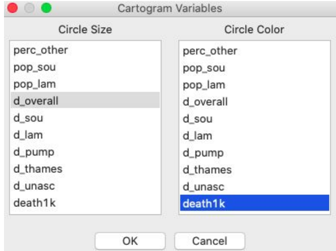


1. Conditional boxmap: variables





2. Cartogram variables



# SOUTH LONDON EXPERIMENT: Unique Values Map and Boxplot

1 shapefile (shp, shx, dbf):

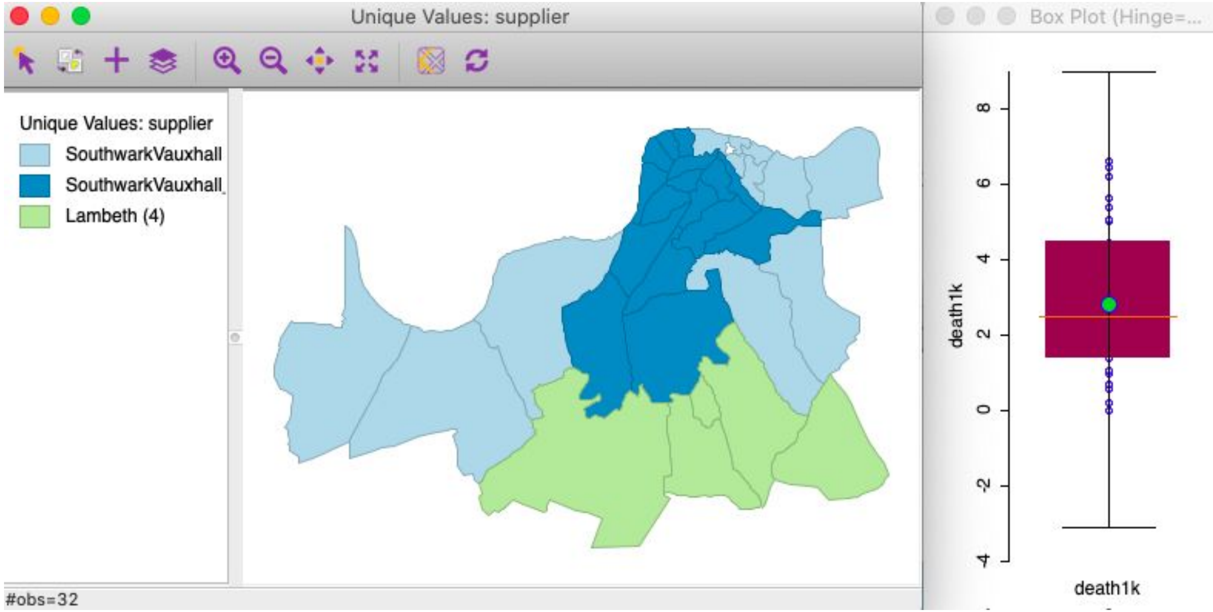
- subdistricts

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/