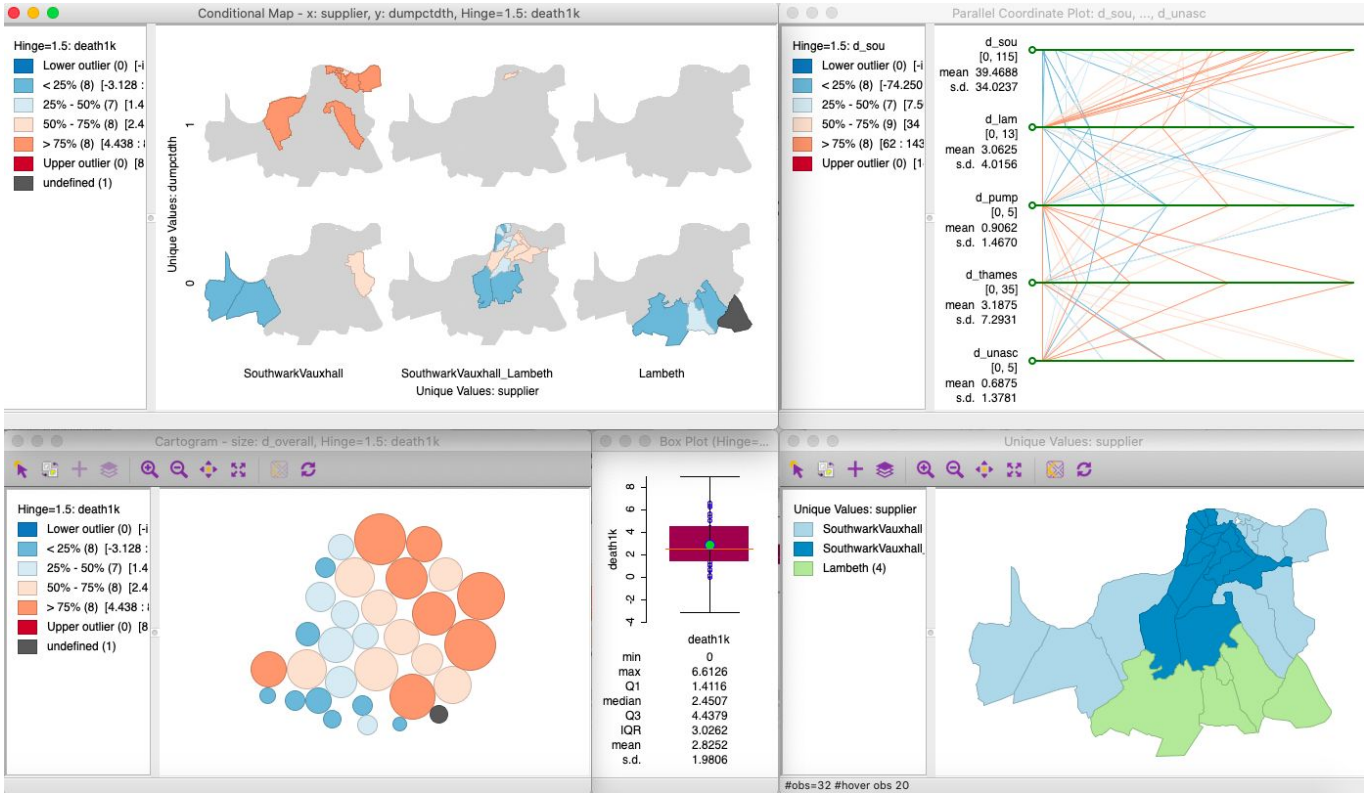
EDA and ESDA with GeoDa

John Snow & the 19th Century Cholera Epidemic

Julia Koschinsky
Marcos Falcone
spatial@uchicago.edu

September 2020
Edited January 2021

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

- Introductory Storymap: https://bit.ly/3mSGZiS (Video: https://youtu.be/lGN8SK1Y1h4)
- Storymap on Research Designs: https://rb.gy/vqjeoq (Video: https://bit.ly/2YmH6lp)
- YouTube Playlist - Spatial Insights Project: https://bit.ly/3loxlhi

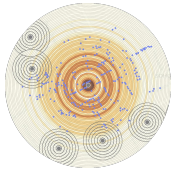
Examples and Spatial Data Files for Use in GeoDa

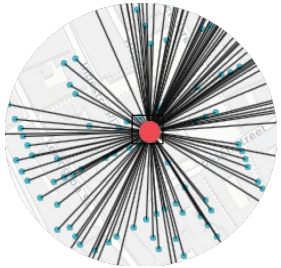
Cholera deaths in Soho

1,852 houses with cholera deaths and non-deaths
Dataset 1

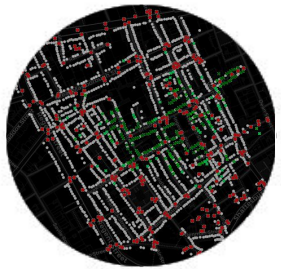
250 cholera deaths by building
Dataset 2

Cholera deaths by block and ring
Datasets 3+4+5





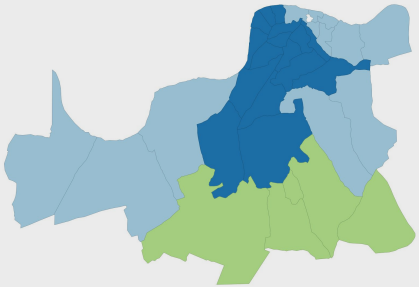
6 pumps
Dataset 6



325 sewer grates and ventilators
Dataset 7

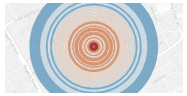
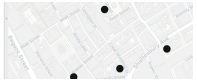
Correlates of deaths in Soho

S. London Experiment



32 subdistricts
Dataset 8

Overview of 8 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_nd_by_house	Deaths and non-deaths aggregated to houses	Broad St Pump	Point	1852	8	**Digitized by CSDS**	General Board of Health 1855	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. deaths_by_bsrings	Deaths aggregated to 5m rings around Broad St pump	Broad St Pump	Polygon	60	4	Tobler 1994, Wilson 2011, Arribas-Bel et al. 2017. **Rings + calculations by CSDS**	Snow 1855 (Map 1)	GPL
	5. 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
	6. pumps	Pumps in the Broad St area	Broad St Pump	Point	6	4	Wilson 2011, Arribas-Bel et al. 2017	Snow 1855 (Map 1)	Unknown
	7. sewergrates_ventilators	Untrapped sewer grates and ventilators	Broad St Pump	Point	325	5	**Digitized by CSDS**	General Board of Health 1855	GPL
	8. 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: **The Soho Outbreak**

CHOLERA DEATHS IN **SOHO:** A SUMMARY

In 1854, a cholera outbreak in the Soho neighborhood (London) took place. Compared to previous outbreaks, this one was particularly deadly, which prompted the medical and research community to further investigate the potential causes of cholera. Since many thought that cholera spread through toxic gases that were emanating from an old pest field, the Metropolitan Commission of Sewers charged Edmund Cooper with the task of discrediting this theory, which resulted in a map based on data we use here. Simultaneously, John Snow believed that cholera was transmitted through ingested water and thus that the culprit was the neighborhood's Broad Street pump. In what follows, we will also use his data to explore his theory with modern statistical tools. The scripts on the right allow you to explore both the airborne and waterborne hypothesis with the original data.

For more context, visit our Snow introductory storymap and our storymap on research designs.

CHOLERA DEATHS NEAR A **PEST FIELD, SEWER GRATES, AND BROAD ST PUMP**

Detecting Spatial Patterns:
Find spatial patterns of cholera deaths with different maps and multiple layers:
Unique Values, Standard Deviation and Natural Breaks Maps

Comparing Averages Across Groups:
Compare deaths counts close to and distant from potential correlates:
Averages Charts

Comparing Distributions Across Groups:
Compare deaths near a pest field, sewer grates, and pumps:
Conditional Box Plots

Identifying Clusters and Spatial Concentrations:
Find out where deaths are concentrated:
Identifying Spatial Concentrations Using the Univariate Local Join Count

MORE CHOLERA DEATHS NEAR **BROAD STREET PUMP?**

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

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

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

Comparing Distributions Across Groups:
Compare deaths near & far from a pump
Conditional Box Plots

Overview of GeoDa Scripts: **The South London Natural Experiment**

CHOLERA DEATHS IN **SOUTH LONDON:** A SUMMARY

In 1854, a different location within London also provided researchers with an opportunity to uncover the mode of transmission of cholera. Indeed, an outbreak that took place in South London was different from another that had occurred in 1849 because one of two water companies that served the area had changed the source of its water in the Thames river, whereas the other had not. Since the river was known to be polluted by sewage and John Snow was convinced that contamination of water was causing cholera to spread, this provided him with a unique opportunity to conduct a natural experiment to test whether differences in water supply led to changes in cholera deaths. The South London scripts on the right allow you to explore this theory in GeoDa with the original data from the natural experiment.

For more context on the rationale behind this research, visit our Snow introductory storymap and our storymap on research designs.

SOUTH LONDON NATURAL EXPERIMENT: MORE DEATHS WITH A SPECIFIC WATER SUPPLIER

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 SOHO OUTBREAK

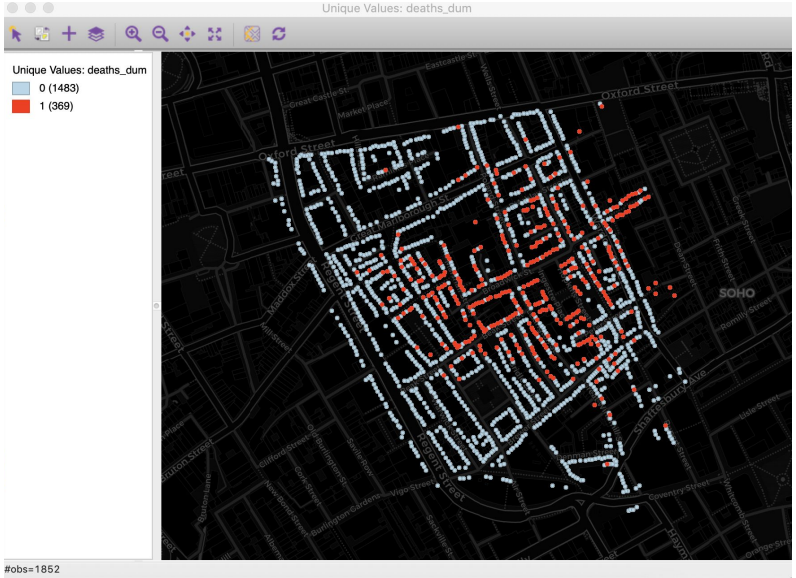
DETECTING SPATIAL PATTERNS: DIFFERENT MAP TYPES & MULTIPLE LAYERS

Detecting spatial patterns with maps (unique values, standard deviations, natural breaks) + multiple layers
Cholera deaths and non-deaths and potential correlates

Resource Links

The 1854 Soho Cholera Outbreak

1,852 houses with cholera deaths or no deaths recorded in the first 10 days of the outbreak



Unique Values Map

Houses with above-average numbers of cholera deaths

Standard Deviation Map

For context, visit our storymap on research designs.

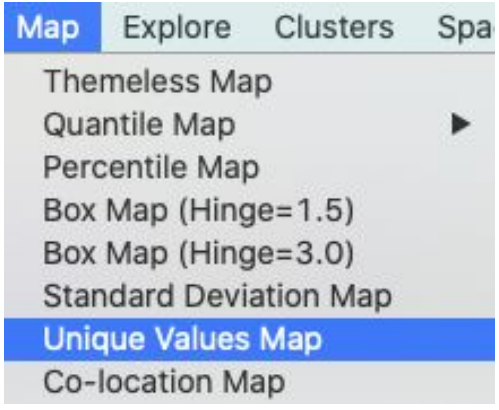
GeoDa Implementation

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

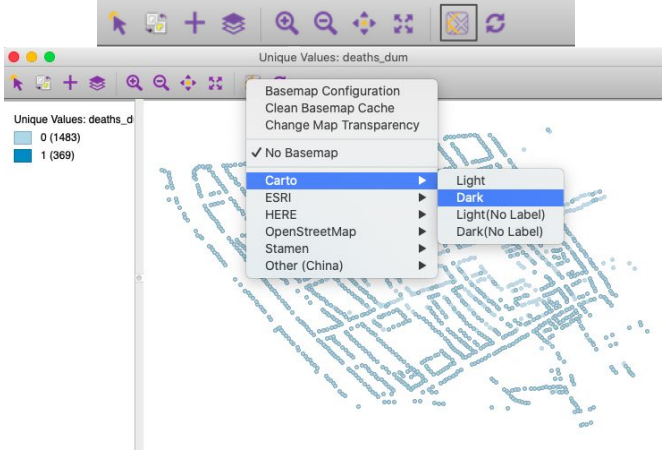
VARIABLES
- deaths_nd_by_house: **deaths**
- binary variable that distinguishes deaths vs. non-deaths: **deaths_dum**

STEPS
1. **Map-Unique Values Map-**deaths_dum
2. Add Basemap (Carto Dark) 
3. Right-click on legend for 1 and change fill + outline colors to red
4. **Map-Standard Deviation Map**-deaths
5. Add Basemap (Carto Dark) 

4. Unique Values Map



5. Add basemap

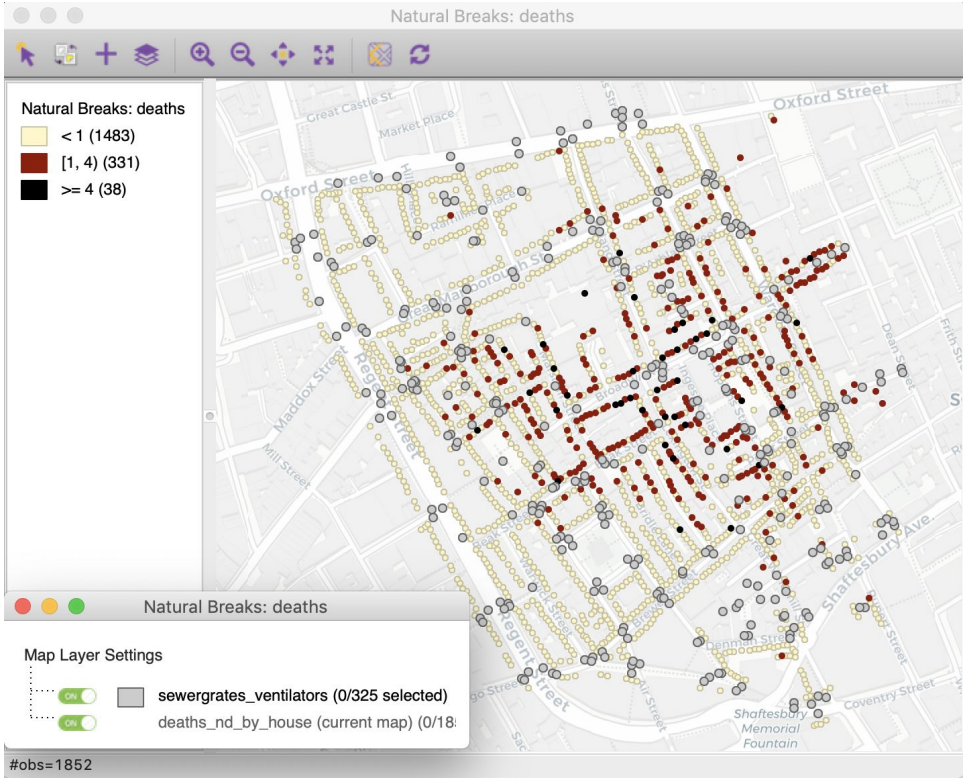


Deaths, Non-Deaths and Sewer Grates

The dominant theory of how cholera was transmitted in mid-19th century London was that it was airborne. If open sewer grates and ventilation shafts were related to the cause of cholera, we should see concentrations of deaths around them.

Let's replicate the [Metropolitan Commission of Sewers Map]:

Where are deaths and non-deaths in relation to sewer grates and ventilators?



For more context, visit our [storymap on research designs].

GeoDa Implementation

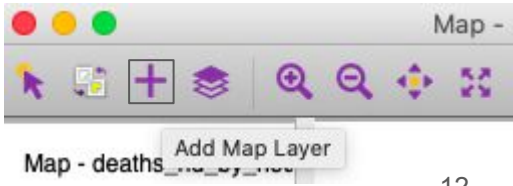
DATA - 2 shapefiles (shp, shx, dbf):
- deaths_nd_by_house
- sewergrates_ventilators

VARIABLE
- deaths_nd_by_house: **deaths**

STEPS
1. Load deaths_nd_by_house
2. Add layer to map: sewergrates_ventilators
3. Right click on sewergrates_ventilators, change point radius to 3
4. Click on sewergrates_ventilators and place on top of deaths_nd_by_house

Variable selection

2. Add layer to map

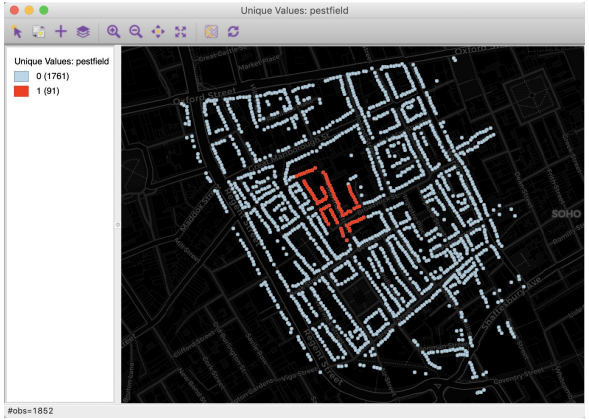


Deaths, Non-Deaths and the Pest Field

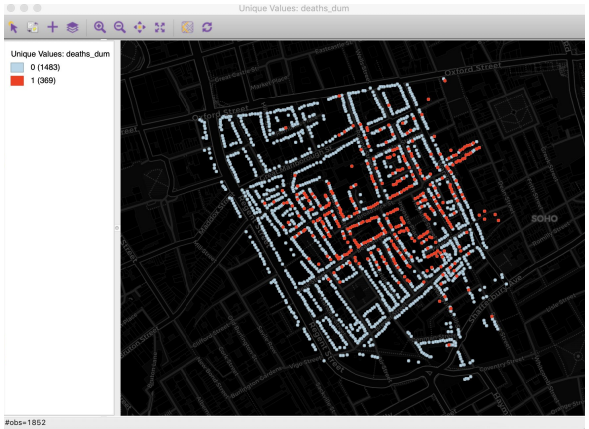
In the specific case of the Soho neighborhood, people thought that toxic gases were emanating from untrapped sewer grates and ventilation shafts located on or close to a 17th-century pest field. If this was true, we would expect to find more deaths near grates and shafts on the former pest field.

So let us replicate another aspect of the [Metropolitan Commission of Sewers Map]:

Where are deaths and non-deaths in relation to the former pest field?



Pest field
Unique Values Map



Cholera deaths
Unique Values Map

For context, visit our storymap on research designs.

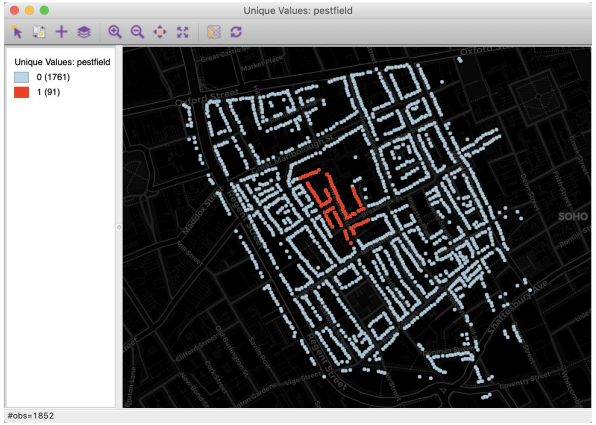
GeoDa Implementation

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

VARIABLE
- deaths_nd_by_house: **deaths**

STEPS
1. Load deaths_nd_by_house
2. **Map-Unique Values Map**
3. Variable: **pestfield**
4. Add basemap (Carto Dark)
5. Right-click on legend for 1 and change fill + outline colors to red

2. Unique values map of **pest field**

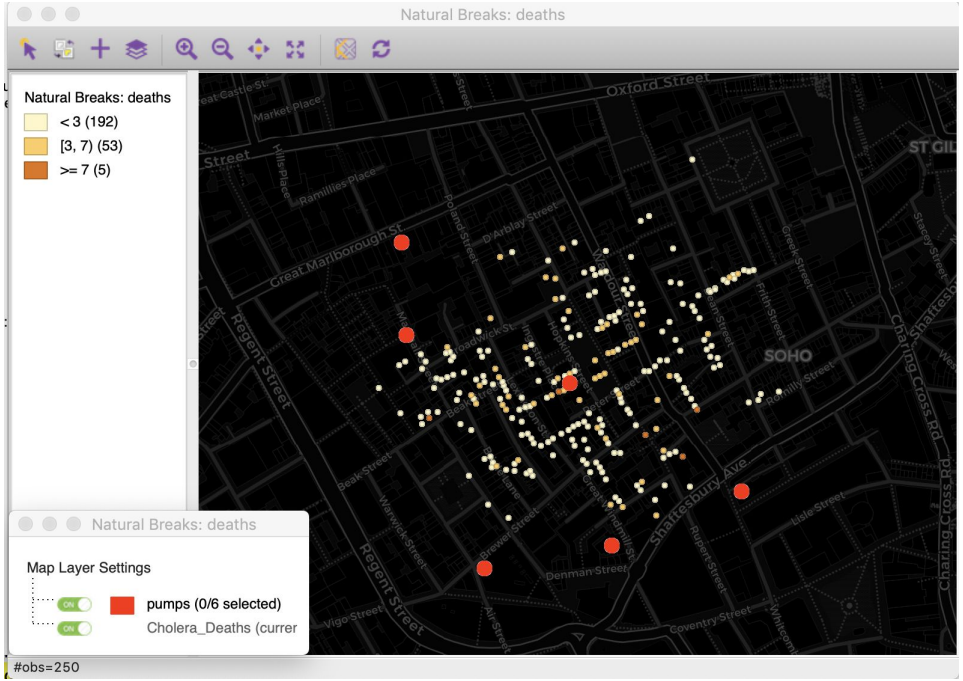


3. Variable selection

Deaths and Water Pumps

In contrast to the dominant airborne theory, John Snow held that cholera was transmitted by ingesting choleraic water. Since many people got their water from public pumps, Snow created a map to show deaths in relationship to water pumps. The famous Broad St pump is in the center of the map.

Let's replicate John Snow's map:
Where are deaths in relation to water pumps?



For context, visit our storymap on research designs.

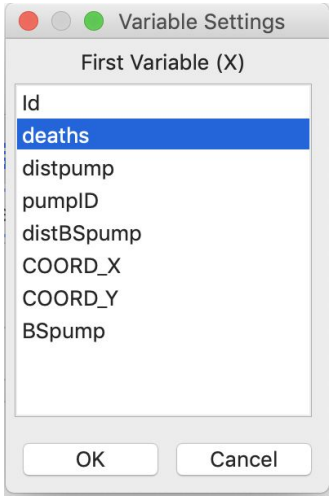
GeoDa Implementation

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

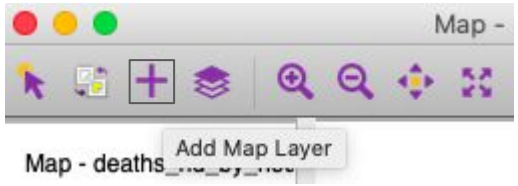
VARIABLE
- deaths_by_bldg: **deaths**

STEPS
1. Load deaths_by_bldg
2. Right-click on map: Choose **Current Map Types - Natural Breaks** (3 categories) - Select **deaths**
3. Choose basemap (Carto Dark)
4. Add layer to map: pumps
5. Right click on pumps, change fill + outline color to red and point radius to 3
6. Click on pumps and place on top of deaths_nd_by_house

1. Variable selection



4. Add layer to map

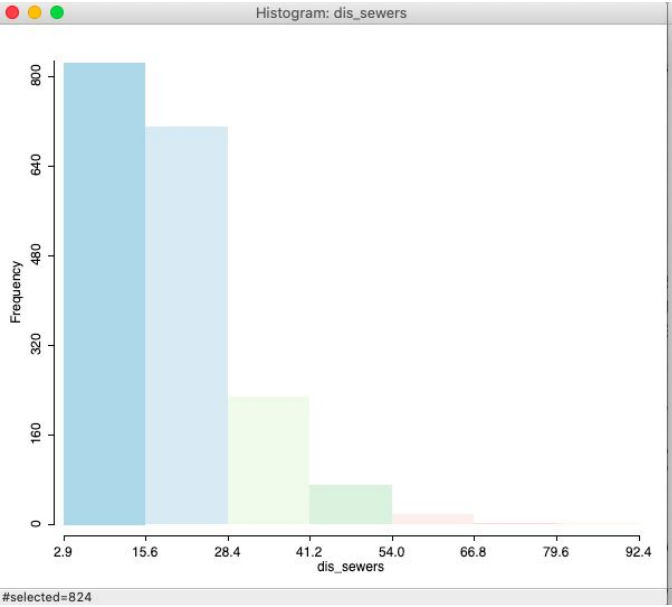


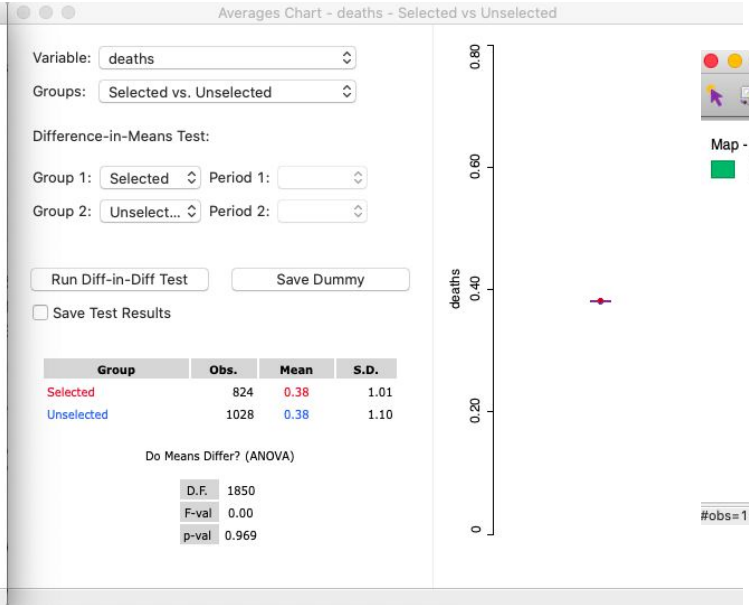
COMPARING AVERAGES ACROSS GROUPS: AVERAGES CHARTS

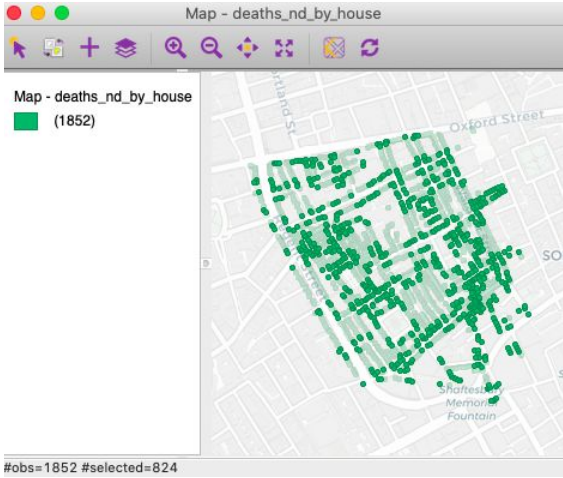
Comparing averages:
Average cholera deaths close to and distant from potential correlates

Resource Links

Let's explore if average deaths were higher near Soho's untrapped sewer grates and ventilators.
Many people thought that gases were polluting the neighborhood through gully holes.







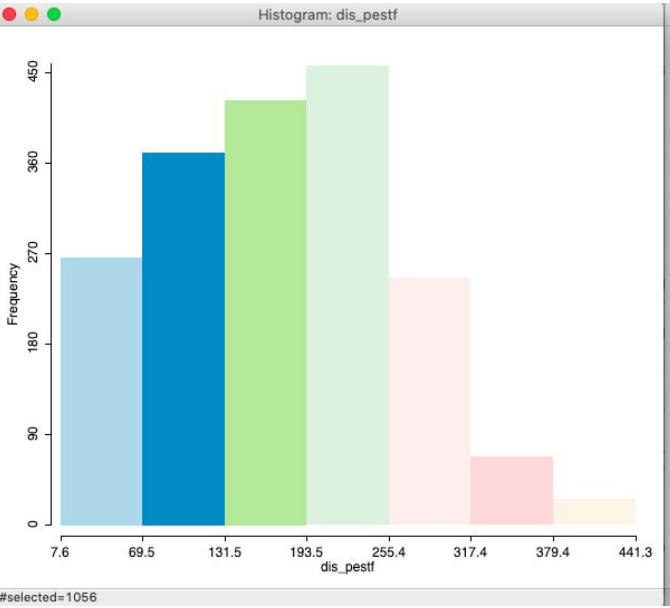
Histogram
with selection

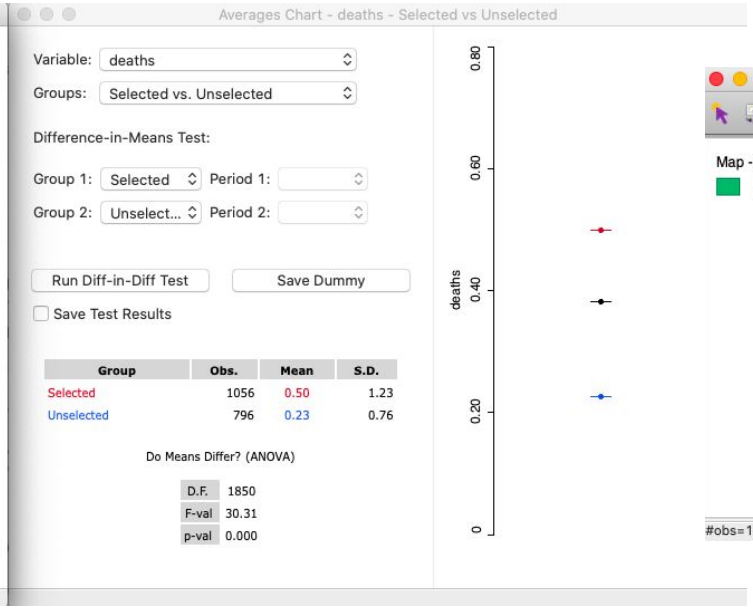
Averages Chart
comparing averages for selected vs unselected

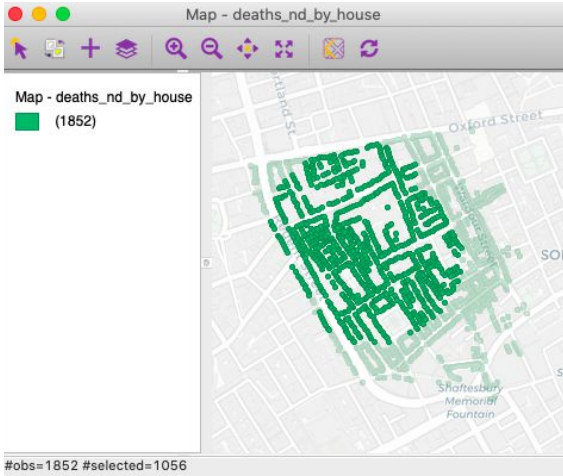
Themeless Map
selected vs unselected

GeoDa's Averages Chart allows us to compare averages for different groups. Here, we make a selection on a histogram to create two groups: houses closest to the nearest sewer grate or ventilator (selected) and those further away (unselected), as shown in the map. This comparison yields insignificant results. The average death count in the Averages Chart is exactly the same for those who were closer to and farther from a sewer grate.

Next, we explore if deaths are higher in houses closer to the subset of sewers in the old pest field. We again make a selection on a histogram to create two groups: houses within 193.5 meters of the old pest field (selected) and those further away (unselected).







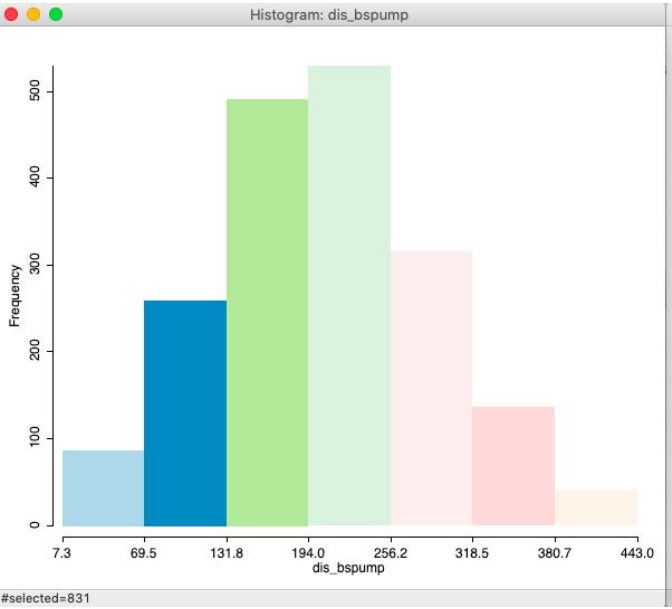
Histogram
with selection

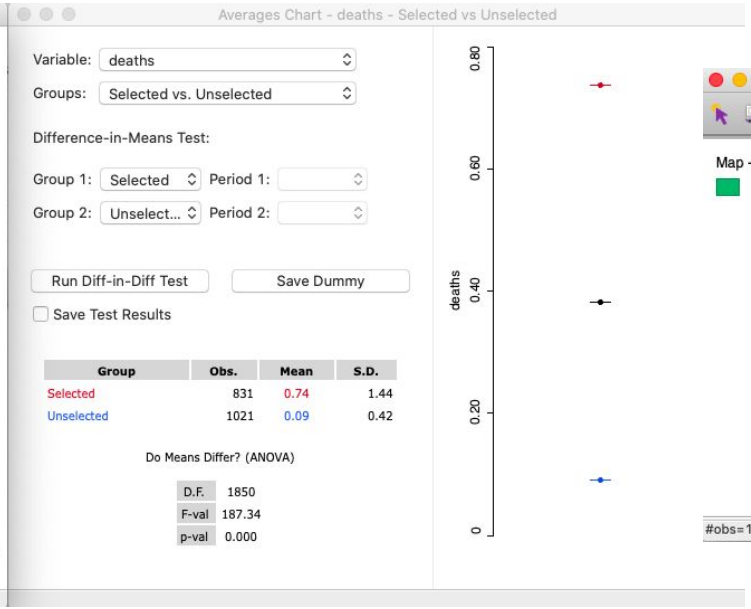
Averages Chart
comparing averages for selected vs unselected

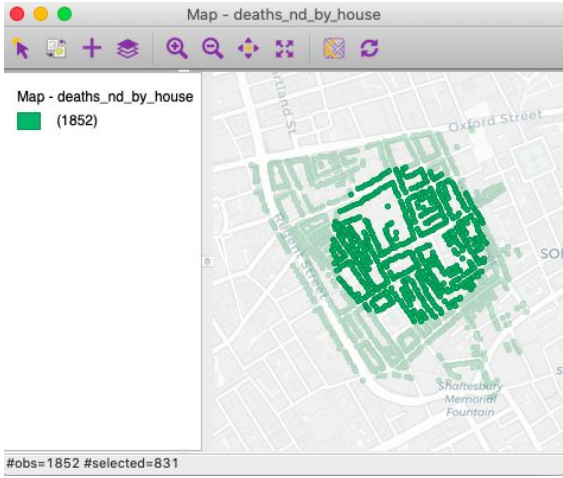
Themeless Map
selected vs unselected

This time, the Averages Chart shows a higher average cholera death count for selected observations than those farther away. This result is statistically significant. However, as we will see in the next example, it turns out to be driven by an alternative explanation to the pest field with an overlapping spatial pattern.

John Snow sought to demonstrate an alternative theory -- that choleraic water was the mode of transmission of cholera. If we measure distances from houses to the Broad Street pump, from which most people in the neighborhood got their water, we find that houses that were closer to the pump showed significantly higher death counts.







Histogram
with selection

Averages Chart
comparing averages for selected vs unselected

Themeless Map
selected vs unselected

By contrast, houses that were farther from the pump showed significantly lower death counts (note that we're using the same scale for all three examples).

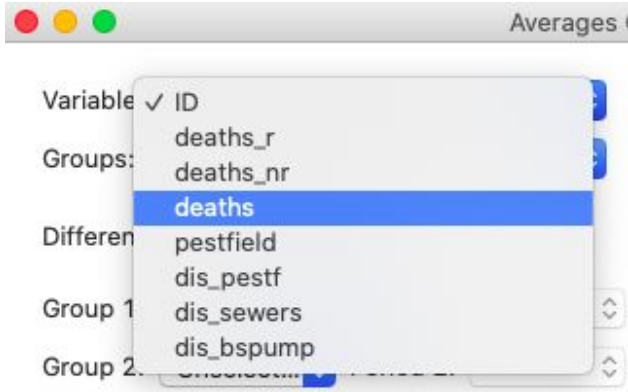
GeoDa Implementation

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

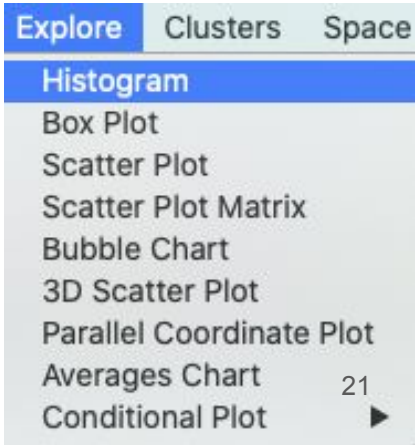
VARIABLES
- deaths_nd_by_house: **deaths, dis_pestf, dis_sewers, dis_bspump**

STEPS
1. **Explore-Averages Chart**-deaths
2. **Explore-Histogram-**dis_pestf
   a. Select the three bars to the left simultaneously (hold shift to add more)
3. **Explore-Histogram-**dis_sewers
   a. Select the bar to the far left
4. **Explore-Histogram-**dis_bspump
   a. Select the three bars to the left simultaneously (hold shift to add more)

1. Selecting the Averages Chart variable



2. Creating a HIstogram



COMPARING DISTRIBUTIONS ACROSS GROUPS: CONDITIONAL BOXPLOTS

Comparing distributions:
Cholera death distributions close to and distant from potential correlates

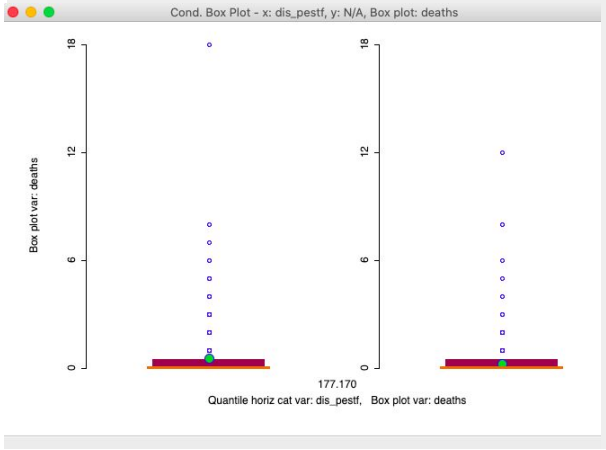
Resource Links

Next, we'll extend the comparison of average deaths to that of death distributions with conditional box plots. We'll compare two groups for each of the three cases: Those below and above the median distance to 1) the old pest field, 2) the nearest sewer grate, and 3) the Broad Street pump.

Similar cholera death counts for houses ...

closer to... ... and farther from

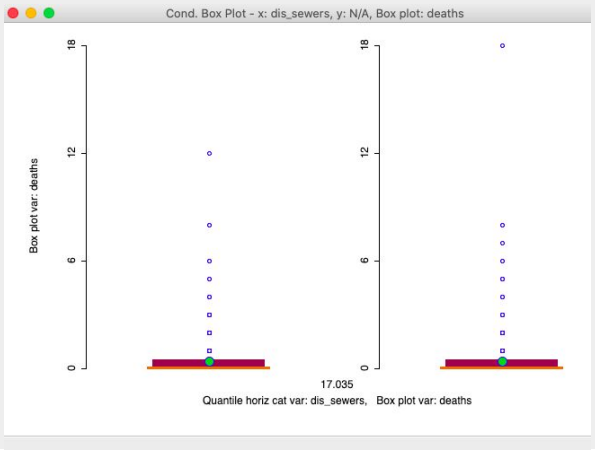
the pest field



Median distance from pest field = 177m
below median above median

closer to... ... and farther from

the sewer grates

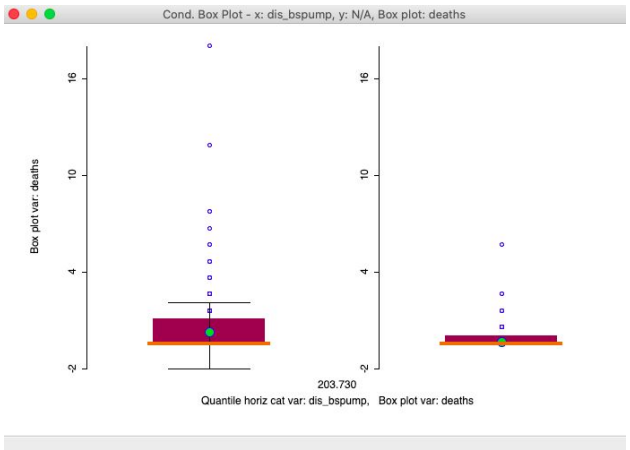


Median distance from sewers = 17m
below median above median

Higher cholera death counts ...

closer to ... ... and farther away from

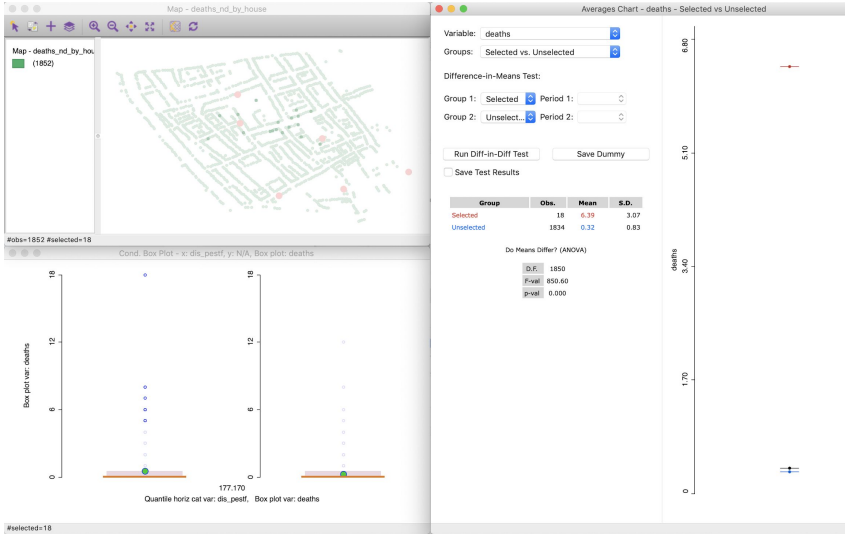
Broad St Pump



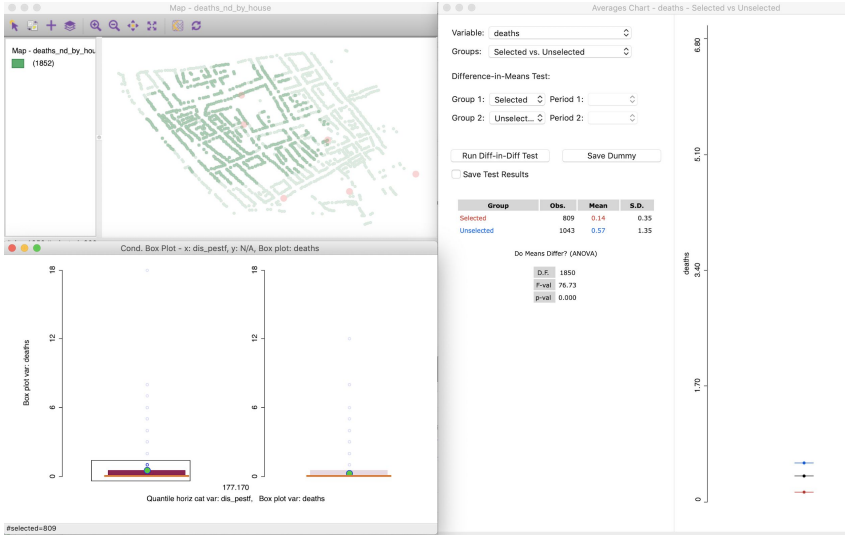
Median distance from Broad St pump = 203m
below median above median

The Averages Chart showed a significant difference between average deaths close to and farther away from the pest field. However, when you explore the full distribution of deaths, it turns out that this result is driven by the largest death counts that are not only close to the pest field but also to Broad St pump.

When you select the 5 largest death counts in the conditional plot, **average deaths for this selection** are **significantly higher** than those for **unselected houses**.



When you select houses with low death counts in the conditional plot, **average deaths for this selection** are **significantly lower** than those for **unselected houses**.



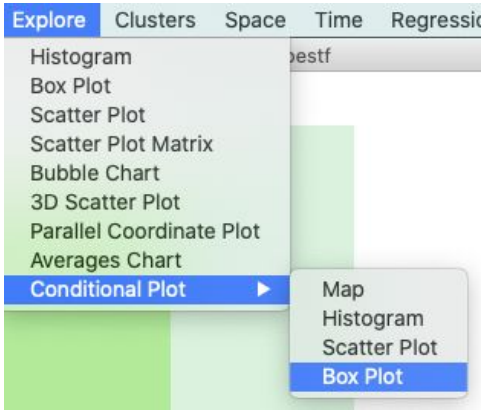
GeoDa Implementation

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

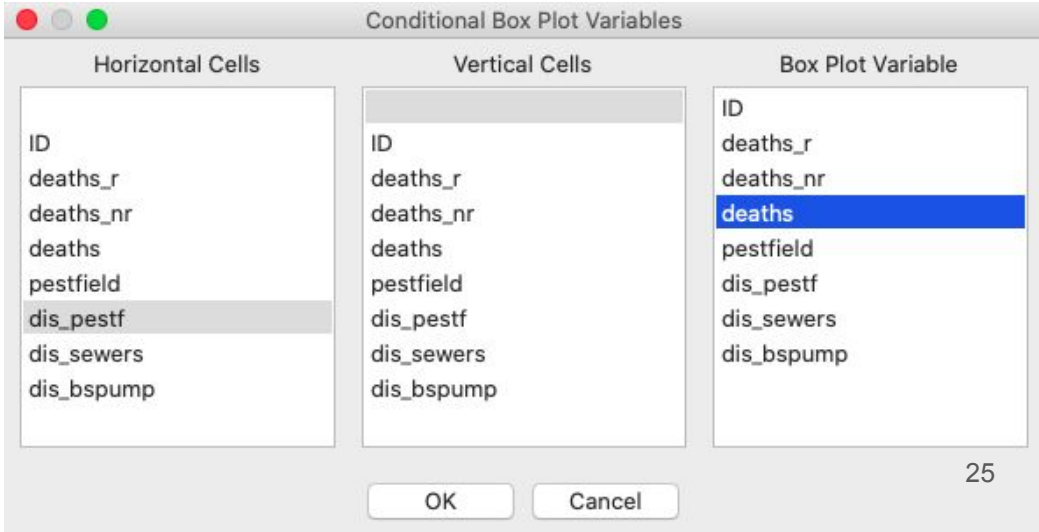
VARIABLES
- deaths_nd_by_house: **deaths, dis_pestf, dis_sewers, dis_bspump**

STEPS
1. **Explore-Conditional Plot-Box Plot**
2. Select dis_pestf for horizontal cells, leave vertical cells blank, and select deaths as box plot variable
3. Repeat step 1
4. Select dis_sewersfor horizontal cells, leave vertical cells blank, and select deaths as box plot variable
5. Repeat step 1
6. Select dis_bspump for horizontal cells, leave vertical cells blank, and select deaths as box plot variable

1. Creating a Conditional Box Plot



2. Selecting Conditional Box Plot variables



IDENTIFYING CLUSTERS AND SPATIAL CONCENTRATIONS

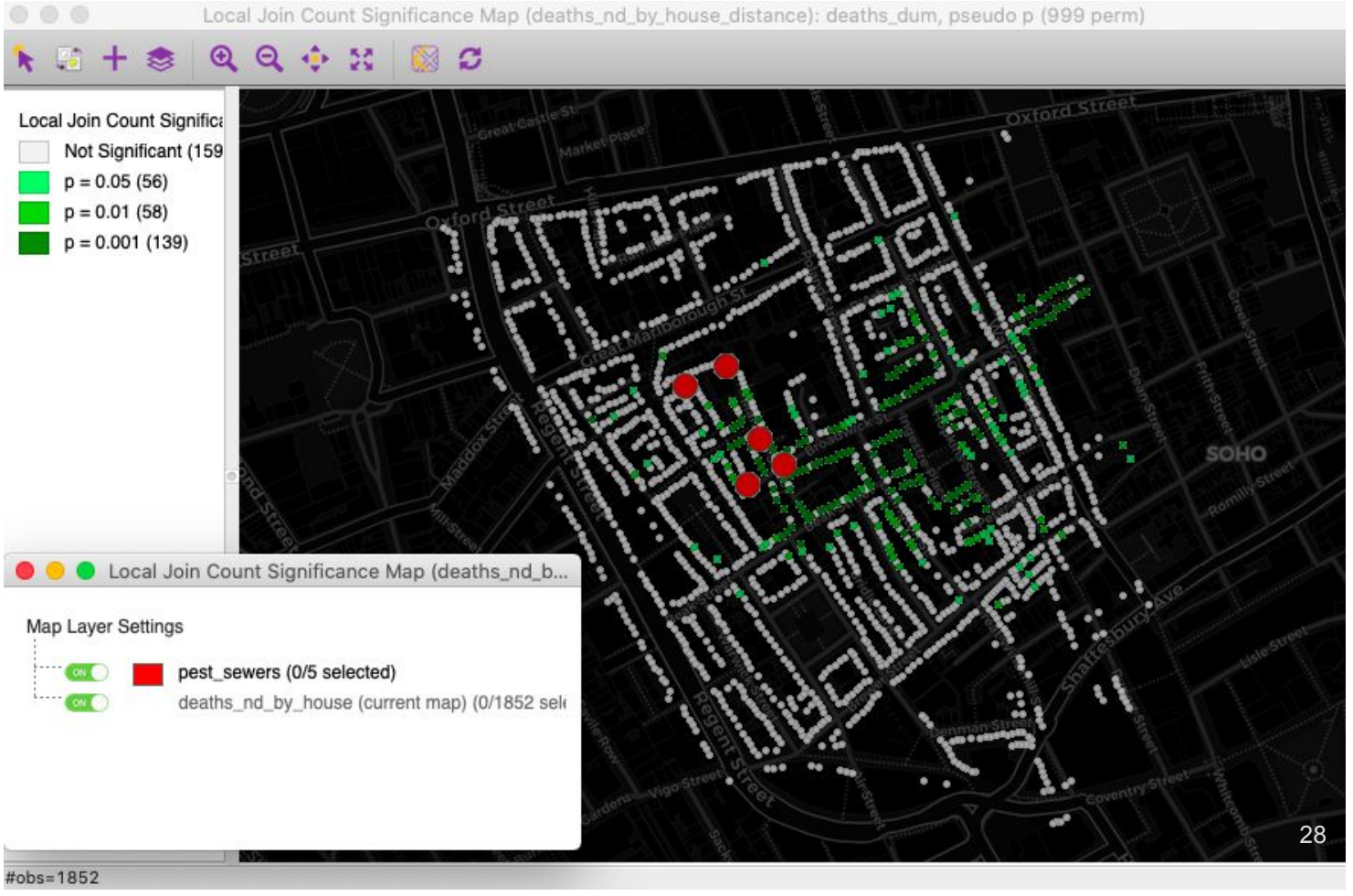
IDENTIFYING SPATIAL CLUSTERS WITH THE UNIVARIATE LOCAL JOIN COUNT STATISTIC

Identifying spatial clusters:
Distinguish the location of cholera deaths and non-deaths

Resource Links

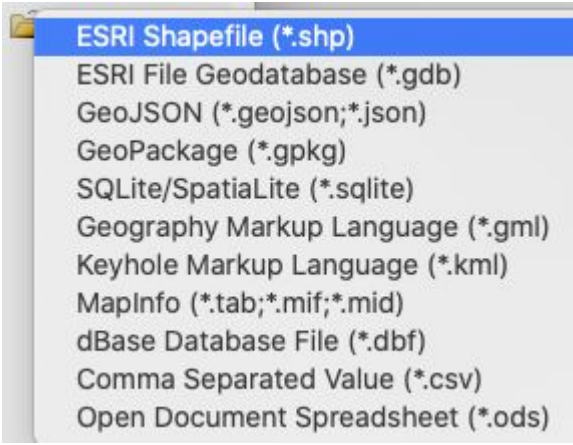
Were cholera deaths concentrated around a supposedly toxic 17th-century pest field?

For this example, we will create a binary variable with the location of cholera deaths, calculate the Local Join Count Statistic to identify concentrations, and then overlay the location of different potential drivers of spatial clusters of cholera deaths.



GeoDa Implementation (1/2)

4. Exporting a new shapefile



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

- deaths_nd_by_house
- sewergrates_ventilators

VARIABLES

- deaths_nd_by_house: **deaths, pestfield,**

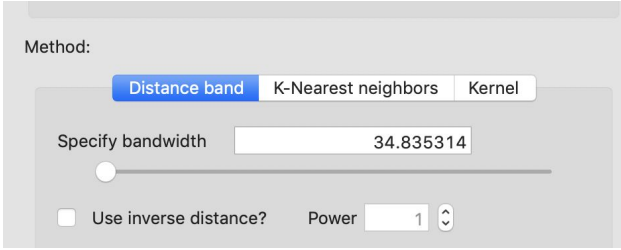
STEPS

Create new shapefile: pest_sewers.shp:

1. **Load** sewergrates_ventilators
2. **Table**-Sort ‘pestfield’ highest to lowest
3. Select observations equal to 1: **File-Save Selected As** (this creates a subset of the 5 sewer grates or ventilators that are on the grounds of the 17th-century pest field)
4. File Path - .shp - Select destination - **Save** as ‘pest_sewers.shp’ - **OK**

GeoDa Implementation (2/2)

12. Distance Weights



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

- deaths_nd_by_house
- pest_sewers

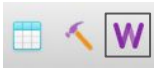
VARIABLES

- deaths_nd_by_house: **deaths_dum**

STEPS

Create distance band weights for deaths_nd_by_house:

10. Create Weights



9. **Load** deaths_nd_by_house
10. **Tools-Weights Manager-Create**
11. Select ID variable (ID)
12. **Distance Weight-Distance Band**: Leave rest of the settings- **Create**

Create Univariate Local Join Count Map

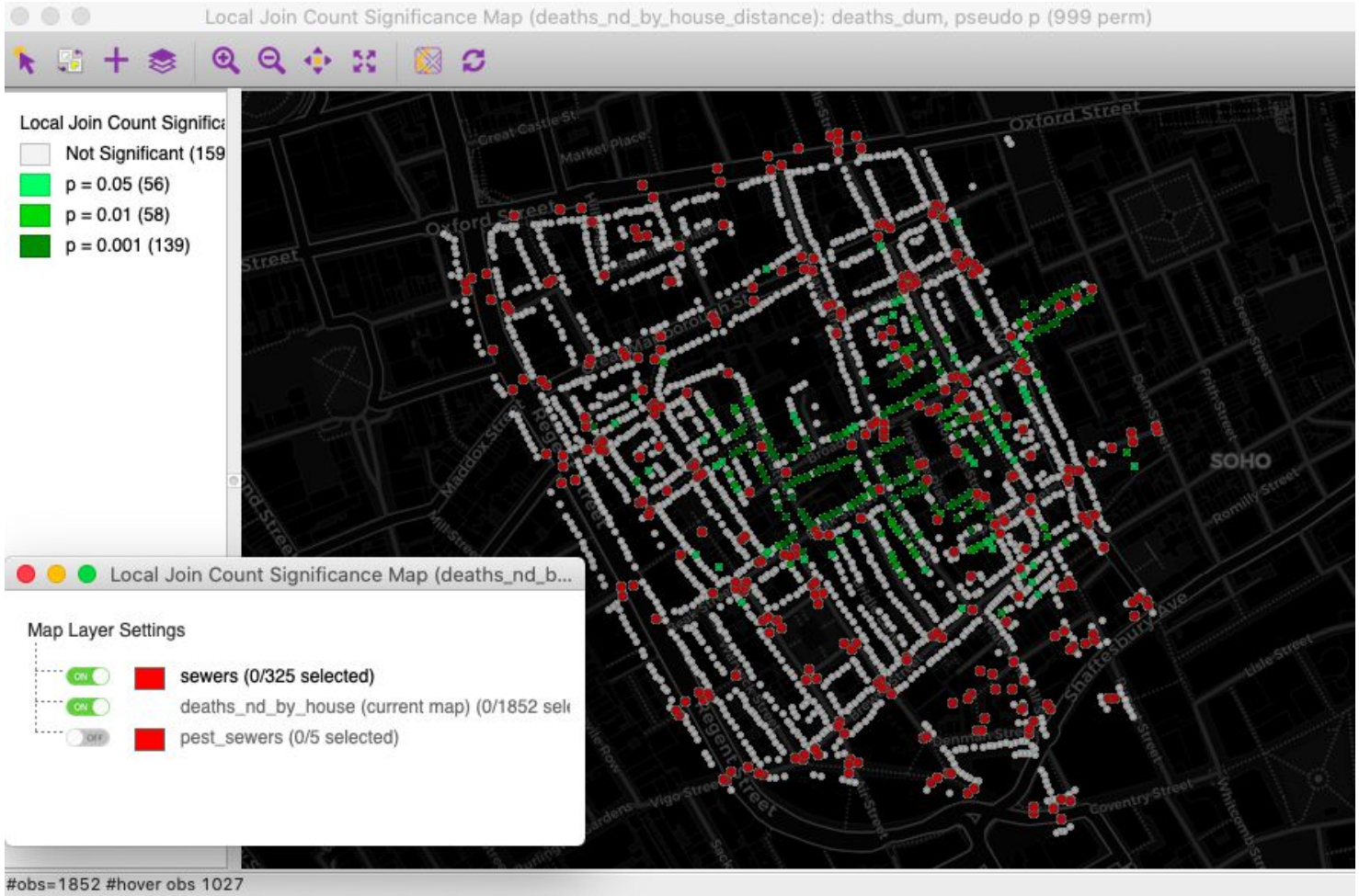
13. **Space-Univariate Local Join Count**
14. Select variable (“deaths_dum”) - **OK**

Add layers

15. **Add basemap** (Carto Dark)
16. **Add layer to map**: pest_sewers
17. Right-click on pest_sewers-Change fill color of pumps to red, change point radius to 5
18. Click on pest_sewers and place on top of deaths_nd_by_house

13. Univariate Local Joint Count

What was the relationship between all of Soho's sewer grates and deaths?



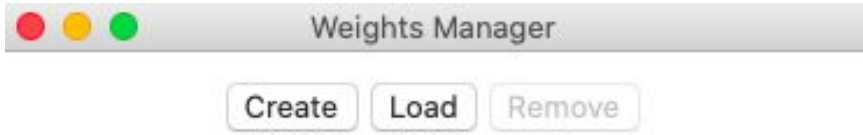
GeoDa Implementation

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

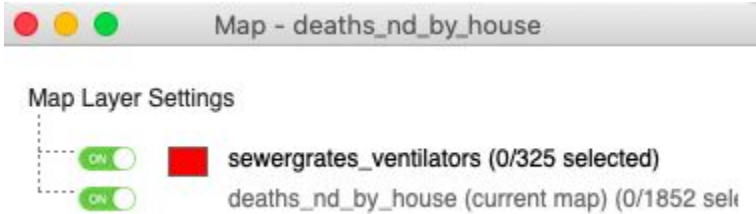
VARIABLES
- deaths_nd_by_house: **deaths**

STEPS
1. Load deaths_nd_by_house
2. Load weights (.gwt file), otherwise repeat steps 9-12 from previous demo
3. **Space-Univariate Local Join Count**
4. Select variable (“deaths_dum”), then **OK**
5. Add basemap (Carto Dark)
6. Add layer to map: sewergrates_ventilators
7. Right click on sewergrates_ventilators-Change fill color of pumps to red, change point radius to 3
8. Click on sewergrates_ventilators and place on top of deaths_nd_by_house

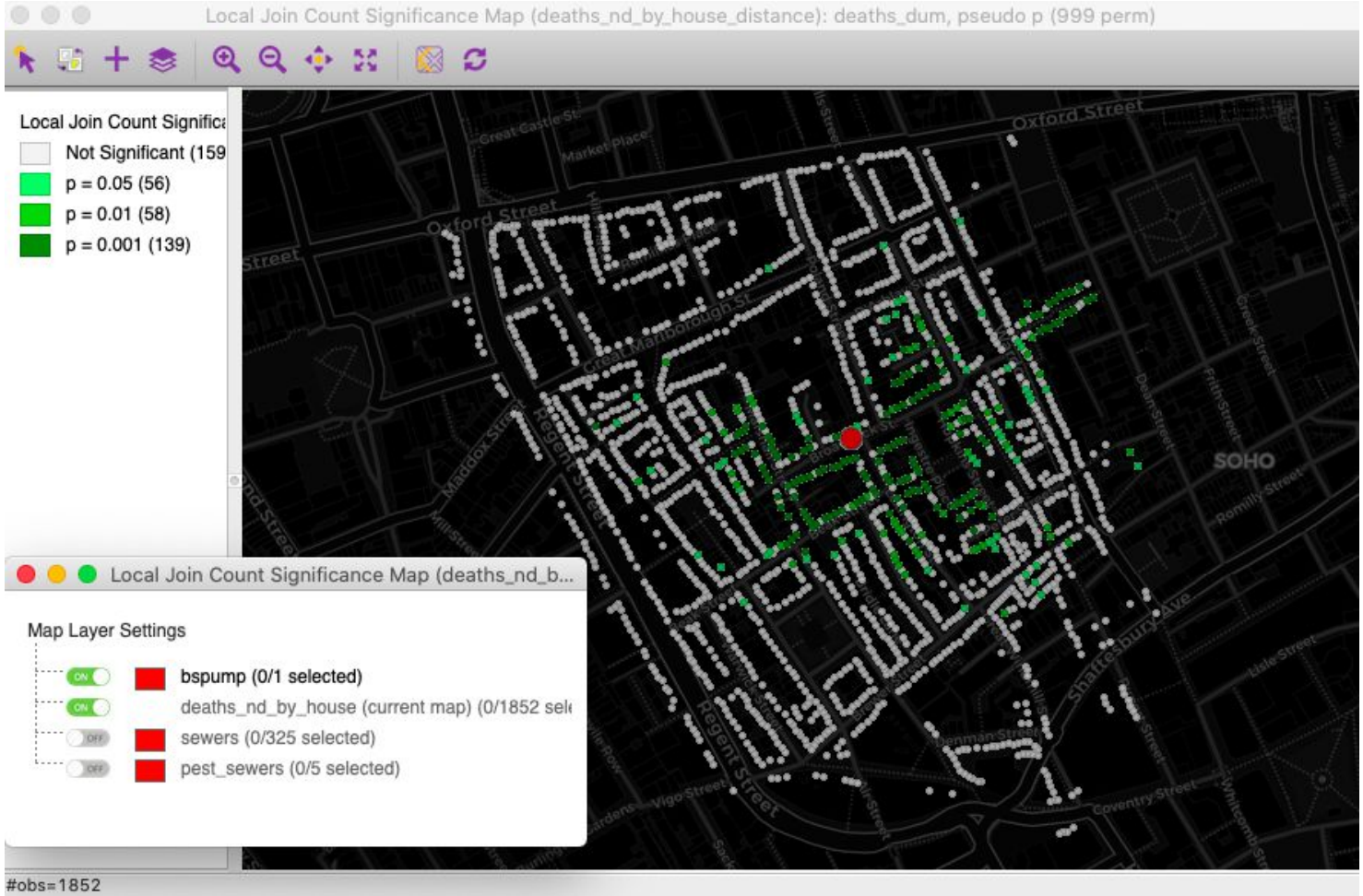
2. Load Weights



8. Place sewergrates_ventilators on top of deaths_nd_by_house



Were deaths concentrated around the Broad Street pump?



GeoDa Implementation

2. Select Broad Street pump

Table - pumps

	ID	x	y	name
1	1	529396.539395	181025.063047	Broad St Pump
2	2	529192.537868	181079.391380	Great Malborough Pump
3	3	529183.739766	181193.735013	Ramilies Place Pump
4	5	529613.205238	180896.804121	Rupert St Pump
5	6	529453.585995	180826.353152	Brewer St Pump
6	8	529296.104419	180794.849037	Warwick St Pump

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

- deaths_nd_by_house
- pumps

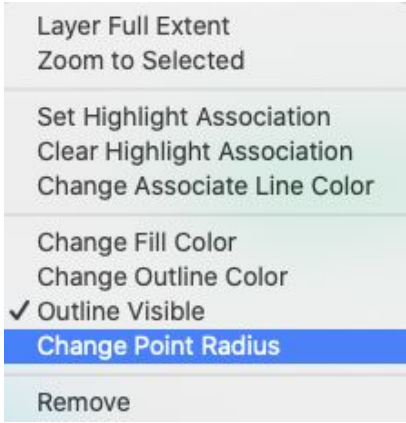
VARIABLES

- deaths_nd_by_house: **deaths**

STEPS

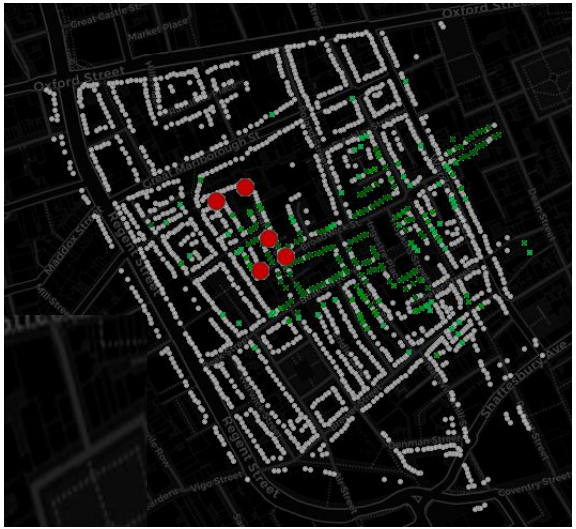
1. Load pumps
2. **Table**-Select 'Broad Street pump'-**File-Save Selected As** (this creates a dataset only with Broad Street pump)
3. File Path - .shp - Select destination - **Save** as 'bspump.shp' - **OK**
4. Load deaths_nd_by_house
5. Load weights (.gwt file), otherwise repeat steps 9-12 from [the first Local Join Count demo]
6. **Space-Univariate Local Join Count**
7. Select variable ("deaths_dum"), then **OK**
8. Add basemap (Carto Dark)
9. Add layer to map: bspump
10. Right click on bspump-Change fill color of pumps to red, change point radius to 5
11. Click on bspump and place on top of deaths_nd_by_house

10. Change Point Radius



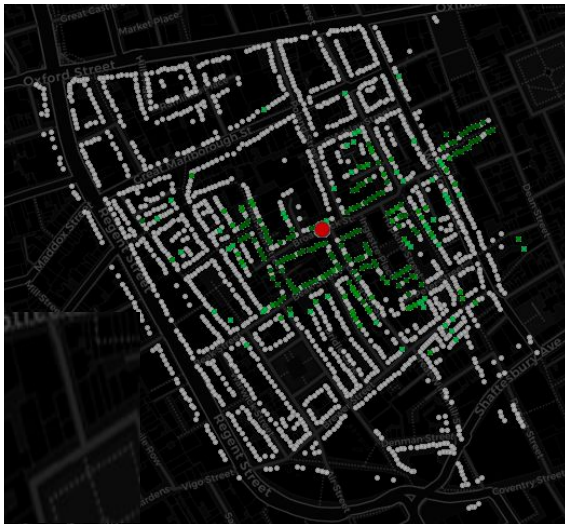
The hypothesized transmission modes of cholera that were analyzed at the time and that you can now explore in GeoDa include toxic gases emanating from 1) a 17th-century pest field, 2) all of Soho's sewer grates and ventilators, and 3) from water of the Broad Street pump. Comparing each of the Local Join Count Statistic's significance map to these potential drivers points towards **water as the main mode of transmission of cholera** (3), as John Snow suspected.

1) Cholera deaths were not concentrated around a supposedly toxic 17th-century pest field



2) Cholera deaths were also not concentrated around sewer grates and ventilators

3) But deaths did seem to be concentrated around the Broad Street pump



Since this was data from the General Board of Health's Cholera Inquiry Committee and given that its members believed in airborne theories of cholera transmission, let's now use data collected by Snow and Henry Whitehead to further investigate the relationship between cholera and water.

EXPLORING THE RELATIONSHIP BETWEEN TWO POINT LAYERS

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

Resource Links

GeoDa Implementation

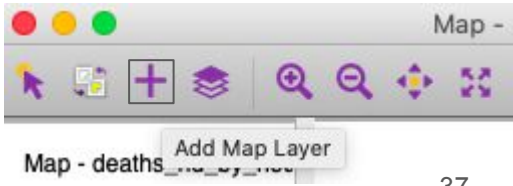
DATA - 2 shapefiles (shp, shx, dbf):
- deaths_by_bldg
- pumps

VARIABLE
- deaths_by_bldg: **deaths**

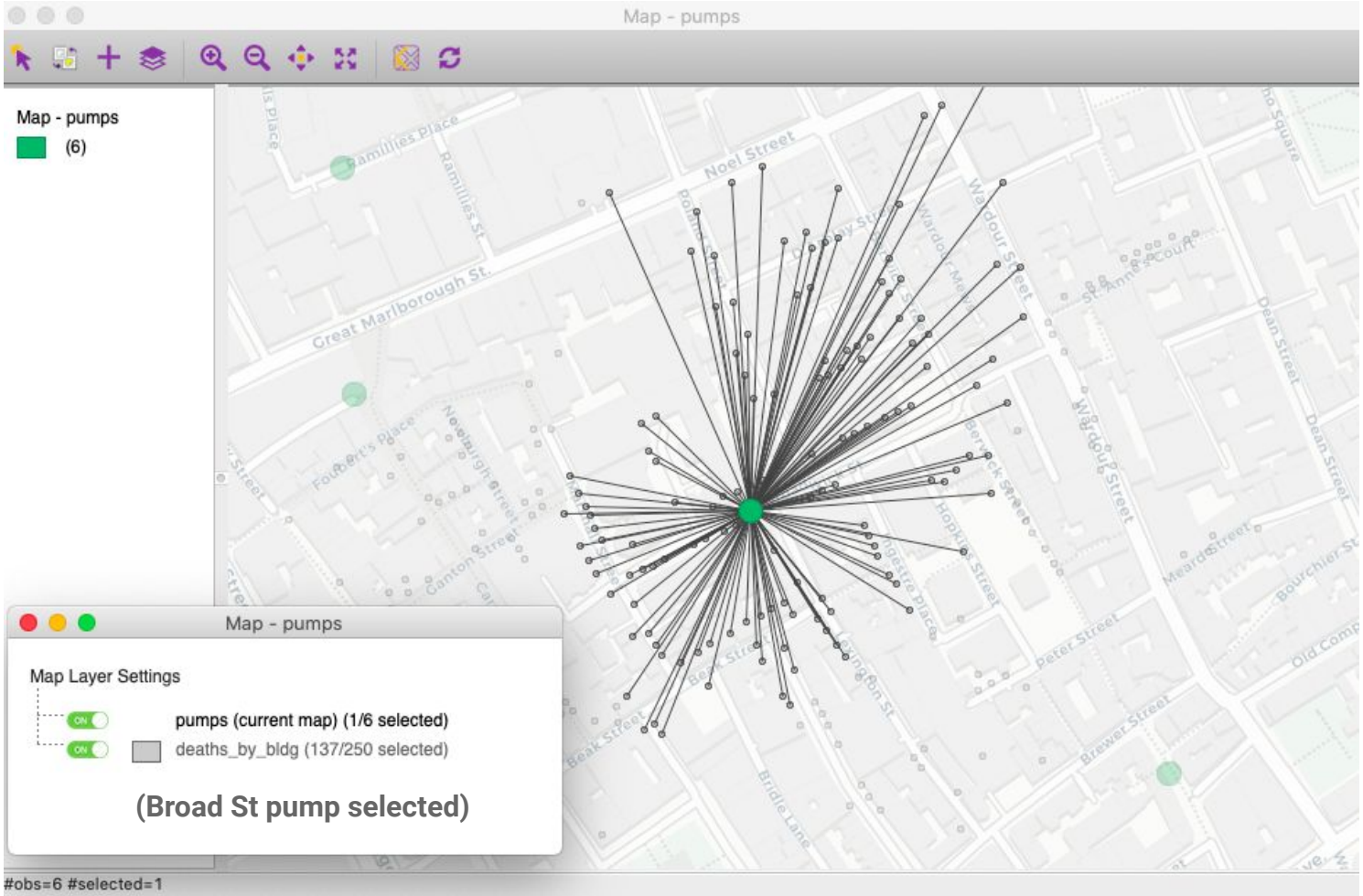
STEPS
1. Load deaths_by_bldg
2. Add layer to map: pumps
3. Right click on pumps-Change fill color of pumps to grey, change point radius to 3
4. Click on pumps and place on top of deaths_nd_by_house

Variable selection

2. Add layer to map



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



GeoDa Implementation

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

- deaths_by_bldg
- pumps

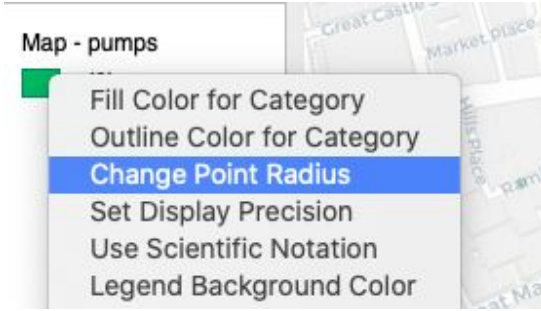
VARIABLES

- pumps: **ID**
- deaths_by_bldg: **pumpID**

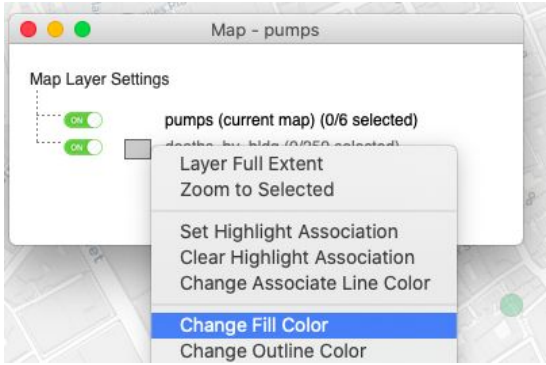
STEPS

1. **Load** pumps
2. **Add basemap** (Carto Light)
3. Change point radius to 8 (right-click on legend (green box))
4. **Add layer to boxmap**: + deaths_by_bldg, then right-click on it:
   a. Change fill color of pumps to black
   b. Set Highlight Association for pumps to link ID of 6 pumps to pumpID of deaths (deaths_by_bldg, pumpID, ID - check on 'show connect line')
5. Linking and brushing: select pump(s)
6. Close map

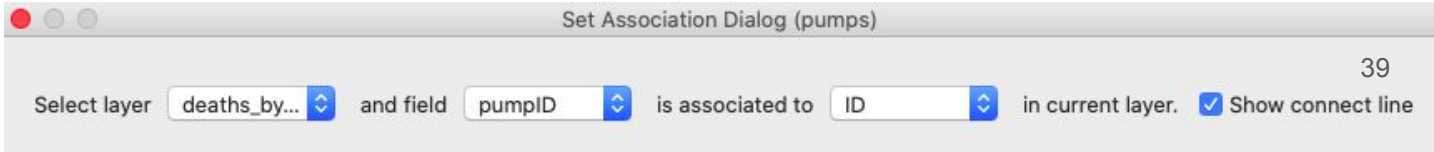
3. Change point radius



4a. Change fill color



4b. Set highlight association

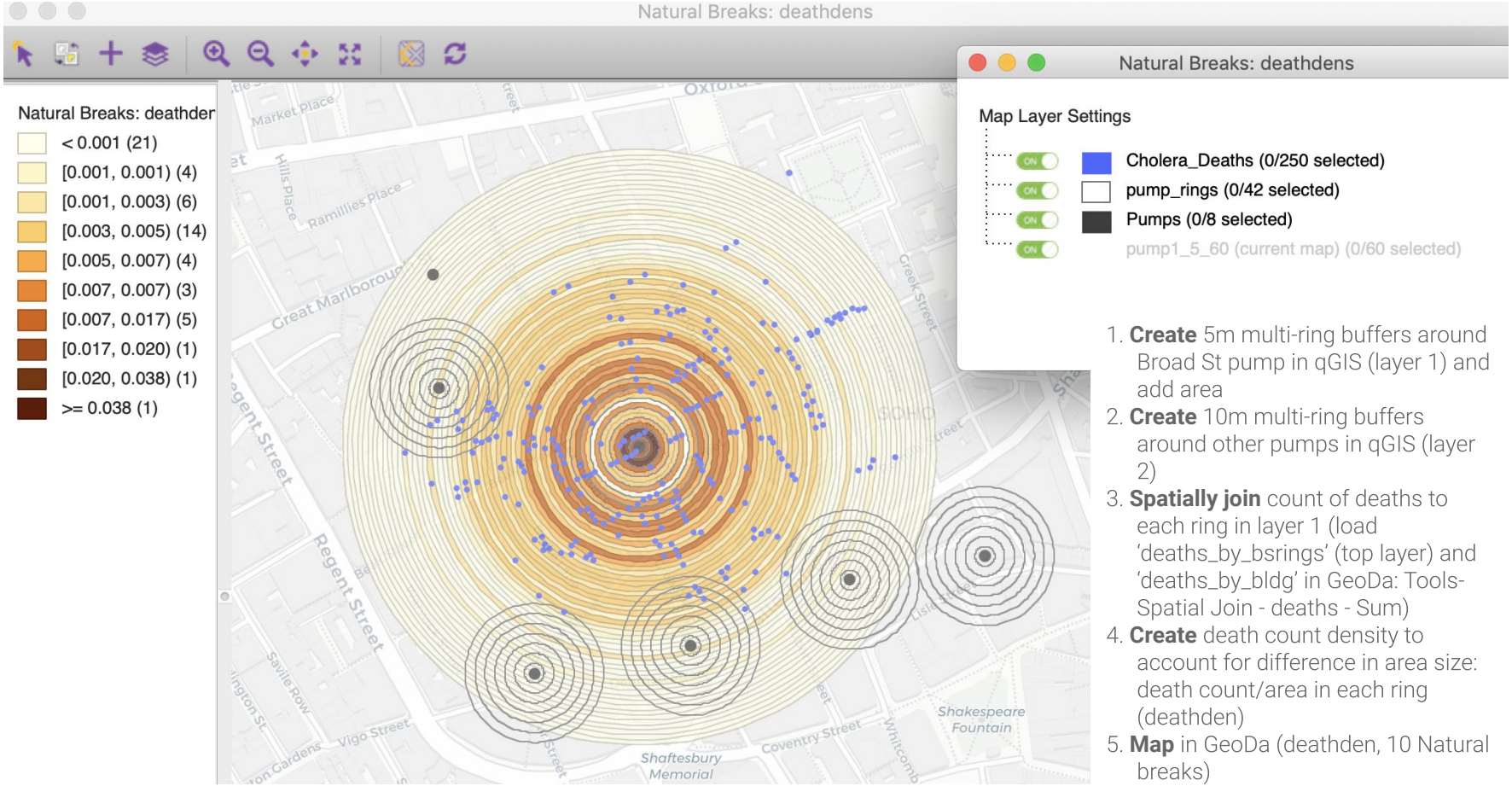


IDENTIFYING DISTANCE DECAY

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

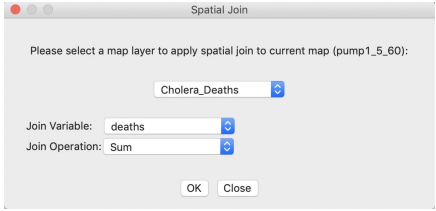
STEPS

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)

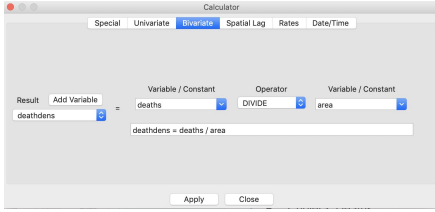
Calculate death density:
7. **Table-Calculator**
8. **Bivariate-Add Variable**: deathden → deaths DIVIDE area (decimals: 6, display 6)
9. **Save** (this adds deaths/area to table)

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



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

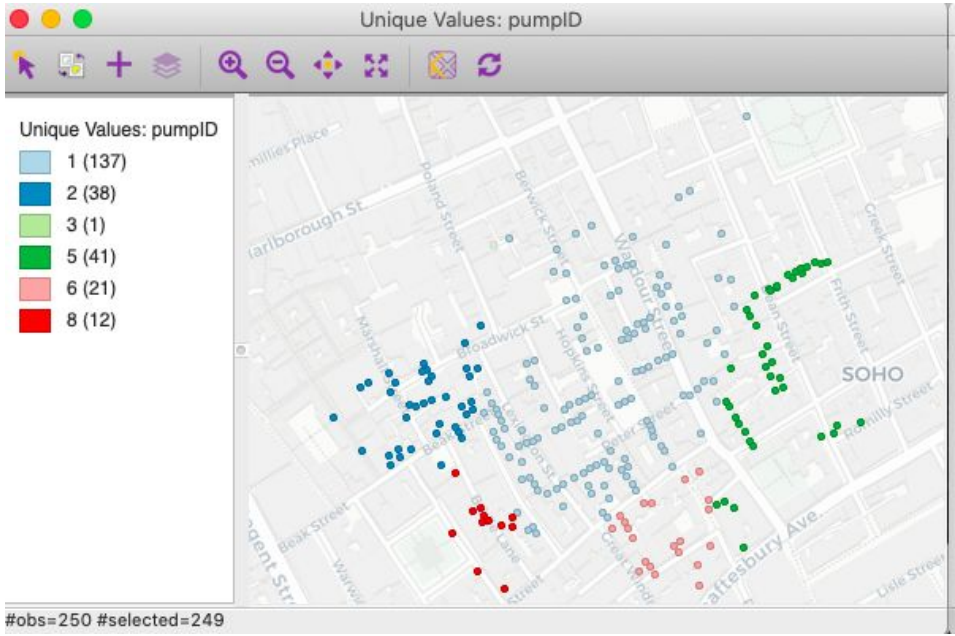
COMPARING DISTRIBUTIONS ACROSS GROUPS

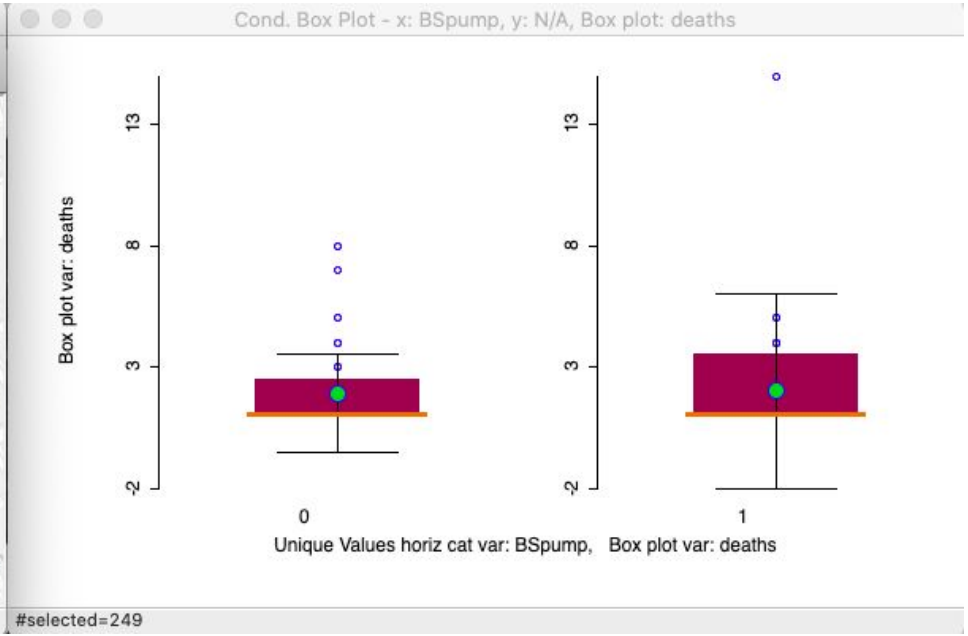
CONDITIONAL BOX PLOTS

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

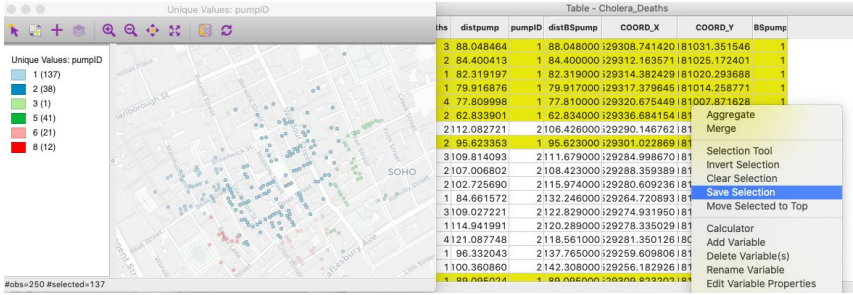
VARIABLES

- deaths
- pumpID

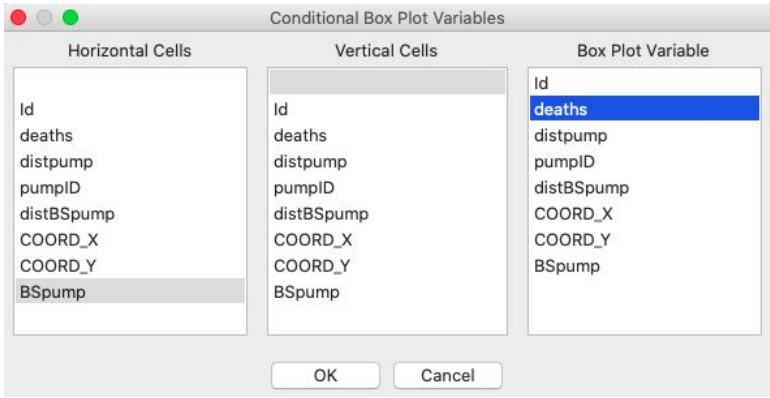
STEPS

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** - Right click and **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

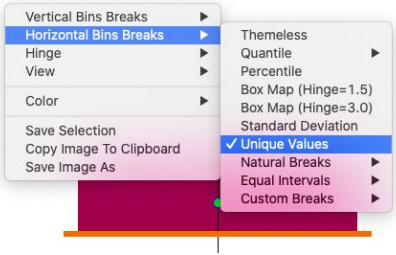
3. Select category 1
4. Right-click to save selection



5. Select variables



5.a. Modify horizontal bin breaks



THE SOUTH LONDON NATURAL EXPERIMENT

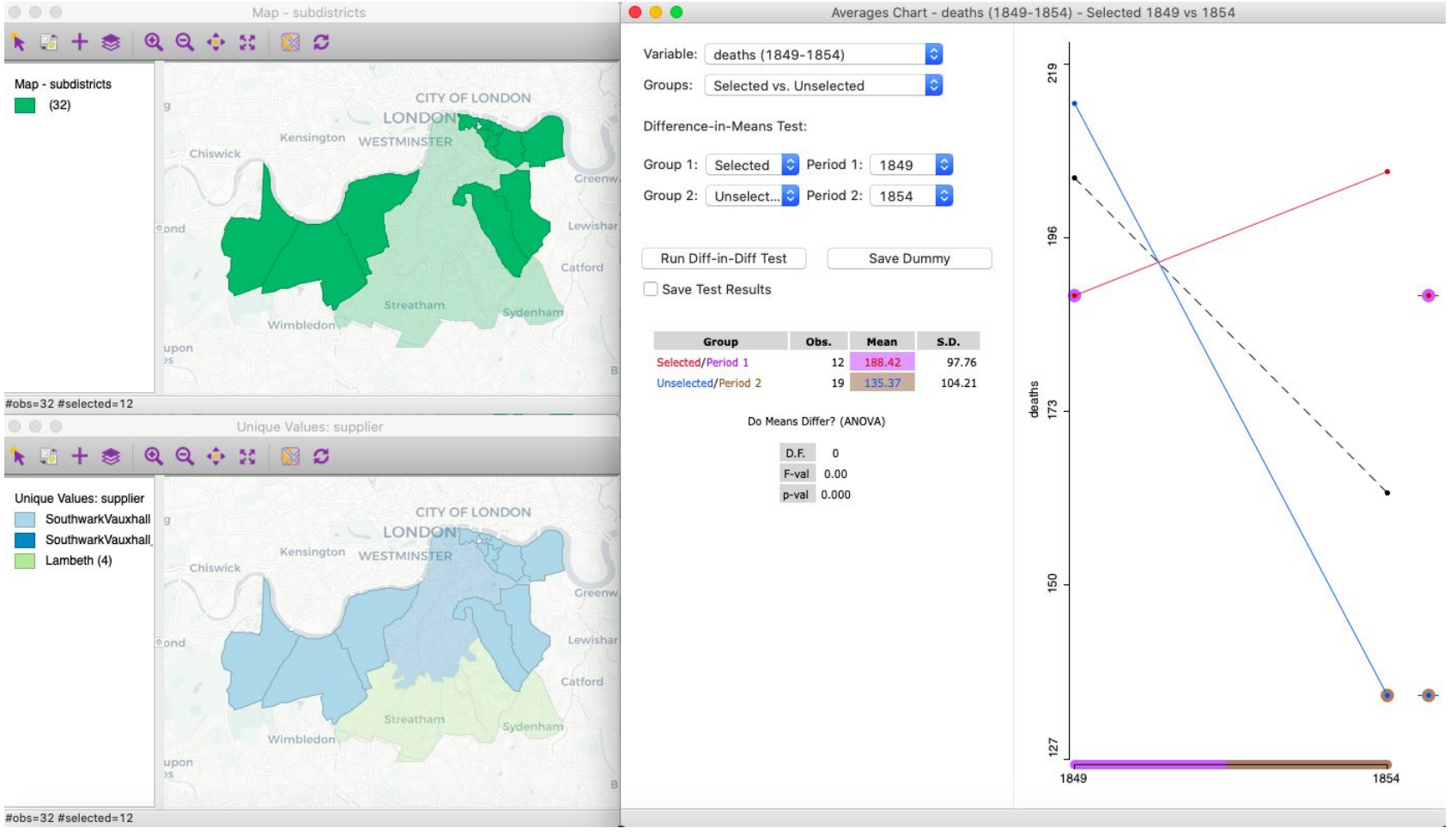
COMPARING TRENDS

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



GeoDa Implementation

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

VARIABLES
- deaths1849
- deaths1854

STEPS

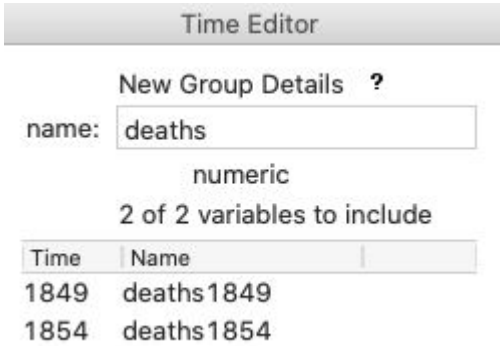
Creating a time variable:

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



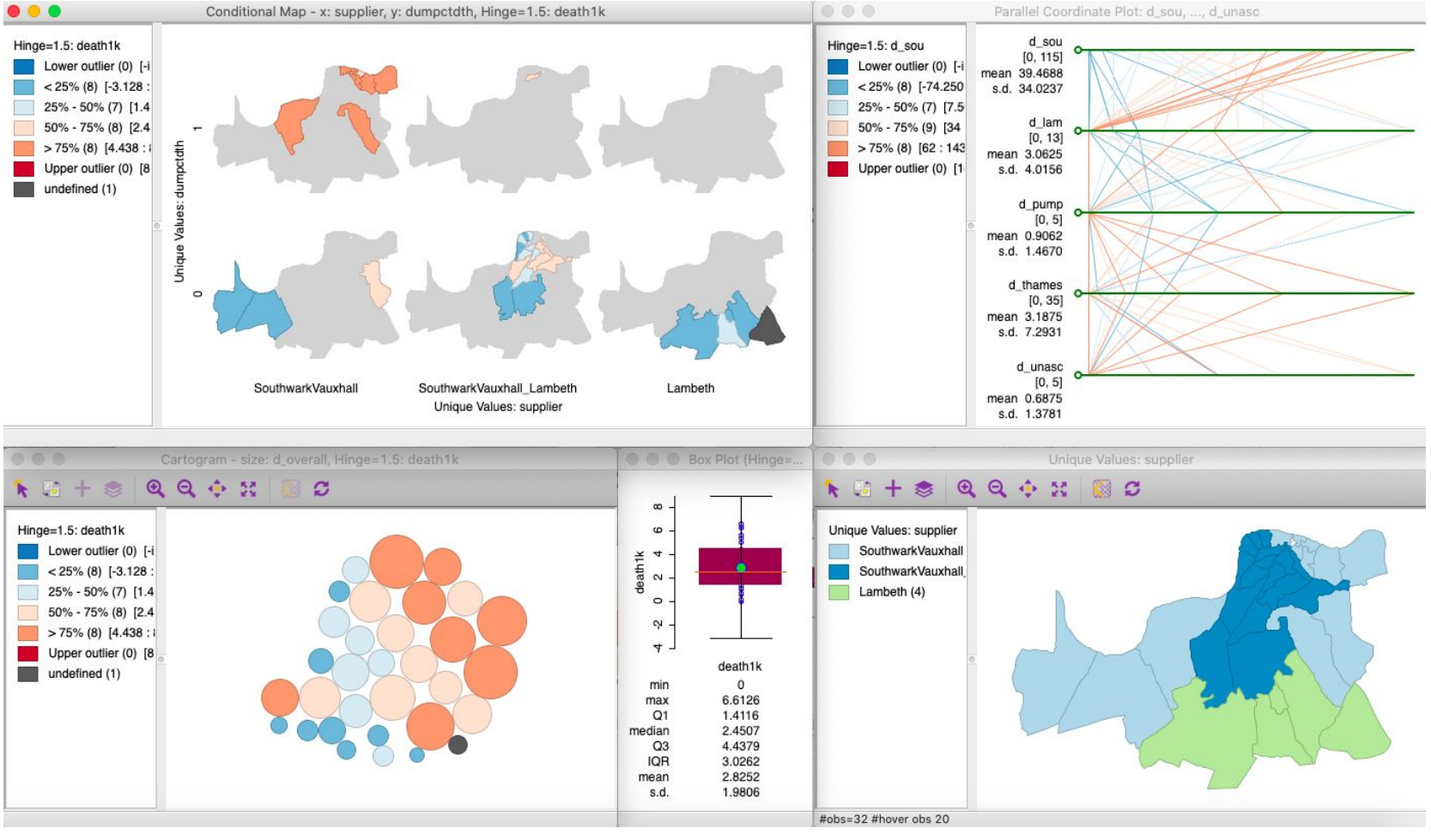
EXPLORING A QUESTION WITH MULTIPLE EDA + ESDA TOOLS

SCATTER PLOTS, BOX PLOTS, PARALLEL COORDINATE PLOTS, CONDITIONAL BOX PLOTS/MAPS, MAPS, AND CARTOGRAMS

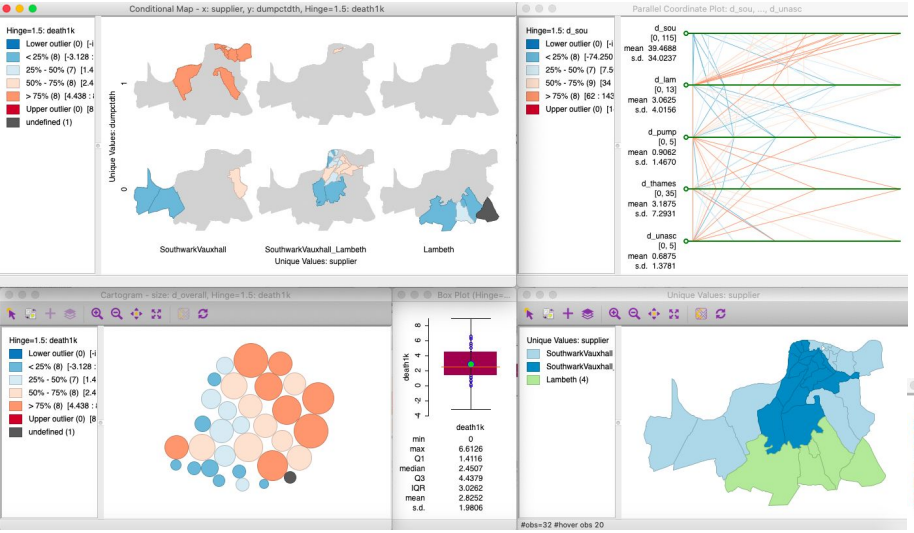
Exploring a question with multiple EDA and ESDA tools:
Explore deaths, causes and water suppliers

Resource Links

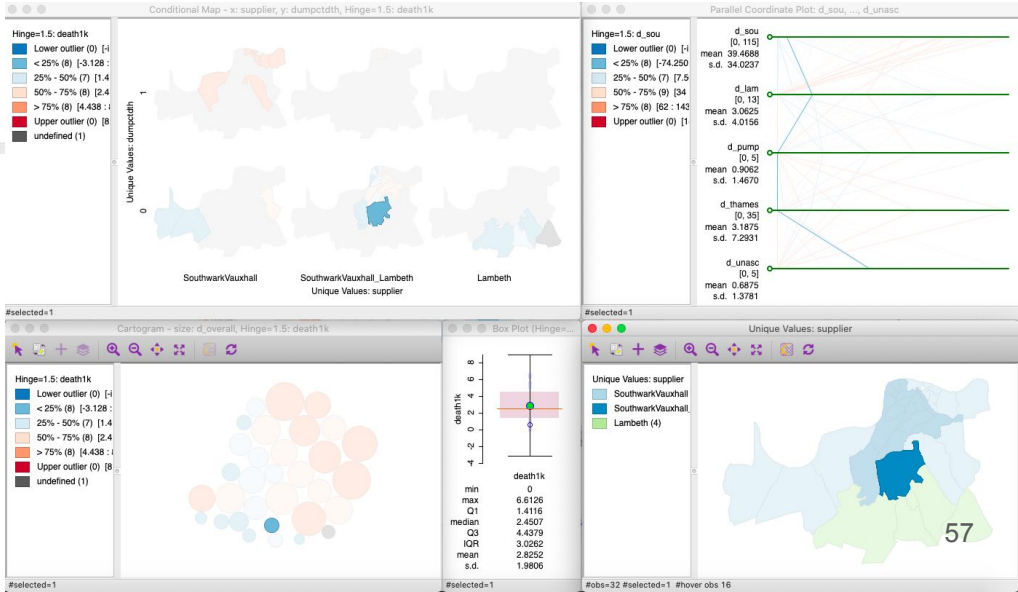
SOUTH LONDON EXP.: ESDA - Multiple Views of Deaths, Death Causes and Water Suppliers



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



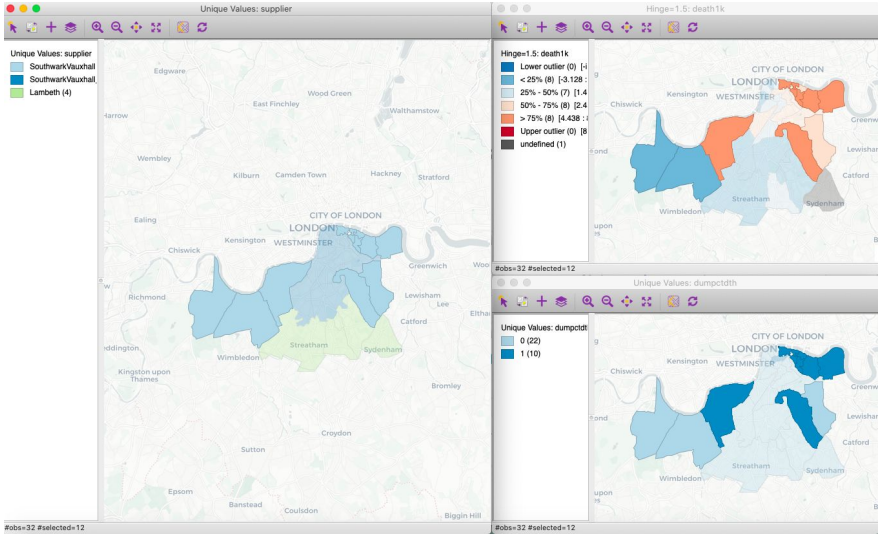
SOUTH LONDON EXPERIMENT

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**



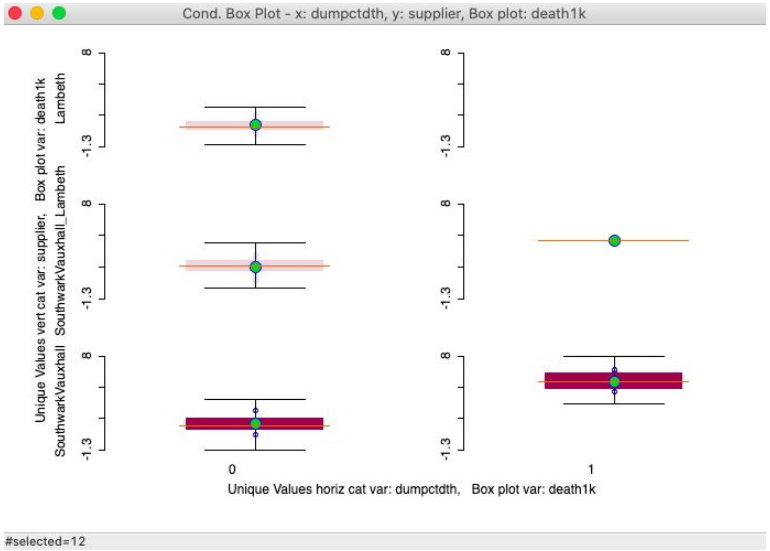
by water **supplier**

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

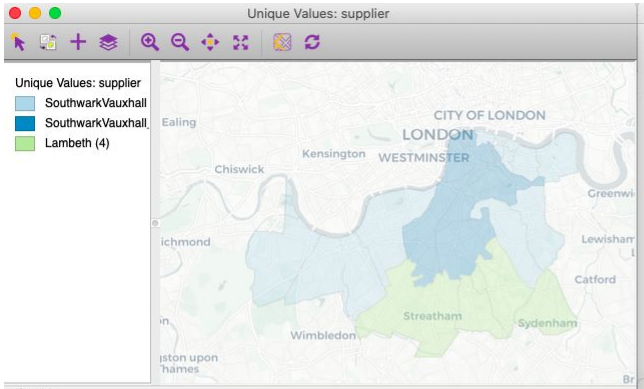


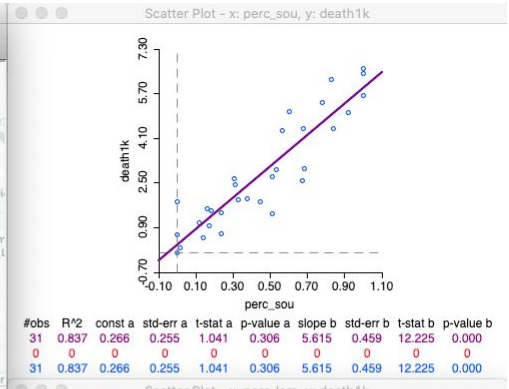
by low-high death1k category
(**dumpctdth**: 0 = 0-3 deaths/1k, 1 = 4-14)

SOUTH LONDON EXPERIMENT
Higher Share of Deaths in Subdistricts Associated with Southwark Water Company

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

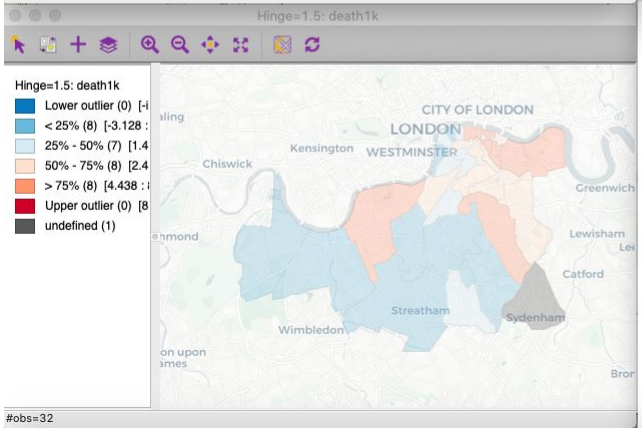
Unique Values Map: Water **supplier**

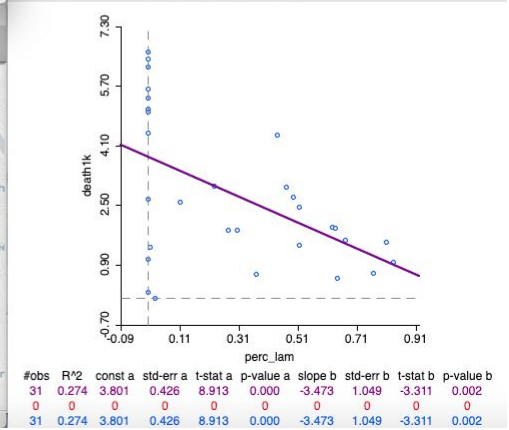




perc_sou: % population served by Southwark & Vauxhall company

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





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: Right click and **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

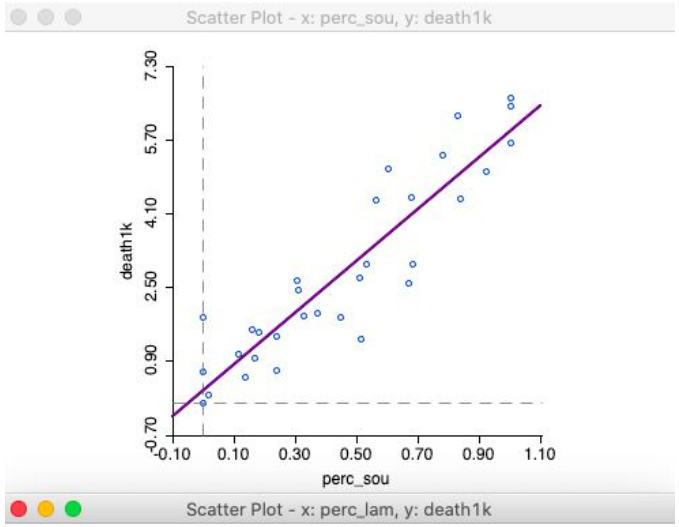
Close conditional boxplot and unique values map (dumpctdth)
Leave other two maps open (death1k and supplier)

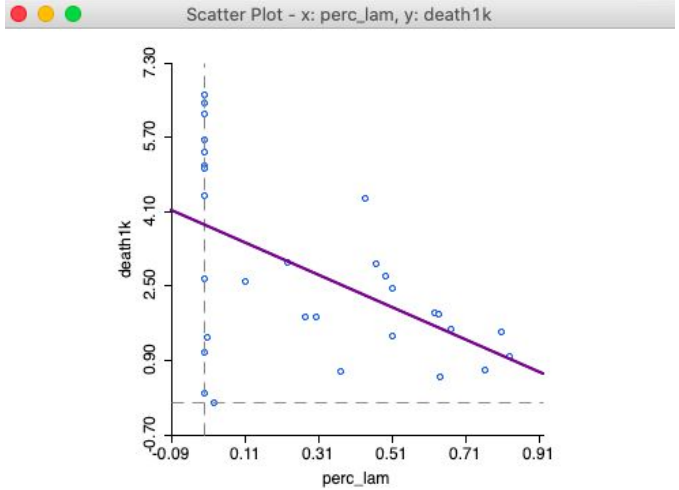
Variables:

- **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

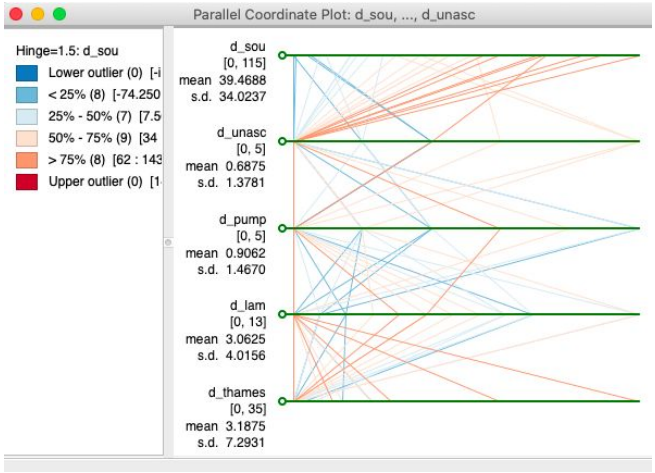
VARIABLES

Deaths attributed to ...

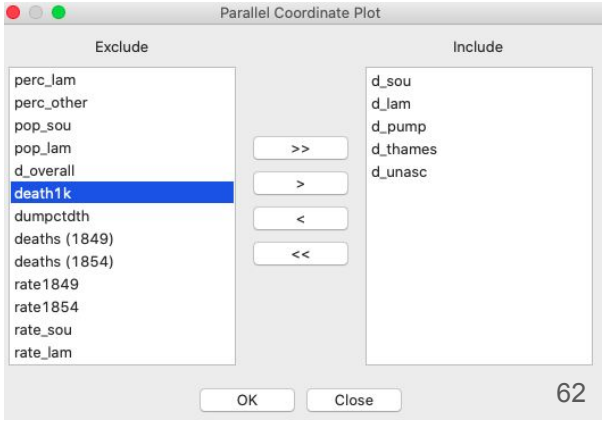
- **d_sou:** … the Southwark company
- **d_lam:** … the Lambeth company
- **d_pump:** … pumps or wells
- **d_thames:** … Thames water
- **d_unasc** … an unknown source

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



SOUTH LONDON EXPERIMENT: Conditional Map and Cartogram

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

- subdistricts

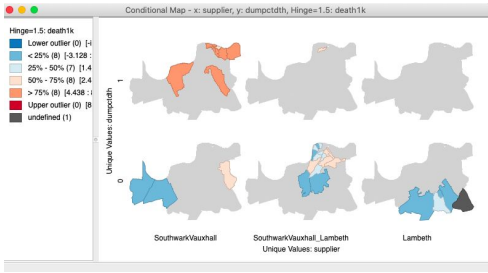
VARIABLES

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

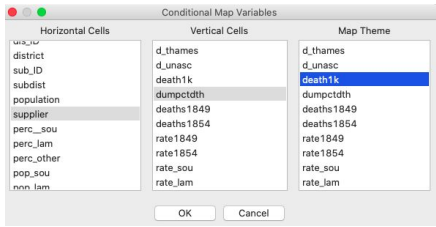
STEPS

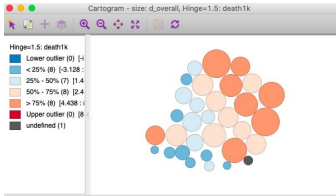
1. **Explore-Conditional Plot-Boxplot** with horizontal = **supplier**, vertical = **dumpctdth**, and map theme = **death1k** (2 rows, 2 columns)
   a. **Right-click: Change vertical bin breaks to unique values** for categorical representation of 0-1

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

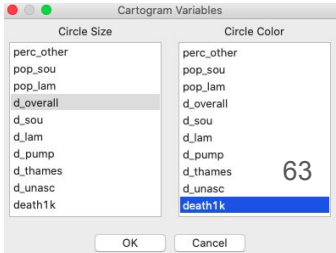


1. Conditional boxmap: variables





2. Cartogram variables



SOUTH LONDON EXPERIMENT: Unique Values Map and Boxplot

1 shapefile (shp, shx, dbf):

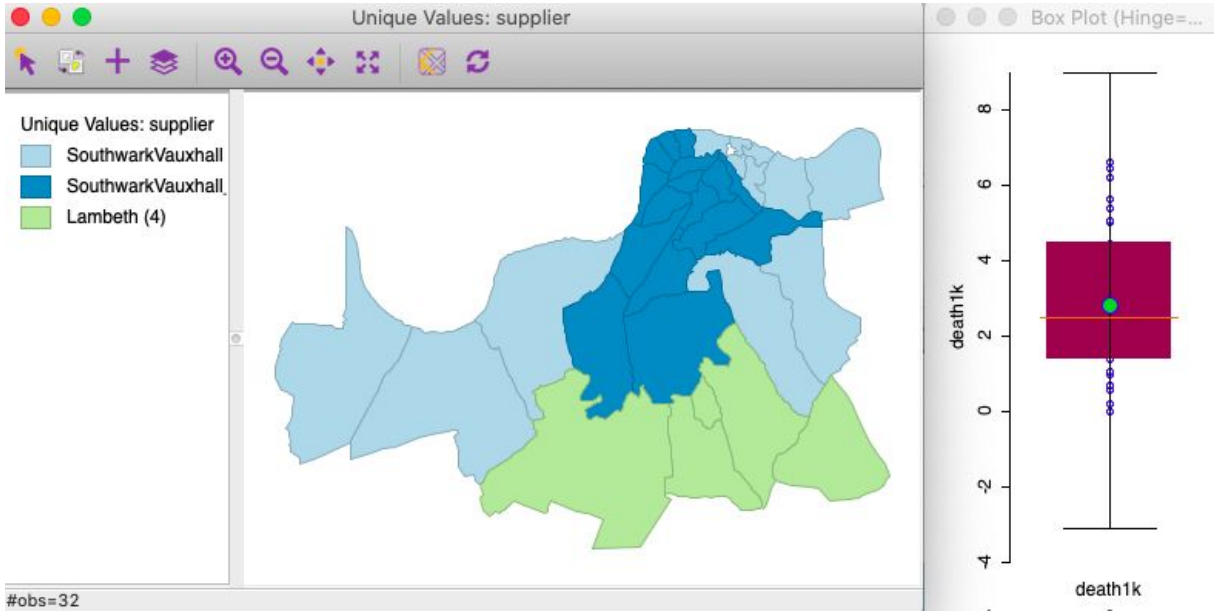
- subdistricts

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).

General Board of Health, Medical Council (1855), Plan shewing the ascertained deaths from cholera in the part of the Parishes of St. James, Westminster and St. Anne, Soho, during the summer and autumn of 1854, in *Appendix to Report of the Committee for Scientific Inquiries in Relation to the Cholera-Epidemic of 1854*, London, HMSO, no. 14, available at http://kora.matrix.msu.edu/files/21/121/15-79-45-30-johnsnow-a0a1b9-a_16430.jpg.

Snow, J. (1855). *On the Mode of Communication of Cholera*, London, second edition, Map 1, available at https://bit.ly/32Az1lW

Snow, J. (1855). *On the Mode of Communication of Cholera*, London, second edition, Map 2, available at https://bit.ly/2lvf9t4

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/