John Snow & the Cholera Epidemic in Mid-19th Century London: 7 Datasets With Documentation for Use in GeoDa

Marcos Falcone
Julia Koschinsky
Peter Vinten-Johansen
Thomas Coleman
Luc Anselin

September 25, 2020
spatial@uchicago.edu

Table of Contents

Overview

John Snow (1855) and his colleagues' quest to discover how cholera was transmitted during the mid-19th century in London has become a classic case for teaching spatial data analysis, causal inference, scientific reasoning, quasi-experimental research design, and spatial epidemiology. Our goal at the University of Chicago's Center for Spatial Data Science has been to make various existing datasets related to the famous Broad Street pump and Grand Experiment cases available in one place for teaching and learning exploratory spatial data analysis in our GeoDa software. This document contains the documentation for these seven datasets we are (re-)sharing, including content, sources, and modifications we undertook.

The reason why we compiled existing data on the Snow case is to illustrate the process of generating explanatory insights with spatial data and make it easy to replicate this analysis in GeoDa for teaching purposes. John Snow and colleagues set out to solve the puzzle of cholera transmission, applying scientific reasoning to develop and test their hypotheses (for details, see our summary video and our story map). To replicate and understand some of the insights gained during and after the epidemic, we prepared scripts with instructions for teaching and learning spatial analysis in GeoDa, which can be found here.

These datasets were brought together from different contemporary sources, based on Snow (1855, Maps 1 and 2). In some cases, we modified the spatial boundaries, as explained below. More easily accessible data pertain to the famous Broad Street pump case: They include individual cholera deaths in the RHist package compiled by Waldo Tobler in 1994 (Dataset 1), and cholera deaths aggregated to buildings and blocks (Datasets 2+3), shared by Robin Wilson (2011) and Arribas-Bel et al. (2017). To illustrate spatial outliers with local cluster statistics, we modified the blocks file to add a building near the Broad Street pump, which was unusual in that it was near the Broad Street pump but had few cholera deaths (this was a workhouse where, as it turned out, people drank water from their own well rather than the pump). In addition, we used the pump locations shared by Wilson (2011) to generate two new datasets that aggregate cholera deaths in concentric rings around pump locations (Datasets 5+6).

The datasets related to the Broad Street pump case are limited in that they contain very few variables: The count or rate of cholera deaths and calculations based on deaths (such as distance to nearest pumps). Since more variables are needed to illustrate exploratory spatial data analysis in GeoDa, we also integrated data at the London subdistrict level from the Grand Experiment that Tom Coleman prepared (2019; 2020) based on Snow (1855). We are grateful to Tom Koch for sharing the boundary files prepared for their Koch and Denike (2006) analysis. We used these boundaries (without any attribute data) as our starting point and then worked with the original maps under consultation of Peter Vinten-Johansen and Tom Coleman to make several adjustments to these boundaries. These adjusted boundaries are what we are sharing with Coleman's (2019; 2020) attribute data (Dataset 7), all the modifications to which are

outlined in this document. The spatial boundaries for these subdistricts were previously not publicly available in electronic format.

The documentation is structured as follows: The table below provides an overview of the seven datasets available for download. In the subsequent sections, each dataset is featured in the order of the table, starting with a brief description of the data, a screenshot of the data in GeoDa, and a list of variable names and descriptions. Since there were several modifications to the subdistrict boundaries, we document each of modification steps in additional detail.

Table: Overview of Data

This table summarizes the main characteristics of the 7 datasets, including name, content, and sources, as well as the number of observations and variables. It also indicates where we added modifications.

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	Grand 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

List of Resources

Data to Download: https://geodacenter.github.io/data-and-lab//snow/
Story Map: https://bit.ly/3mSGZiS
Video: https://bit.ly/365giRY
GeoDa Scripts: https://geodacenter.github.io/data-and-lab//data/geoda_scripts_snow.pdf
Snow Data (Tom Coleman): https://github.com/tscoleman/SnowCholera
Snow 1855 Map: https://bit.ly/341fbQH
**GeoDa Download:** https://geodacenter.github.io/

7 Datasets

Deaths (Broad Street Pump)

### 1. Individual cholera deaths ('deaths')

This dataset contains 578 individual deaths during the 1854 cholera epidemic that were compiled by Tobler (1994). We used the projected version distributed through Arribas-Bel et al. (2017). Deaths are recorded as points and are located in the vicinity of the Broad Street pump. Besides the ID and the coordinates of each point, the dataset includes a categorical variable called 'cl' which indicates which of the 6 pumps is closest (by pump ID) (see pumps).
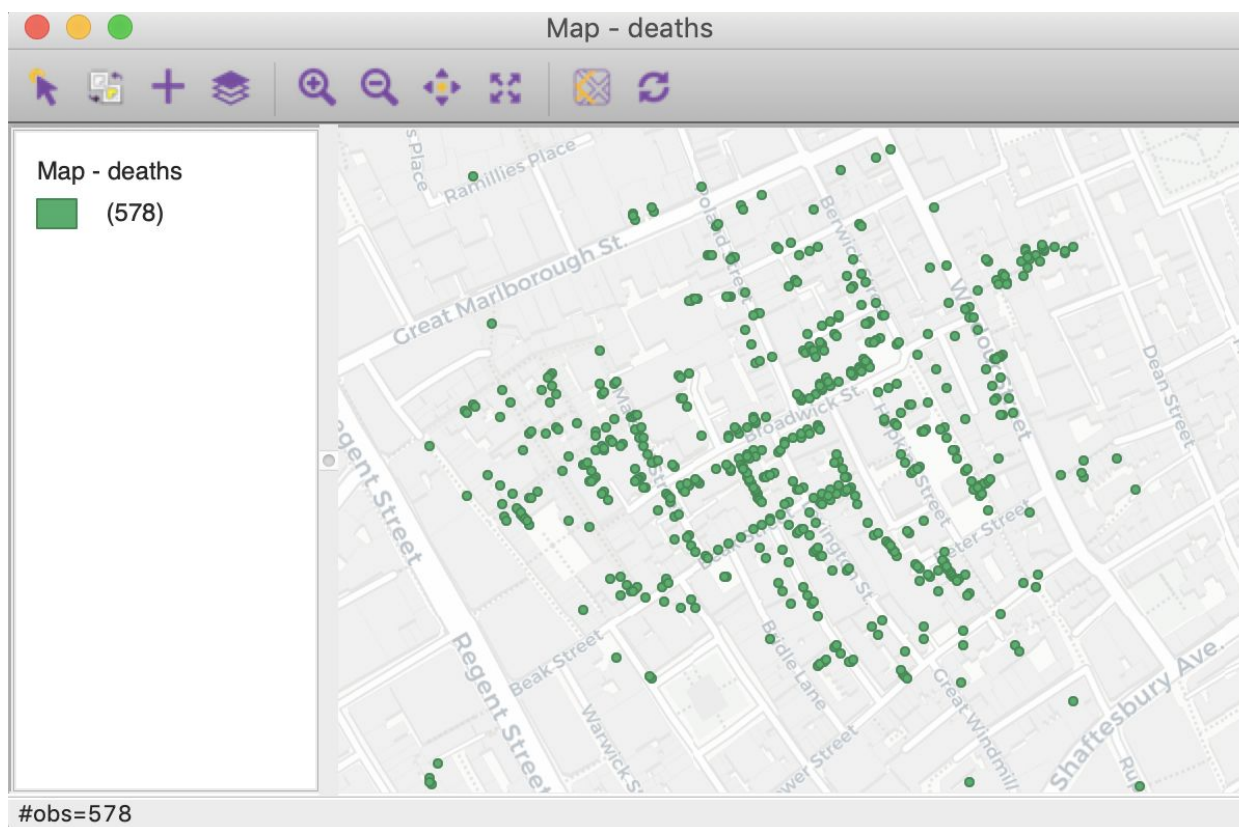
Variable name	Description
ID	ID
lon	Longitude
lat	Latitude
CL	Creates categories depending on which pump is closest (see 'pumps' dataset)

## 2. Cholera deaths aggregated by building ('deaths_by_bldg')

This dataset contains 250 points that correspond to buildings where cholera deaths were recorded near the Broad Street pump. These data were shared publicly by Wilson (2011). The variables include an ID for each building, its coordinates, and a death count. We added the following variables: an ID for the closest pump, as well as the distance to it and to the Broad Street pump (in meters), and, finally, a dummy variable that classifies observations in terms of whether the Broad Street pump was the closest pump to them.
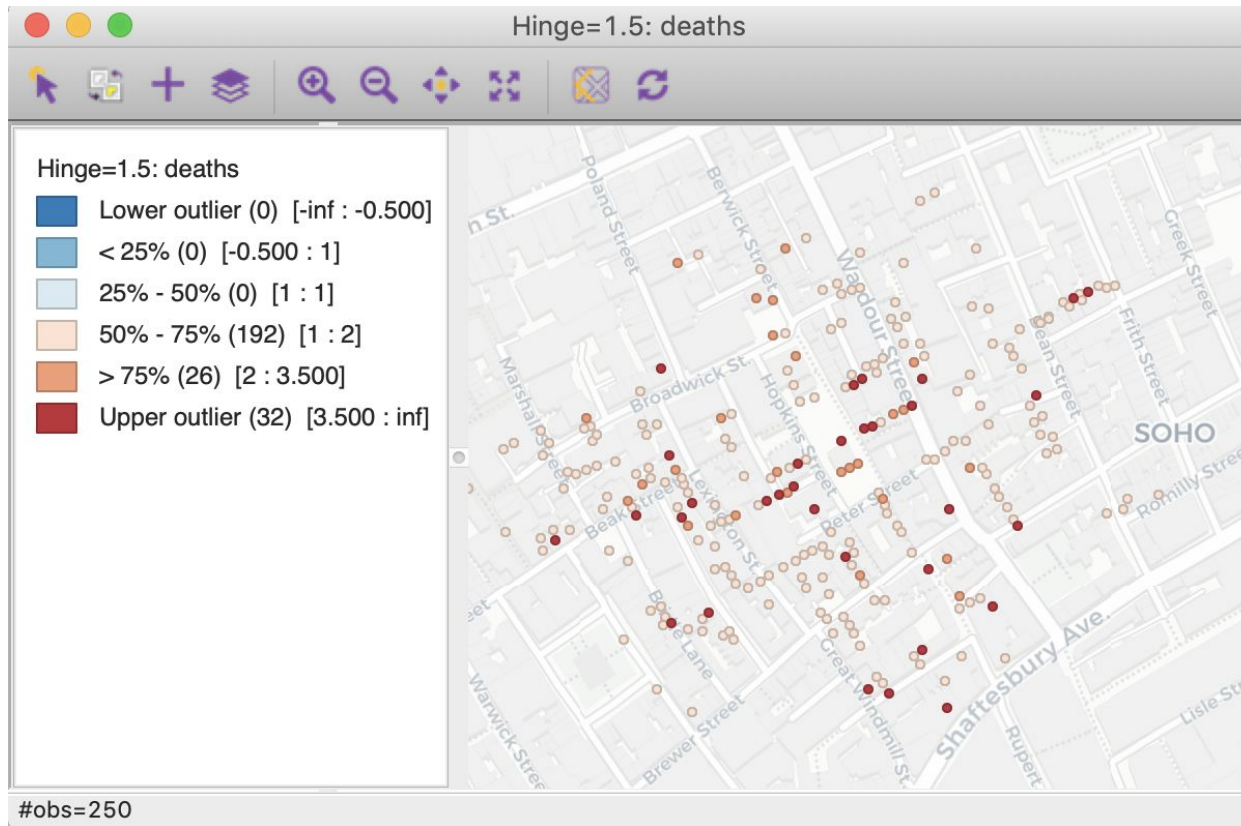
Variable name	Description
ID	ID
x	X coordinates (in meters)
y	Y coordinates (in meters)
deaths	Number of deaths per building
pumpID	ID of the nearest pump (see 'pumps' dataset)
distpump	Distance to the nearest pump (in meters - see 'pumps' dataset)
distBSpump	Distance to Broad St pump (in meters - see 'pumps' dataset)
BSpump	Create categories depending on whether the Broad Street pump is closest (1) or not (0)

3. Cholera deaths aggregated by block ('deaths_by_block')

This dataset contains housing blocks, in the forms of polygons, which aggregate cholera deaths in the vicinity of the Broad Street pump. Originally, 39 observations were provided by Wilson (2011), with an ID for the polygons, a death count and the death density (in terms of population). We also created one additional observation (ID=1) to account for a particular building where John Snow found that people did not drink water from the Broad Street pump, which is the workhouse selected in yellow in the map below.
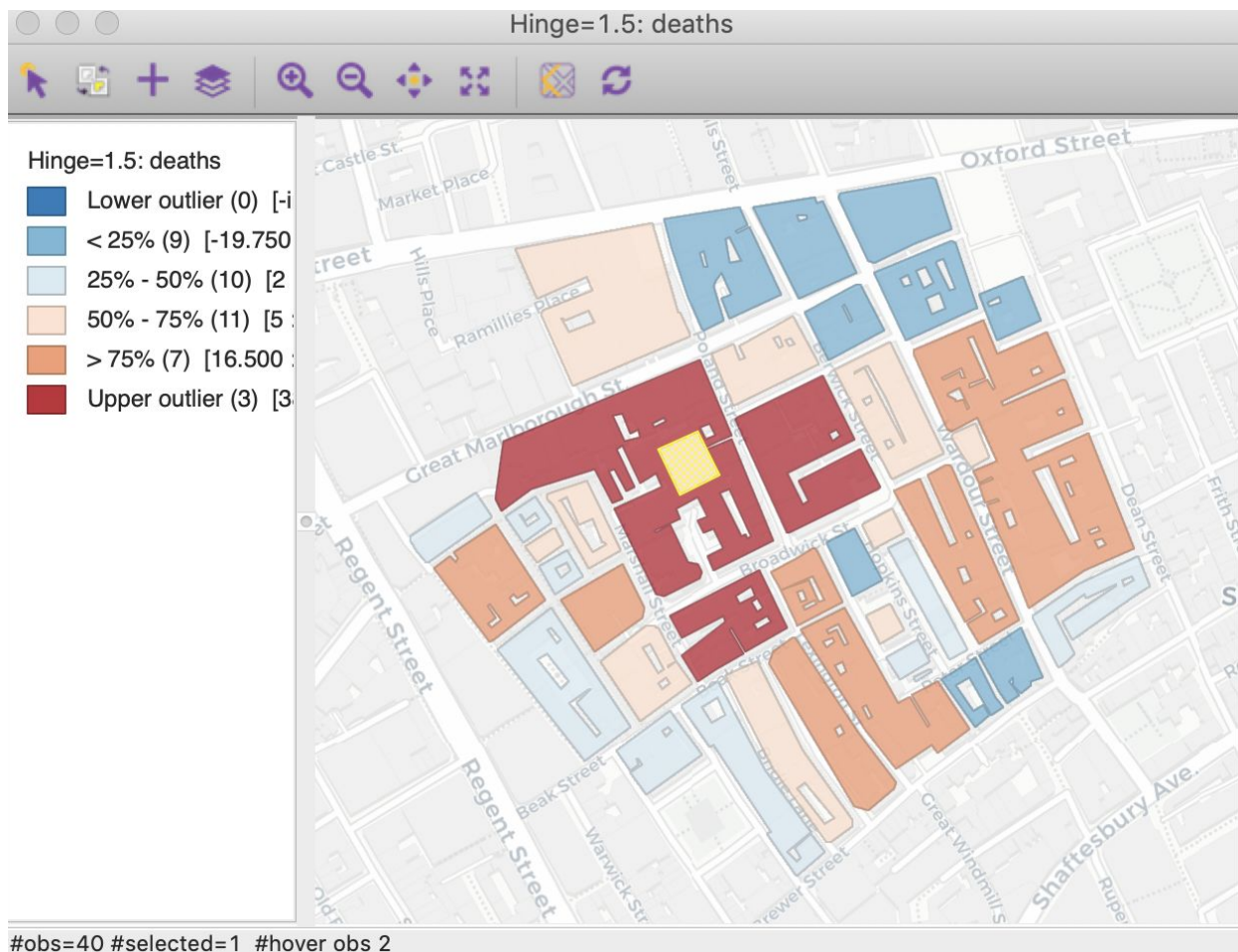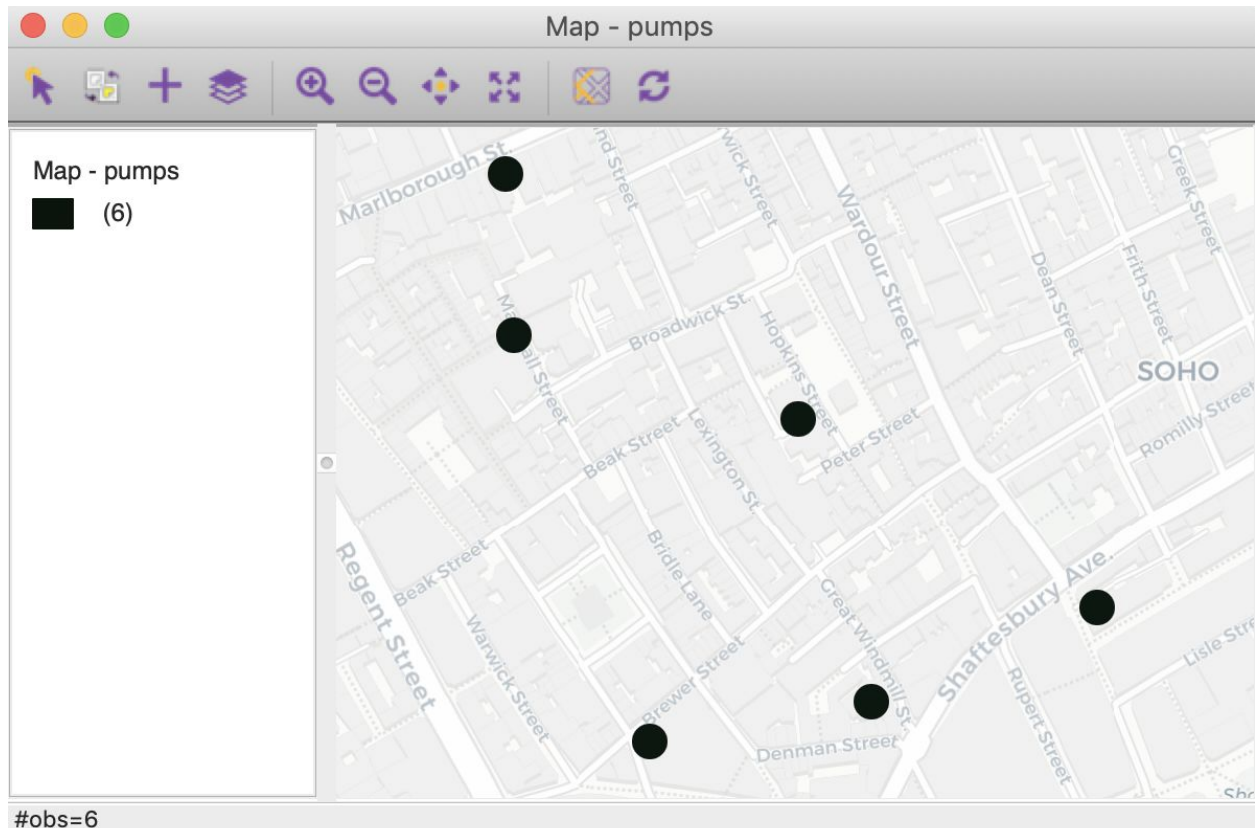

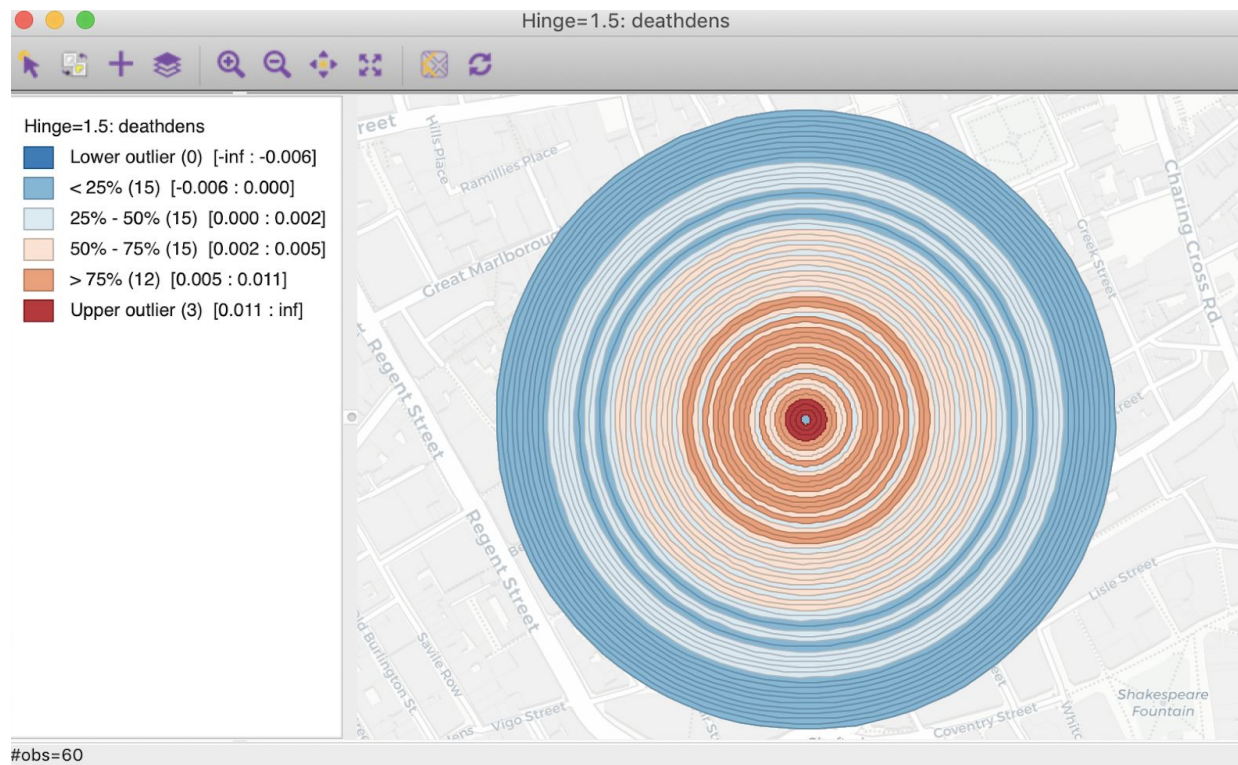| Variable name	Description
ID	ID
deaths	Number of deaths per polygon
deathdens	Number of deaths per polygon divided by population

7

Pumps

4. Six pumps ('pumps')

This dataset contains six points that represent the Broad Street pump and the five pumps closest to it. Originally, the dataset compiled by Wilson (2011) consisted of eight observations, two of which were removed because they showed no deaths in their vicinity (see deaths_by_otherrings). We assume that the spatial extent of the deaths data does not include the other pumps. Variables include an ID for the pumps, their coordinates, and their names.

Variable name	Description
ID	Pump ID
x	X coordinates (in meters)
y	Y coordinates (in meters)
name	Name of the pump

5. Cholera deaths aggregated by Broad Street pump rings ('deaths_by_bsrings')

This dataset combines two other datasets: Individual cholera deaths (see deaths) and the Broad Street pump, extracted from the 6 pumps dataset (see pumps). We created 60 polygons that represent rings that start at the location of the Broad Street pump and progress in 5-meter increments around the pump. We created these rings in QGIS to compare them in terms of cholera deaths, thus the dataset contains the count of deaths per ring. Since the outer rings cover more area than then inner rings, we also included the ring area (in squared meters). This was used to create a variable for death density, i.e. deaths per square meter, shown below.

Variable name	Description
ID	ID
area	Area (in squared meters)
deaths	Number of deaths per ring
deathdens	Number of deaths per ring divided by area

6. Cholera deaths aggregated by other pump rings ('deaths_by_otherrings')

This dataset combined the same two datasets as above: Individual cholera deaths (see deaths) and the five pumps except for the Broad Street pump, extracted from the 6 pumps dataset (see pumps). The dataset contains 35 polygons which represent seven rings that progress in 10-meter increments from each of the 5 pumps (see deaths_by_bsrings). Its variables include an ID for the rings and for the pumps, respectively, the coordinates of the pumps, the cholera death count per ring and the distance from each ring to the closest pump.
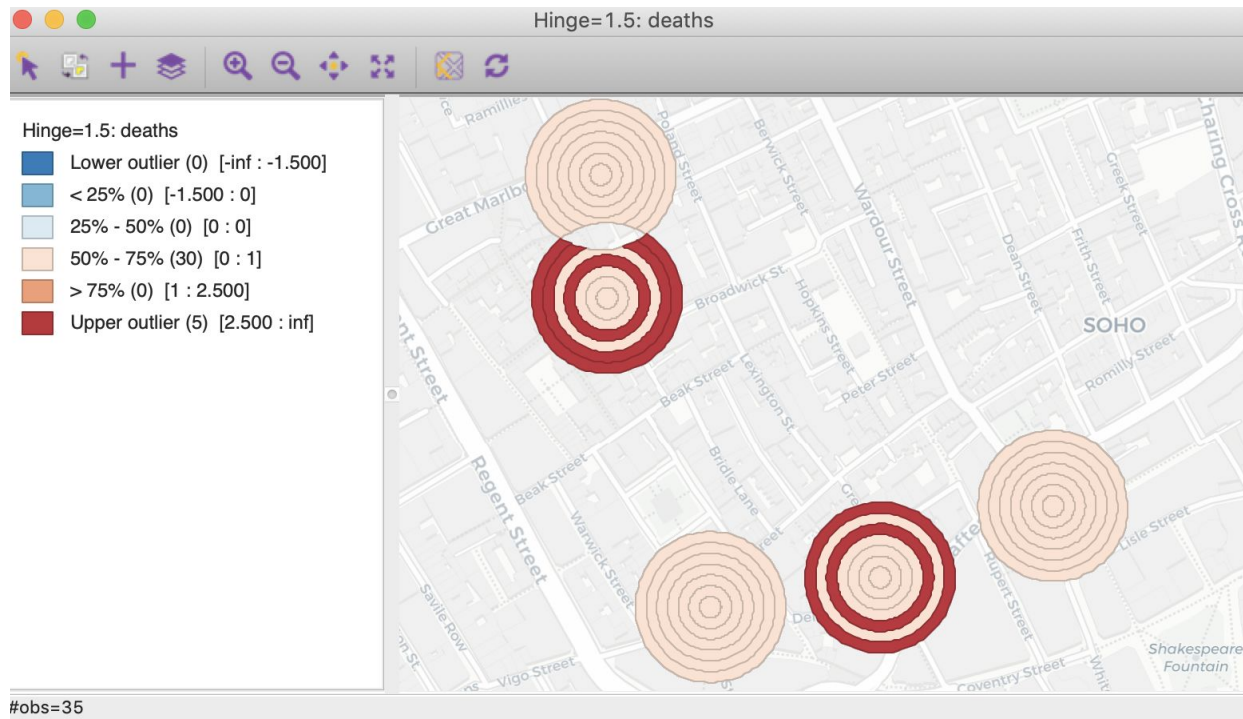

| Variable name	Description
ID	ID
pump_ID	ID of corresponding pump
x	X coordinates of corresponding pump (in meters)
y	Y coordinates of corresponding pump (in meters)
dist	Distance to pump (in meters)
deaths	Number of deaths per ring

Subdistricts (Grand Experiment)

7. 28 subdistricts ('subdistricts')

Since GeoDa is primarily designed for the exploratory analysis of variables associated with spatial areas, we were interested in adding South London subdistricts to the previous datasets. The rich set of attribute data we wanted to use for this purpose was prepared by Tom Coleman (2019; 2020) based on the Grand Experiment. We used the spatial boundary files from Koch and Denike (2006) and modified these boundaries by consulting the original maps and the historian Peter Vinten-Johansen (2020), as described below.

We refer to South London subdistricts as the boundaries in John Snow's Map 2, which appeared in his 1855 report (Snow 1855). You can access this map here.
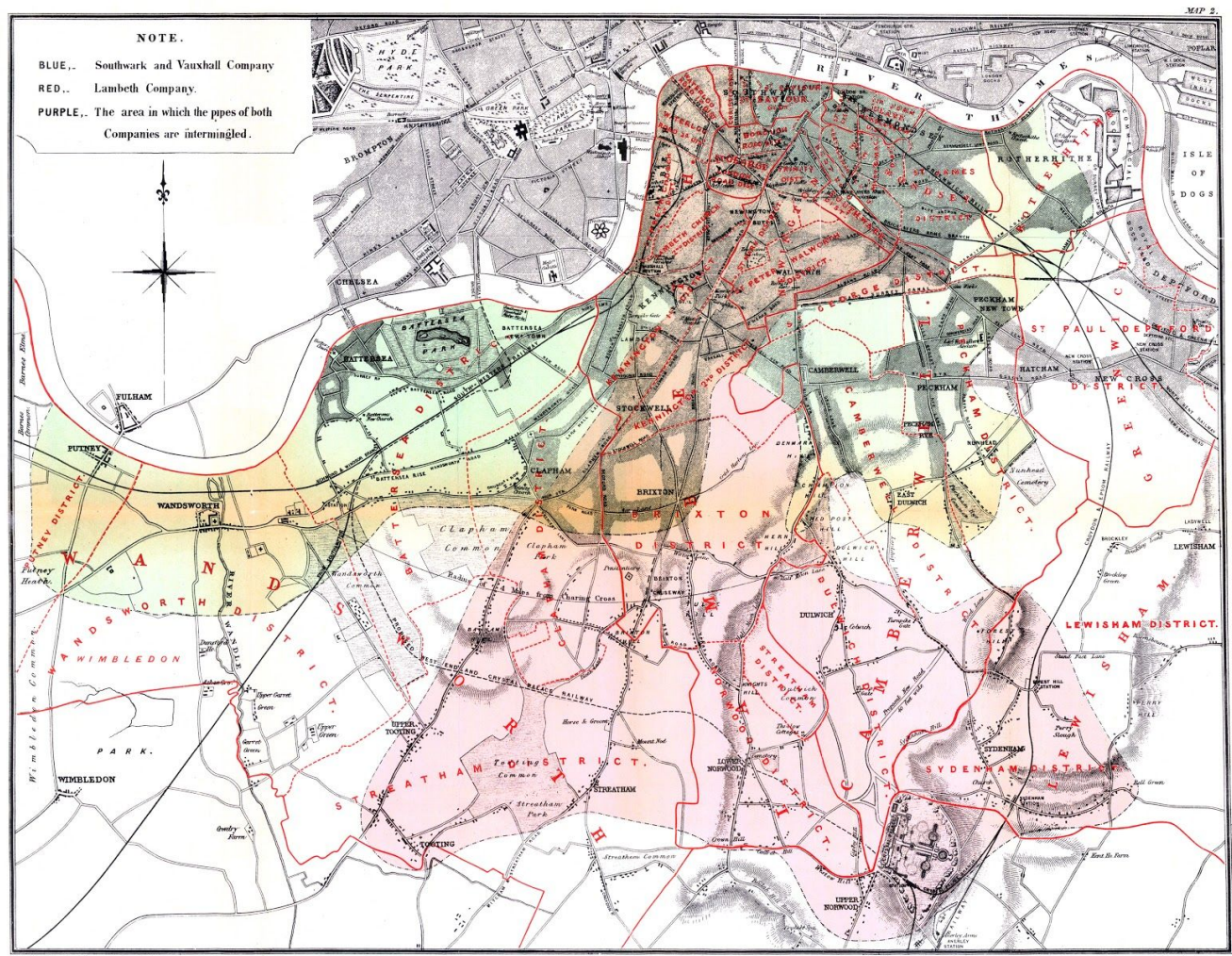
We georeferenced the original Snow map in QGIS. By adding almost 30 control points, the location of the map was very close to that of its actual features. Here is what the outcome looked like in QGIS on top of a current basemap:

11
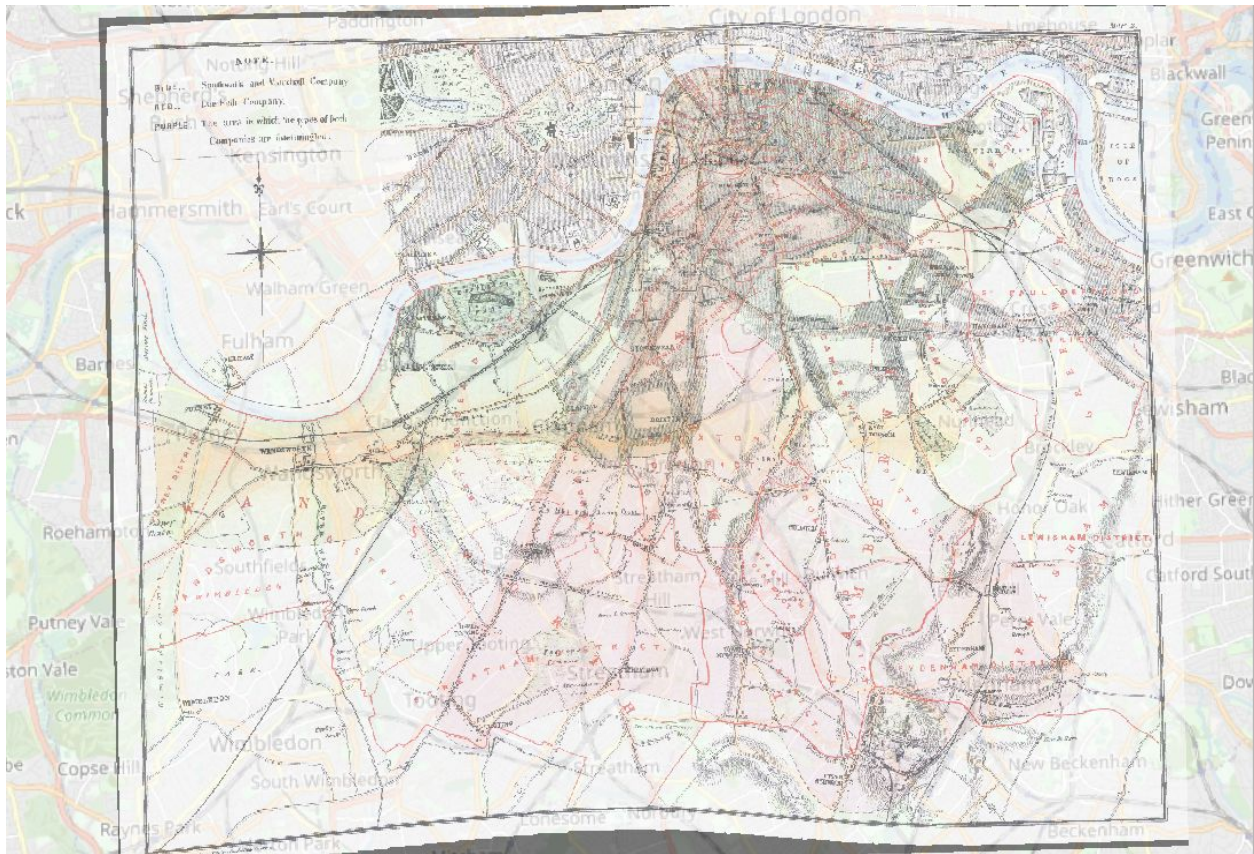
The spatial boundary files (without attribute data) that we used as initial input for the project were originally digitized by Koch and Denike (2006) and shared by the authors. Their original unprojected file, which consists of 41 observations, looks like this:
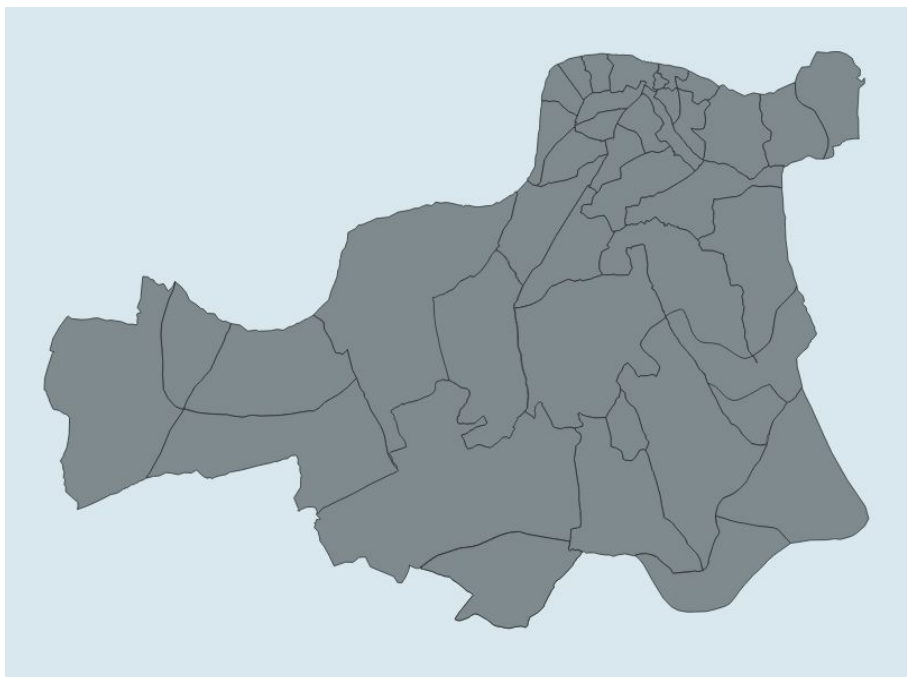
Since the Grand Experiment attribute data compiled by Coleman (2019) that we are using is only available for 28 subdistricts, we excluded the rest of the observations for which there was no attribute data. The resulting (now projected) map consists of the 28 adjacent subdistricts below:
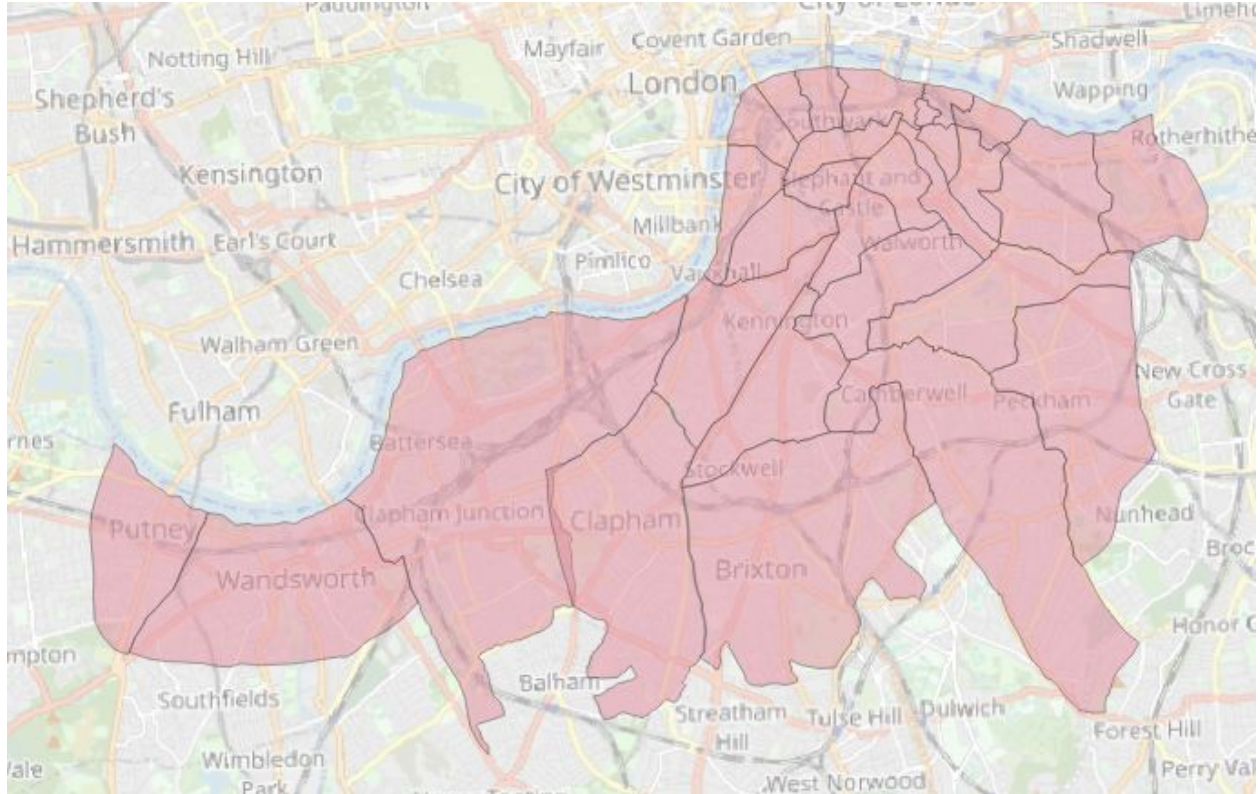

After overlaying Koch and Demike's 2006 shapefile with the 1855 map, we noticed mismatches between the boundaries of the spatial file and the reference map, particularly in subdistricts that were located near the edge of the map. Additionally, some of these subdistricts were stored as multi-polygons in QGIS. We also realized that the boundaries of some of the subdistricts reflected the extent of South London's watersheds instead of their actual administrative borders. We therefore adjusted the subdistrict boundaries as described below.

The boundaries of Putney, Wandsworth, Peckham, and Rothertithe were extended to align them more closely with the 1855 map. In the case of Putney, we presume that its boundaries extend to the Northwest. However, Snow's map does not show them, as he was concerned with the extent of the watersheds, so we kept the end of the 1855 map as Putney's northwestern boundary. The resulting map looks as follows:
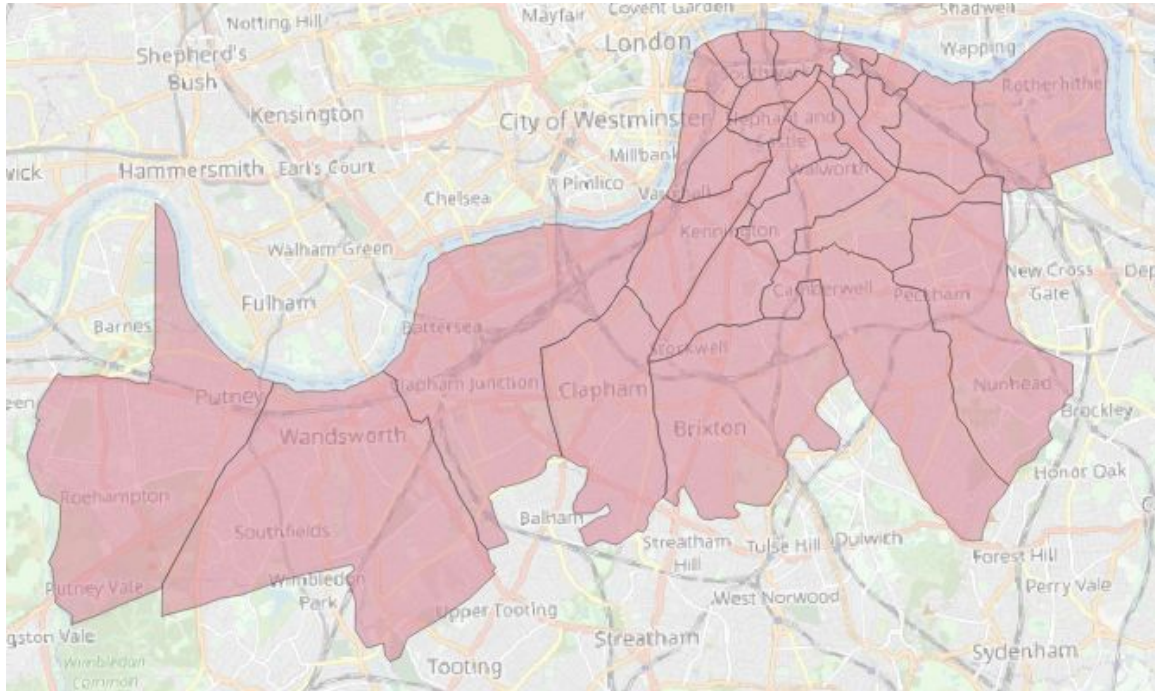
You will notice a small hole in the northeast quadrant which does not correspond to a residential area. This turns out to be Guy's Hospital, which was founded in 1721.

When running spatial weights, we noticed that three observations were neighborless. This sometimes happens during the digitization process if small gaps are left between areas and their neighbors. We fixed this problem in QGIS by cleaning the layer and obtained a final layer (in pink) which can be compared to the initial one (in green) below. The maps also display the 1855 map and a current London basemap as base layers.
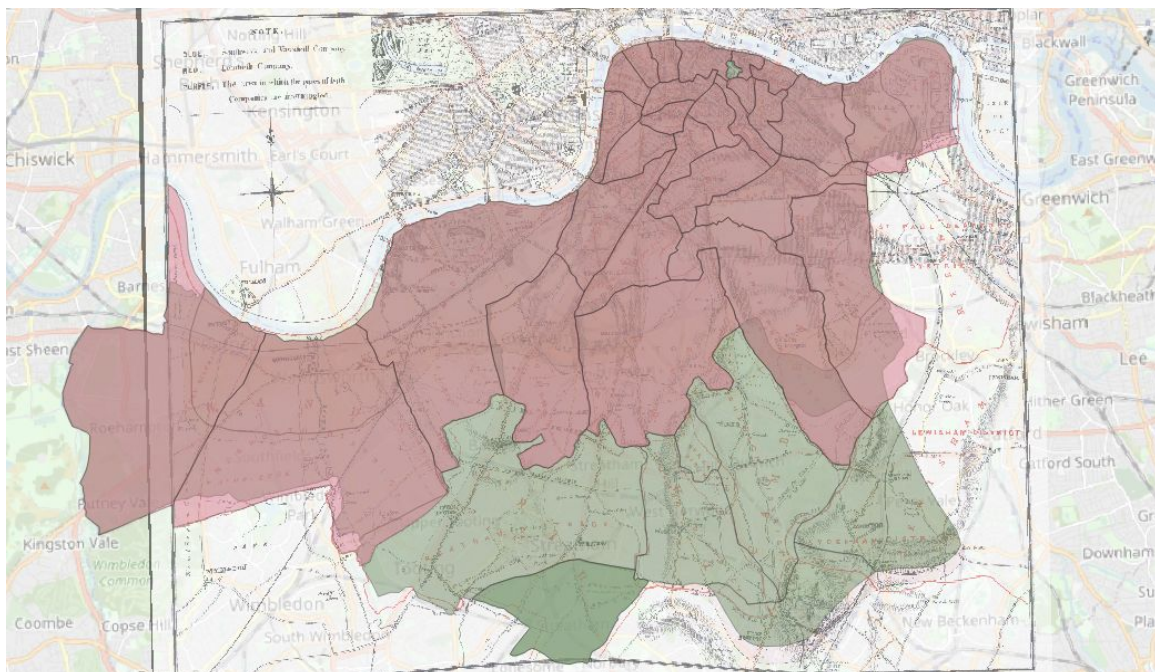
This is what the final layer looks like in GeoDa:
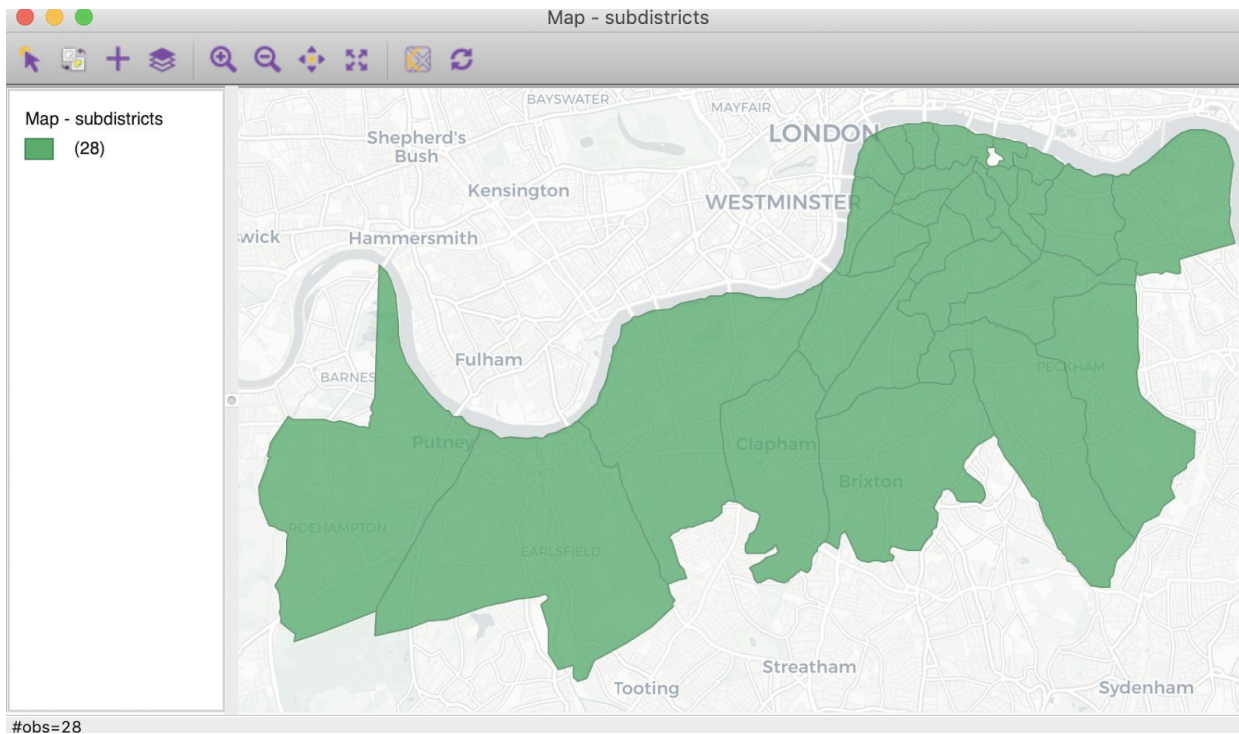
Variable name	Description
dis_ID	London district ID
district	London district
sub_ID	London subdistrict ID
subdist	London subdistrict
pop1851	Population for 1851
supplier	Water company suppliers that served the subdistrict
supplierID	Water company supplier ID
perc_sou	Proportion of the population that was served by the Southwark & Vauxhall company
perc_lam	Proportion of the population that was served by the Lambeth company
perc_other	Proportion of the population that was served by a company other than Southwark & Vauxhall or Lambeth
lam_degree	Creates categories for the proportion of the population that was served by the Lambeth company
d_overall	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854
d_sou	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854 for the Southwark & Vauxhall company
d_lam	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854 for the Lambeth company
d_pump	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854 originating in pump-wells

Variable name (continued)	Description
d_thames	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854 from water from the Thames River and ditches
rate_sou7w	Southwark & Vauxhall cholera death rate per 10000 people in the seven weeks ending August 26, 1854
rate_lam7w	Lambeth cholera death rate per 10000 people in the seven weeks ending August 26, 1854 - Missing values are undefined and should not be converted to 0
rate_oth7w	Cholera death rate per 10000 people for 'other' category in the seven weeks ending August 26, 1854 - Missing values are undefined and should not be converted to 0
deaths1849	Number of deaths attributed to the cholera epidemic in 1849
deaths1854	Number of deaths attributed to the cholera epidemic in 1854
rate1849	Cholera death rate per 10000 people in 1849
rate1854	Cholera death rate per 10000 people in 1854
pop1849	Population for 1849
pop1854	Population for 1854
rAvSupR_49	Average supplier-region-specific cholera mortality rate per 10000 people in 1849
rAvSupR_54	Average supplier-region-specific cholera mortality rate per 10000 people in 1854
pred_Snow	Snow's cholera death count prediction (from his 1856 Table VI)
pred_DiD49	Cholera death count prediction from Difference-in-Difference regression analysis for 1849
pred_DiD54	Cholera death count prediction from Difference-in-Difference regression analysis for 1854

Acknowledgements

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 at 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).

Koch, T. and K. Denike (2006). Rethinking John Snow's South London study: A Bayesian evaluation and recalculation. *Social Science and Medicine*, 63(1), 271-283. Subdistrict boundary files provided by the author.

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/.