John Snow & the Cholera Epidemic in Mid-19th Century London: 9 Datasets With Documentation for Use in GeoDa

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

September 25, 2020
[Version 2: Edited November 12, 2020]
spatial@uchicago.edu

Table of Contents

Overview

John Snow's 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 South London cases available in one place for teaching and learning exploratory spatial data analysis via our GeoDa software and other spatial analysis programs. This document contains the documentation for these nine datasets we are (in most cases re-)sharing, including content, sources, and modifications we undertook.

The reason we compiled existing data on 19th century cholera outbreaks 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 applied scientific reasoning to develop and test hypotheses about the nature and communication of cholera (for details, see our summary video and our story map). To replicate and understand some of the insights Snow gained during and after the epidemic, we prepared scripts with instructions for teaching and learning spatial analysis in GeoDa, which can be found here.

For the first time, we digitized an 1855 map by the General Board of Health (GBoH), which contains deaths and non-deaths per house as well as the location of sewer grates and ventilators (datasets 1 + 8). Because the goal of the GBoH was to discredit the idea that cholera was spreading through sewer grates near newly constructed lines that ran in what had been a pest-field during the 17th century, this map can be used to test an alternative explanation to the waterborne theory of cholera, namely a variant of the miasma theory. With 1,852 houses, this is the largest dataset in this document.

Additionally, we brought together existing datasets from several sources, based on Snow (1855, maps 1 and 2). In some cases, we modified the spatial boundaries, as explained below. Easily accessible data pertain to the famous Broad Street pump case: Individual cholera deaths in the RHist package compiled by Waldo Tobler in 1994 (dataset 2) and cholera deaths aggregated to buildings and blocks (datasets 3+4), 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 (the parish workhouse) that had few cholera deaths because the inmates drank water from their own well. In addition, we used the pump locations shared by Wilson (2011) (dataset 7) 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 South London investigation that Tom Coleman prepared (2019; 2020) based on Snow (1855). We are

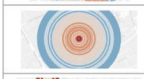
grateful to Tom Koch for sharing the boundary files prepared for his analysis with Denike (2006). We used them (without any attribute data) as our starting point and then worked with the original maps and consulted with Peter Vinten-Johansen and Tom Coleman to adjust the boundaries, which what we are sharing with Coleman's (2019; 2020) attribute data (dataset 9); all the modifications are outlined in this document. The spatial boundaries for these subdistricts were previously not publicly available in electronic format.

The table below provides an overview of the nine datasets available for download. Subsequent sections discuss each dataset, starting with a brief description of the data, a screenshot of the data in GeoDa, a list of variable names and descriptions, and potential modifications of the data. We document each modification of the subdistrict boundaries in additional detail.

Table: Overview of Data

This table summarizes the main characteristics of the 9 datasets, including name, content, and sources, as well as the number of observations and variables. It also indicates where we added modifications.

Overview of 9 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	Individual deaths	Broad St Pump	Point	578	4	Tobler 1994, Arribas-Bel et al. 2017	Snow 1855 (Map 1)	GPL
	3. 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
	4. 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
	5. 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
	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. 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
	8. sewergrates_ventilators	Untrapped sewer grates and ventilators	Broad St Pump	Point	325	5	**Digitized by CSDS**	General Board of Health 1855	GPL
	9. 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

List of Resources

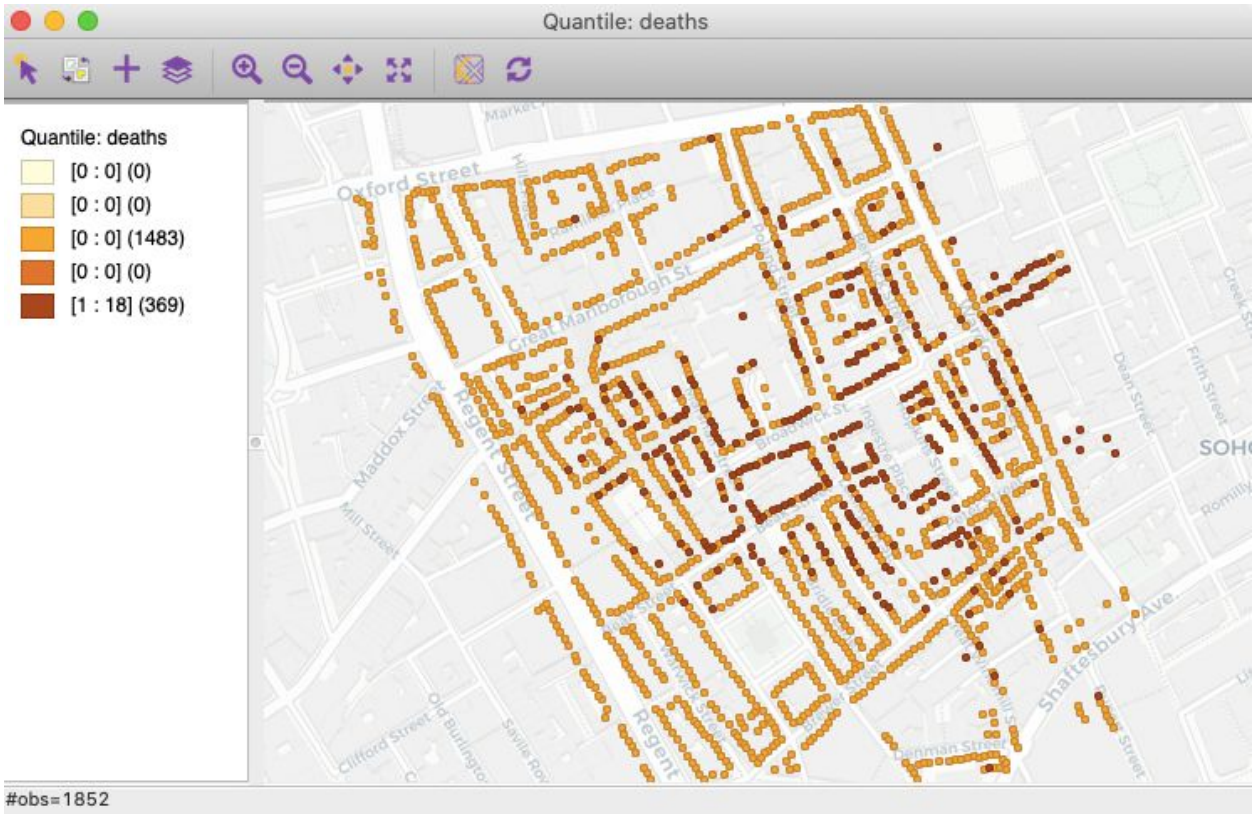
Data to Download: https://geodacenter.github.io/data-and-lab//snow/
Story Map: https://bit.ly/3mSGZiS and **video**: https://youtu.be/lGN8SK1Y1h4
GeoDa Scripts: https://bit.ly/3pqK7no
Snow Data (Tom Coleman): https://github.com/tscoleman/SnowCholera
Snow 1855 Maps: https://bit.ly/32Az1lW (1) and https://bit.ly/2Ivf9t4 (2)
GeoDa Download: https://geodacenter.github.io/

9 Datasets

The Soho Outbreak: Deaths

1. Cholera deaths and non-deaths aggregated by house ('deaths_nd_by_house')

This dataset contains 1,852 points that represent houses in the Soho neighborhood as featured in the 1855 map by the General Board of Health. For each house, the variables include an ID, the number of cholera deaths by residents and non-residents of the building (including non-deaths), a total death count for both types of residents, and a dummy variable that indicates whether the house was located on top of Craven Estate, a 17th-century pest-field. They also include three variables that measure distances to the pest-field, the nearest open sewer grate or ventilator, and the Broad Street pump.

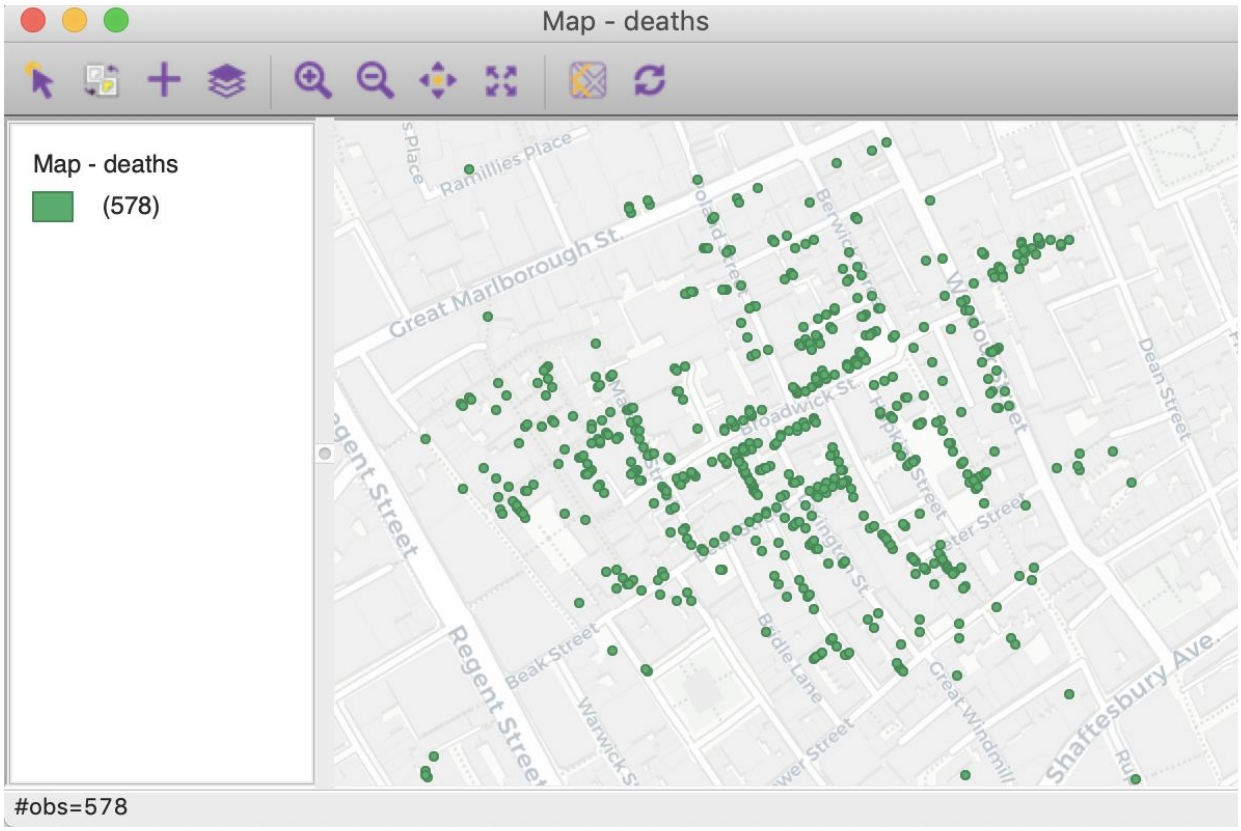


Variable name	Description
ID	ID
deaths_r	Number of deaths of residents per house
deaths_nr	Number of deaths of non-residents per house
deaths	Number of total deaths per house
pestfield	Creates categories depending on whether the house is located on top of the former Craven Estate (1) or not (0).

dis_pestf	Distance to the former Craven Estate, in meters.
dis_sewers	Distance to the nearest open sewer grate or ventilator, in meters.
dis_bspump	Distance to the Broad Street pump, in meters.

2. Individual cholera deaths ('deaths')

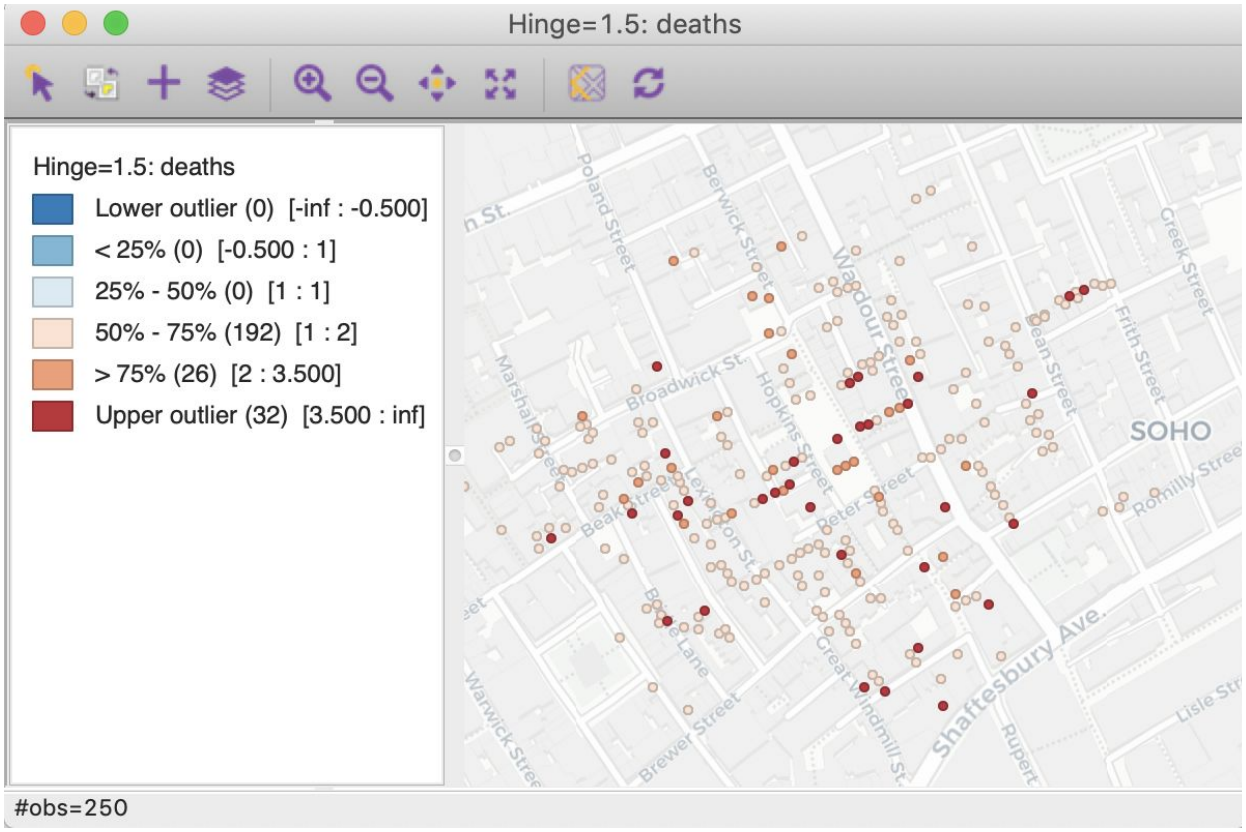
This dataset contains 578 individual deaths that occurred during the 1854 cholera epidemic compiled by Tobler (1994). We used the projected version distributed through Arribas-Bel et al. (2017). Deaths are recorded as points and are located in the cholera field surrounding the Broad Street pump. Besides 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)

3. Cholera deaths aggregated by building ('deaths_by_bldg')

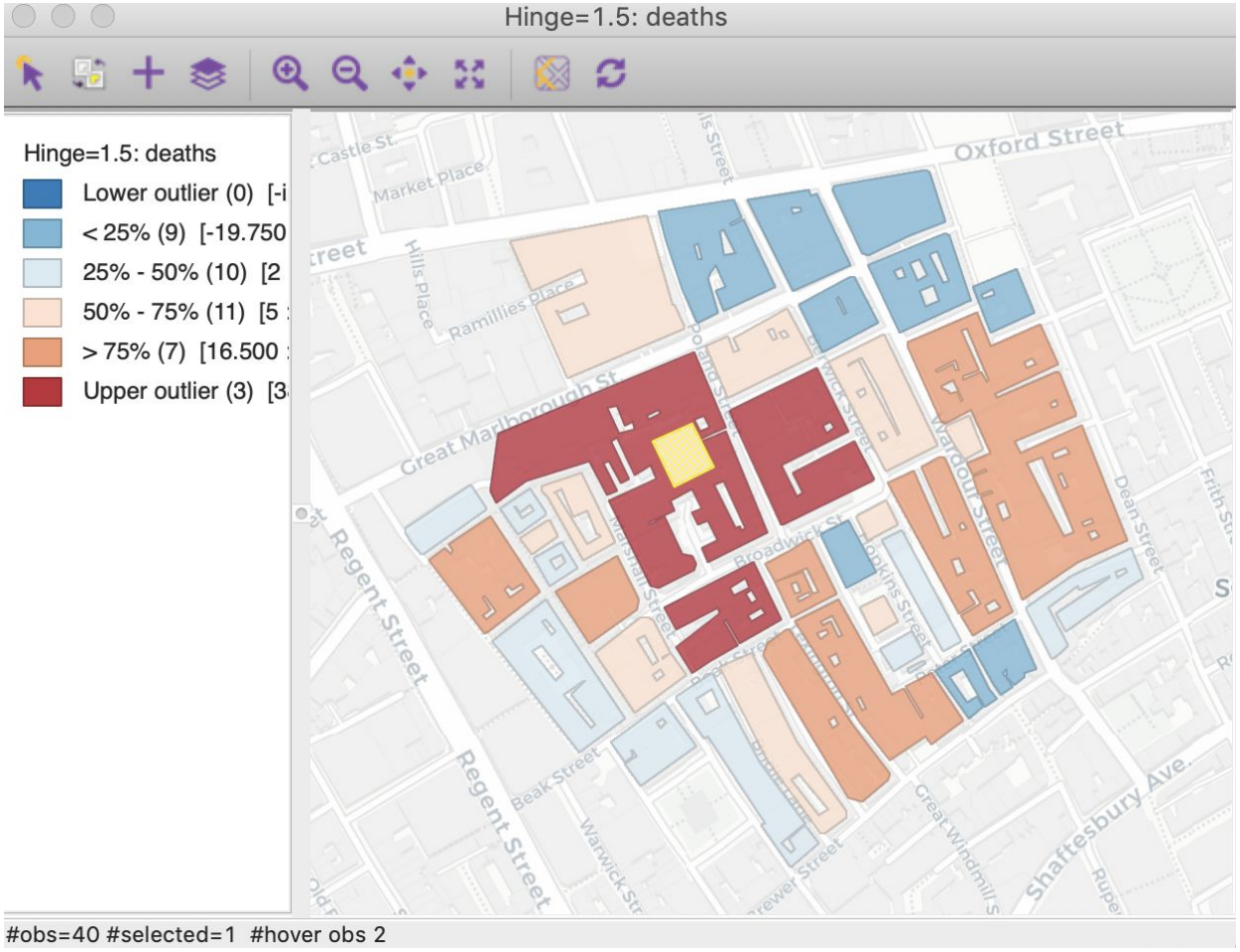
This dataset contains 250 points that correspond to buildings where cholera deaths were recorded near the Broad Street pump. These data were shared publicly by Wilson (2011). The variables include an ID for each building, its coordinates, and a death count. We added the following variables: an ID for 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.



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	Creates categories depending on whether the Broad Street pump is closest (1) or not (0)

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

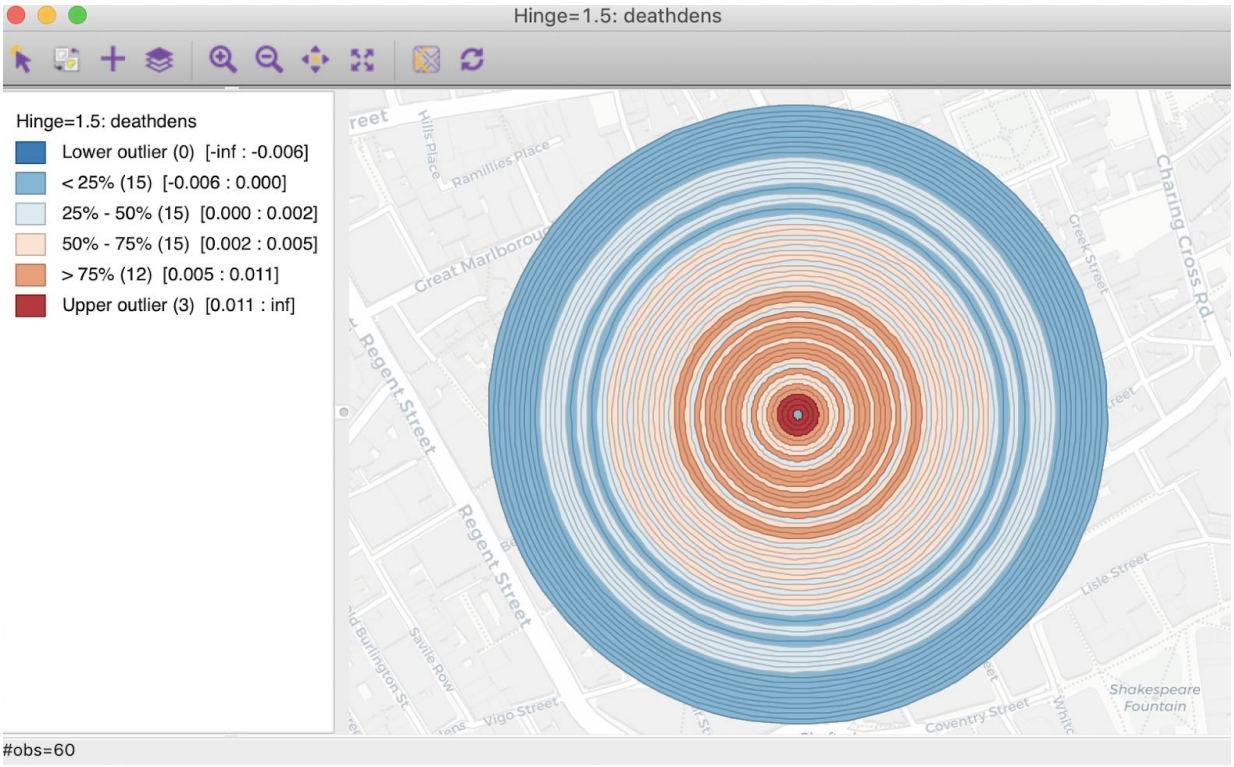
This dataset contains housing blocks, in the form of polygons, which aggregate cholera deaths in the vicinity of the Broad Street pump. Originally, 39 observations were provided by Wilson (2011), with an ID for the polygons, a death count and the death density (in terms of population). We also created 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 — the workhouse depicted 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

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

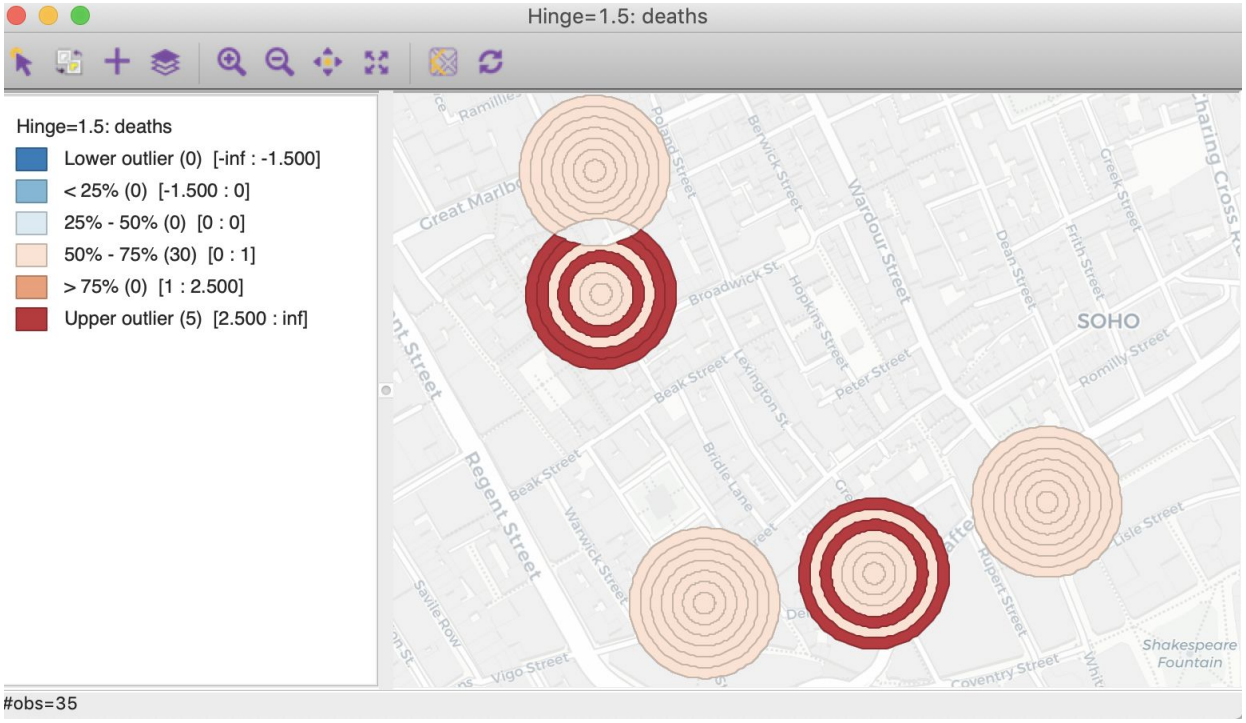
This dataset combines two of the previous datasets: Individual cholera deaths (see deaths) and the Broad Street pump, extracted from the 6 pumps dataset (see pumps). We created 60 polygons that represent rings that start at the location of the Broad Street pump and progress in 5-meter increments around the pump. We created these rings in QGIS to compare them in terms of cholera deaths; thus, the dataset contains the number of deaths per ring. Since the outer rings cover more area than the inner rings, we also included the ring area in square meters. This variable was used to create another one 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: Individual cholera deaths (see deaths) and the five pumps other than 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 the pumps (respectively), the coordinates of the pumps, the cholera death count per ring, and the distance from each ring to the closest pump.

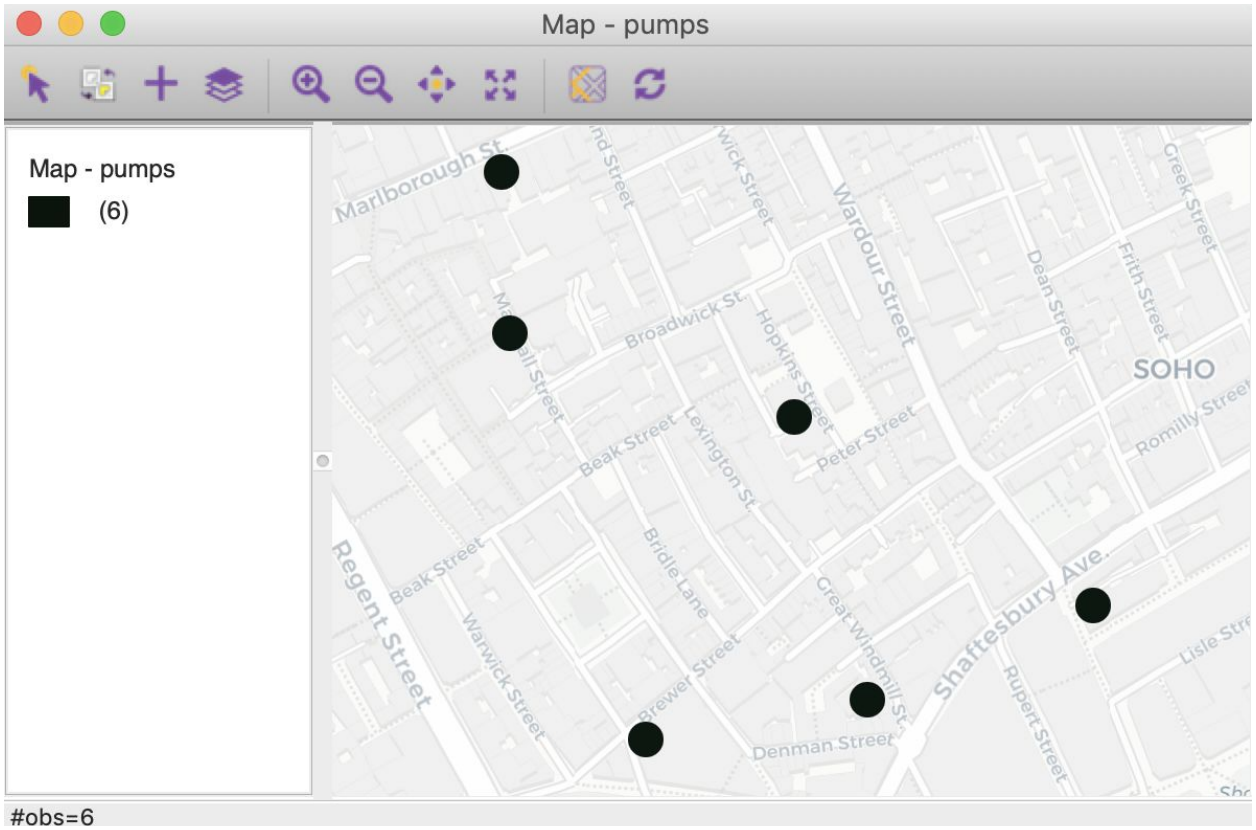


Variable name	Description
ID	ID
pump_ID	ID of corresponding pump
x	X coordinates of corresponding pump (in meters)
y	Y coordinates of corresponding pump (in meters)
dist	Distance to pump (in meters)
deaths	Number of deaths per ring

The Soho Outbreak: Potential Correlates

7. Pumps ('pumps')

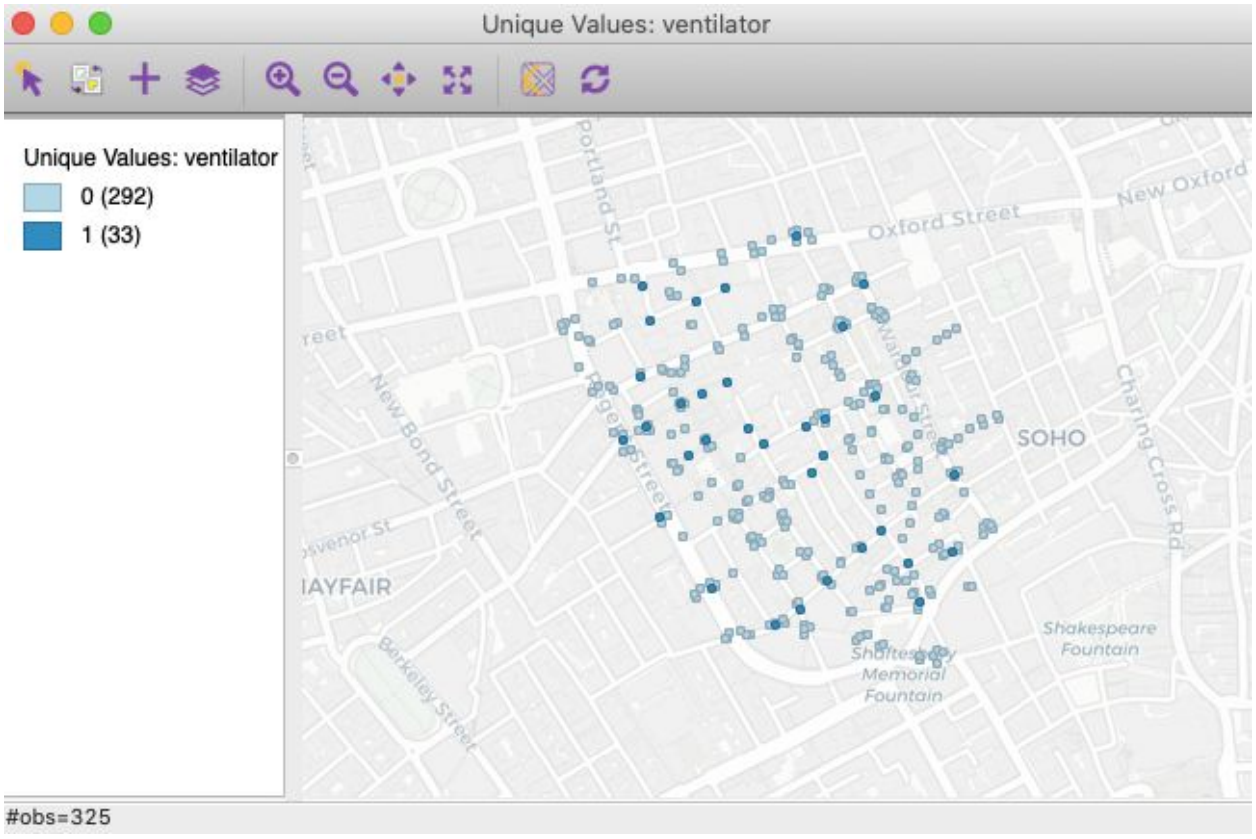
This dataset contains six points that represent the Broad Street pump and the five pumps closest to it. Originally, the dataset compiled by Wilson (2011) consisted of eight observations, two of which were removed because they showed no deaths in their vicinity (see deaths_by_otherrings). We assume that 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

8. Sewer grates and ventilators ('sewergrates_ventilators')

This dataset contains 325 points that represent untrapped sewer grates and ventilators in the Soho neighborhood as represented in the 1855 map by the General Board of Health. For each observation, the variables include an ID, the date where the corresponding sewer line was built both as a string and as a categorical value, and a dummy variable that distinguishes untrapped sewer grates (which allowed sewage to get into the sewer lines) from ventilators (which were supposed to pull fresh air into the sewers but about which neighbors complained since foul smells emanated from them). Another dummy variable indicates whether the house was located on top of Craven Estate, a 17th-century pest field, since people thought toxic gases were coming from the field through the untrapped sewer grates and ventilators.



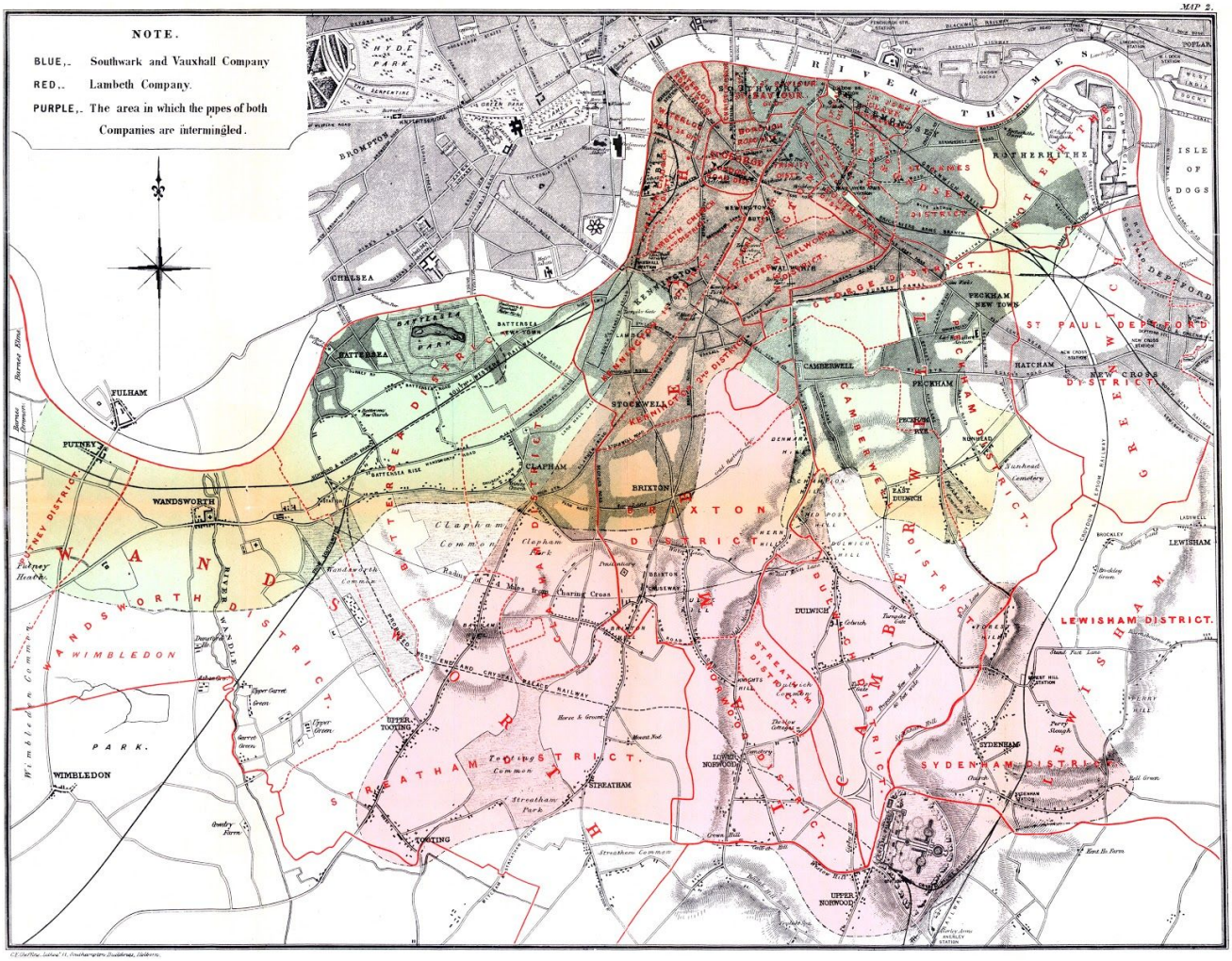
Variable name	Description
ID	ID
date	Date where corresponding sewer line was built
date_code	Date where corresponding sewer line was built as categories: 1=Before 1851, 2=1851, 3=1854
ventilator	Creates categories depending on whether the observation is a ventilator (1) or an untrapped sewer grate (0)
pestfield	Creates categories depending on whether the sewer grates or ventilator is located on top of the former Craven Estate (1) or not (0).

Subdistricts (South London natural experiment)

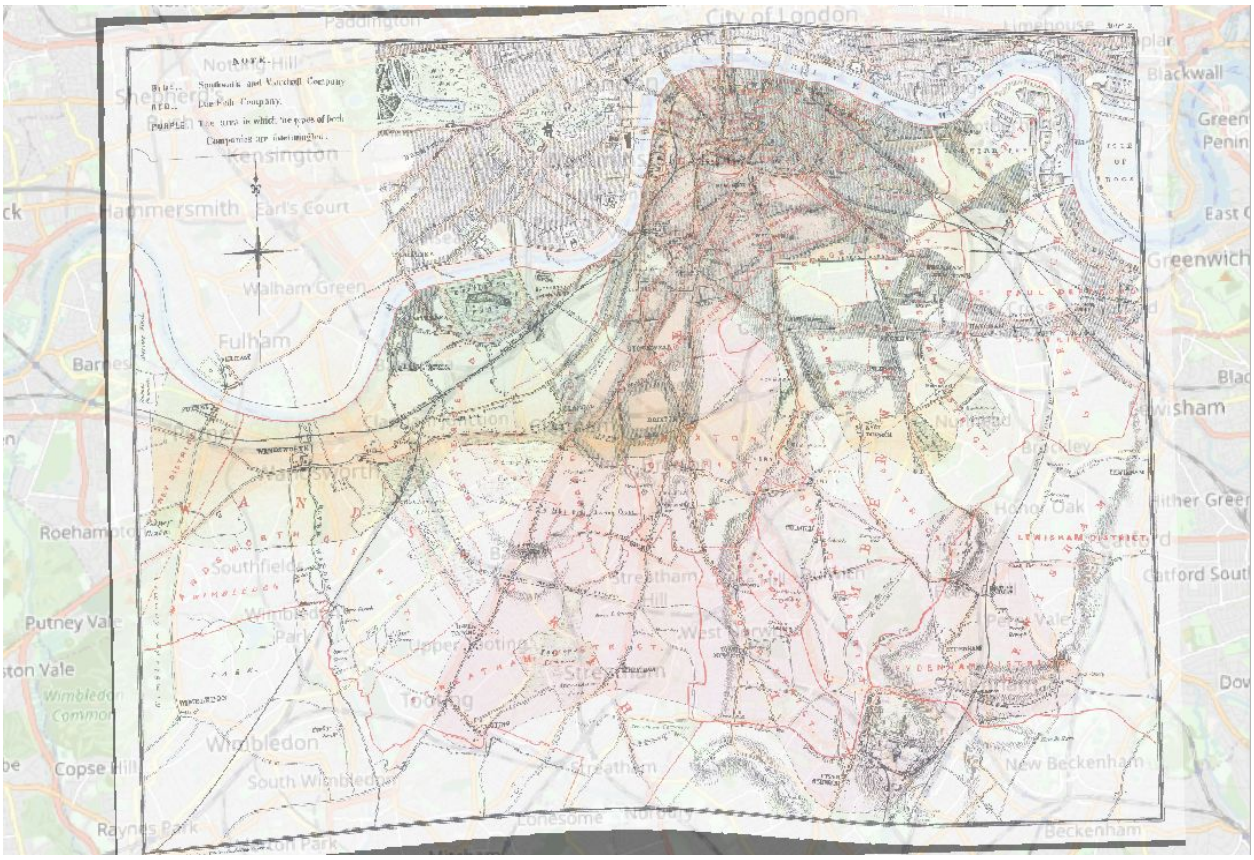
9. The 32 subdistricts ('subdistricts')

Since GeoDa is primarily designed for the exploratory analysis of variables associated with spatial areas, we decided to include a dataset related to Snow's investigation of the natural experiment in South London. The rich set of attribute data we wanted to use for this purpose was prepared by Tom Coleman (2019; 2020). We used the spatial boundary files from Koch and Denike (2006) and modified these boundaries after consulting the original maps and the historian Peter Vinten-Johansen (2020).

The South London subdistricts we refer to are shaped by the boundaries in John Snow's Map 2, which appeared in his 1855 report (Snow 1855). You can access this map here.



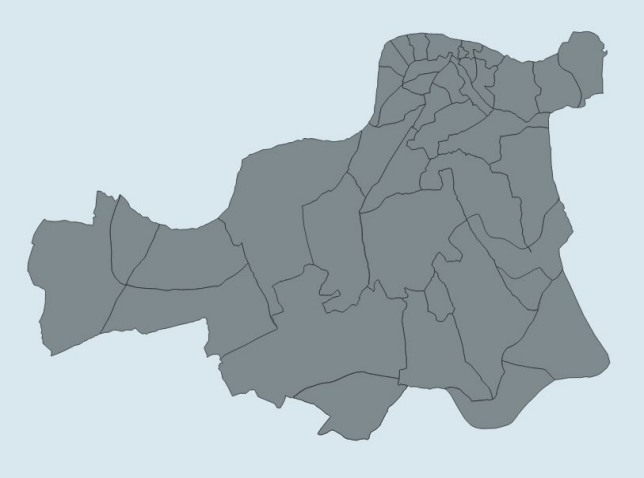
We georeferenced the original Snow map in QGIS. By adding almost 30 control points, the location of the map was very close to that of its actual features. Here is what the outcome looked like in QGIS on top of a current basemap:



Map 2 includes 32 South London subdistricts in which the Southwark & Vauxhall (S&V) and the Lambeth water companies supplied piped water in 1854. S&V was the sole provider in 12 subdistricts and Lambeth was the sole provider in 4 subdistricts; active pipes from both companies existed in 16 subdistricts (either the entire subdistrict or a portion of it). You can see an adaptation of Table XII from Snow (1855:90) at the end of this section.

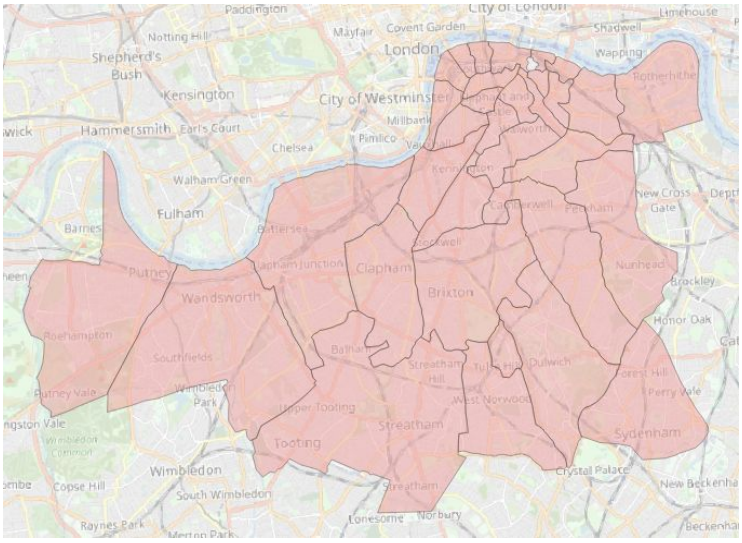
While we decided to keep the 32 subdistricts in this dataset, it should be noted that attribute data compiled by Coleman (2019) for a difference-in-differences comparison of cholera mortality during the 1849 and 1854 cholera epidemic in the watersheds of these two water companies is limited to only 28 of the subdistricts. In 1849, the Lambeth Company did not provide piped water to three subdistricts (Norwood, Streatham, and Sydenham) in which it was sole provider in 1854; in the fourth (Dulwich), no cholera deaths were recorded in 1854.

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:



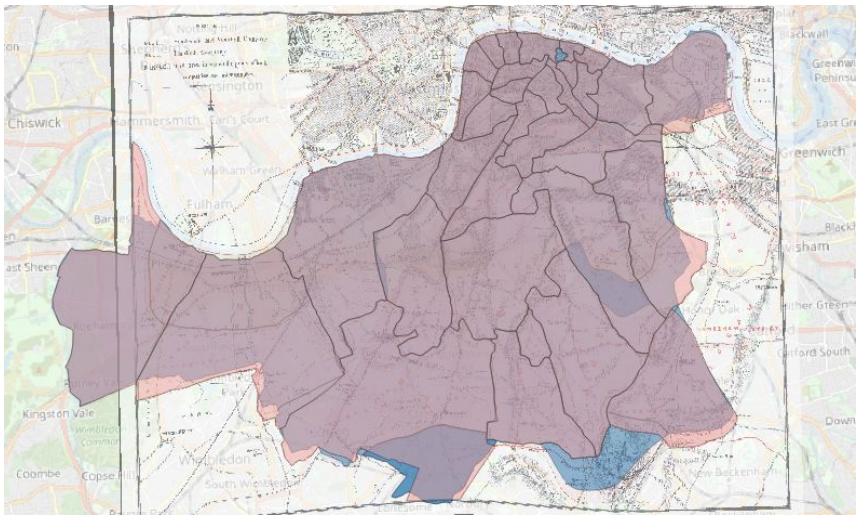
After overlaying Koch and Demike's 2006 shapefile with the 1855 map, we noticed mismatches between the boundaries of the spatial file and the reference map, particularly in subdistricts that were located near the edge of the map. Additionally, some of these subdistricts were stored as multi-polygons in QGIS. We also realized that the boundaries of some of the subdistricts reflected the extent of South London's watersheds instead of subdistrict registration borders. We therefore adjusted the 32 subdistrict boundaries of interest, 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. Snow's map does not show them since he was concerned with the extent of the watersheds, and we kept the end of the 1855 map as Putney's northwestern boundary. In total, we kept 31 polygons and one multipolygon (Streatham, which consisted of a main area in the Wandsworth district and an exclave called Knight's Hill which was located within the Lambeth district). Streatham, as well as Sydenham, were redrawn by us with the QGIS software. The resulting (now projected) map looks as follows:

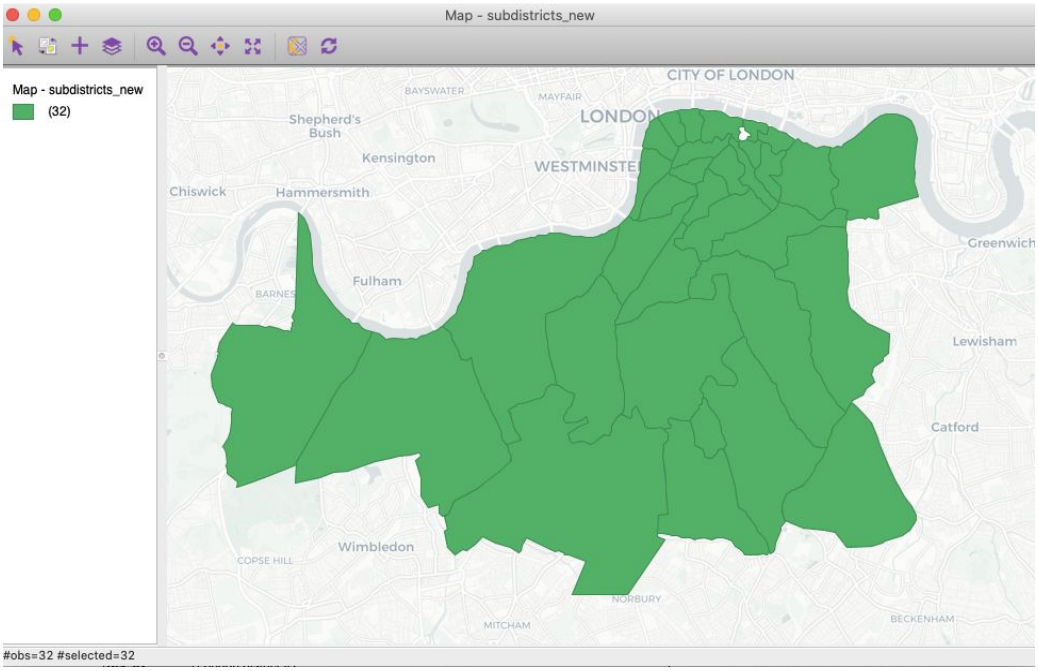


You will notice a small hole in the northeast quadrant which does not correspond to a residential area. This turns out to be Guy's Hospital, which was founded in 1721.

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 blue) below. The maps on this page 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
pop_house	Population per house
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
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
d_unasc	Number of deaths attributed to the cholera epidemic in the seven weeks ending August 26, 1854 of unascertained origin
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

The South London natural experiment
[Adapted from Table XII in Snow (1855:90)]

Subdistrict	Water supplier in 1854
St. Saviour, Southwark	Southwark & Vauxhall
St. Olave	Southwark & Vauxhall
St. John, Horsleydown	Southwark & Vauxhall
St. James, Bermondsey	Southwark & Vauxhall
St. Mary Magdalen	Southwark & Vauxhall
Leather Market	Southwark & Vauxhall
Rotherhithe	Southwark & Vauxhall
Wandsworth	Southwark & Vauxhall
Battersea	Southwark & Vauxhall
Putney	Southwark & Vauxhall
Camberwell	Southwark & Vauxhall
Peckham	Southwark & Vauxhall and Lambeth
Christchurch, Southwark	Southwark & Vauxhall and Lambeth
Kent Road	Southwark & Vauxhall and Lambeth
Borough Road	Southwark & Vauxhall and Lambeth
London Road	Southwark & Vauxhall and Lambeth
Trinity, Newington	Southwark & Vauxhall and Lambeth
St. Peter, Newington	Southwark & Vauxhall and Lambeth
Waterloo Road, 1st	Southwark & Vauxhall and Lambeth
Waterloo Road, 2st	Southwark & Vauxhall and Lambeth
Lambeth Church, 1st	Southwark & Vauxhall and Lambeth
Lambeth Church, 2nd	Southwark & Vauxhall and Lambeth
Kennington, 1st	Southwark & Vauxhall and Lambeth
Kennington, 2nd	Southwark & Vauxhall and Lambeth
Brixton	Southwark & Vauxhall and Lambeth
Clapham	Southwark & Vauxhall and Lambeth
St. George, Camberwell	Southwark & Vauxhall and Lambeth
Norwood	Lambeth
Streatham	Lambeth
Dulwich	Lambeth
Sydenham	Lambeth

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/.

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).

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/21/120/15-78-1DF-22-1855-GBoH-BrSt-Map.pdf.

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://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/.

Acknowledgements

Thank you to Tom Koch for sharing the spatial boundary file from Koch and Denike (2006) as well as other authors who published the Broad Street pump spatial files. We also gratefully acknowledge the support of Luc Anselin, Karina Acosta and the Center for Spatial Data Science.