Exploratory spatial data analysis with GeoDa

Spatial Data Analysis Center

2020-12-04

Contents

Introduction

Lorem Ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

It is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

website

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 1

Project Overview

Lorem Ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.1 Example Subsection1

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter ??.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

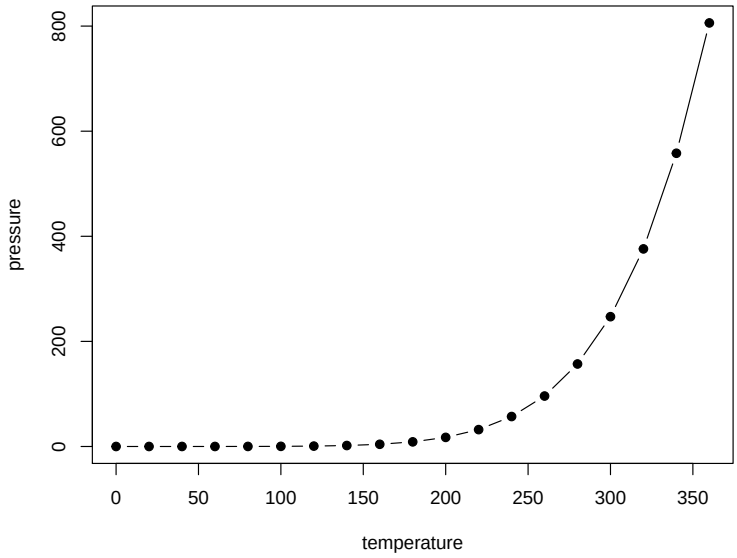


Figure 1.1: Here is a nice figure!

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2020) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 2

The Cases

2.1 John Snow and the Cholera Epidemic

How demarcation helped John Snow figure out that water caused cholera to spread in the 19th century

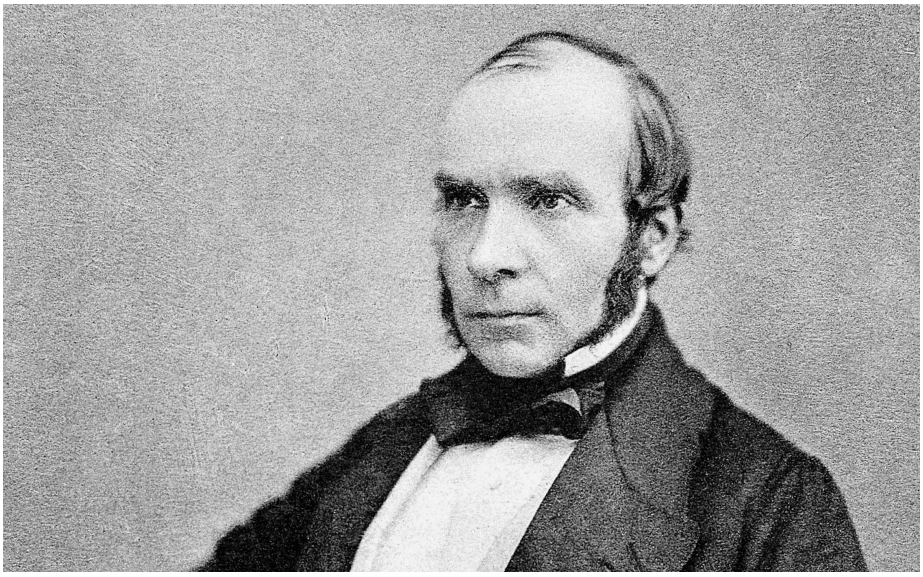


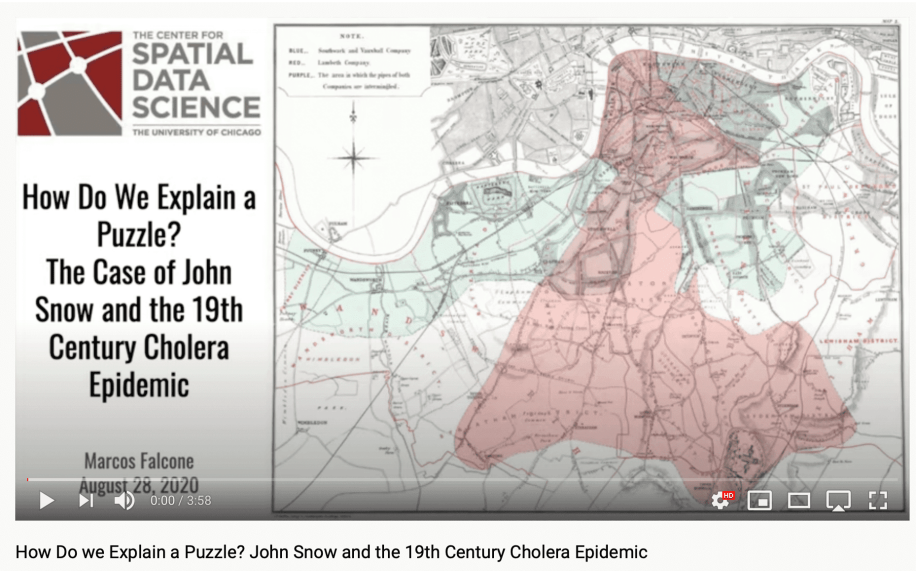
Figure 2.1: Source: Wikipedia

The Puzzle

In the mid-19th century, cholera was claiming the lives of thousands in London. But how did the disease spread? In other words, what was the main mode of

transmission of cholera?

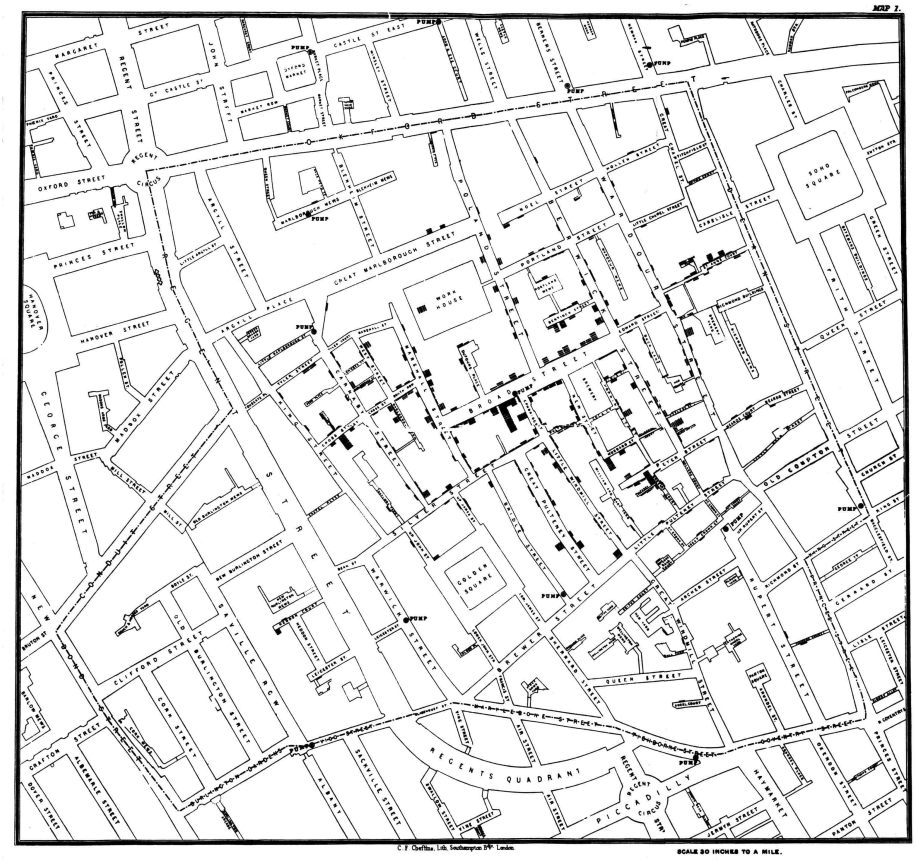
For an overview of this case, see our introductory story map and video.



How Do we Explain a Puzzle? John Snow and the 19th Century Cholera Epidemic

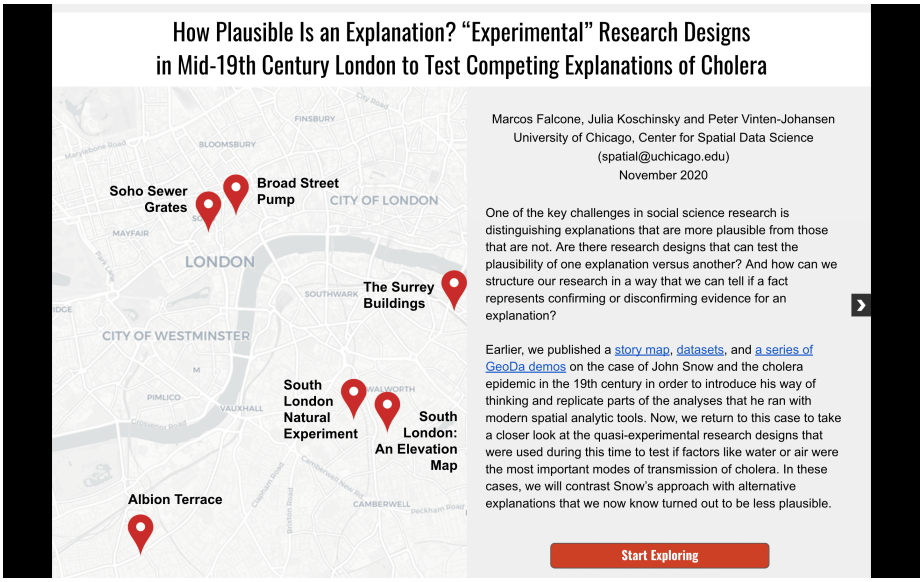
The Research Design

To answer that question, a doctor named John Snow developed a waterborne theory of cholera and then studied the locations where the disease was prevalent as well as those where it was not, along with specific locations where water suppliers varied.



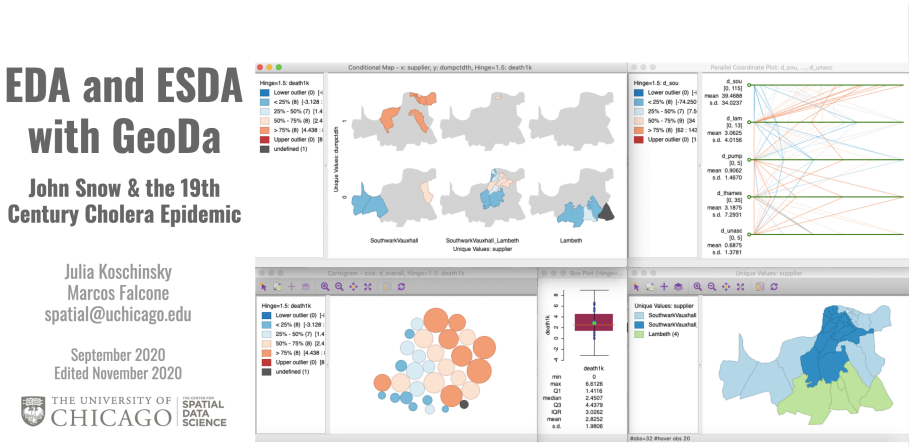
By demarcating cases in this way and testing his theory at both a micro (Soho) and a macro level (South London), Snow was able to gather evidence that was compatible with the waterborne theory of cholera but was harder to account for by the airborne theory.

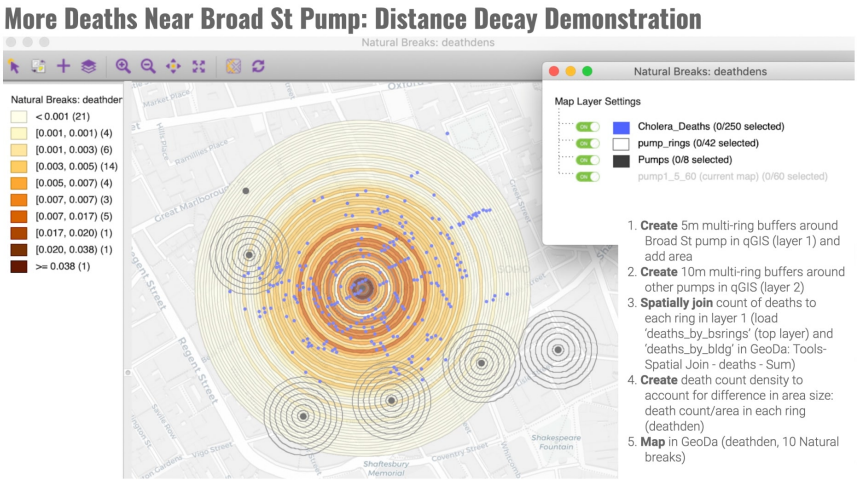
For more detail on the research designs devised by Snow and his contemporaries, see our specialized story map and video.



The Tools

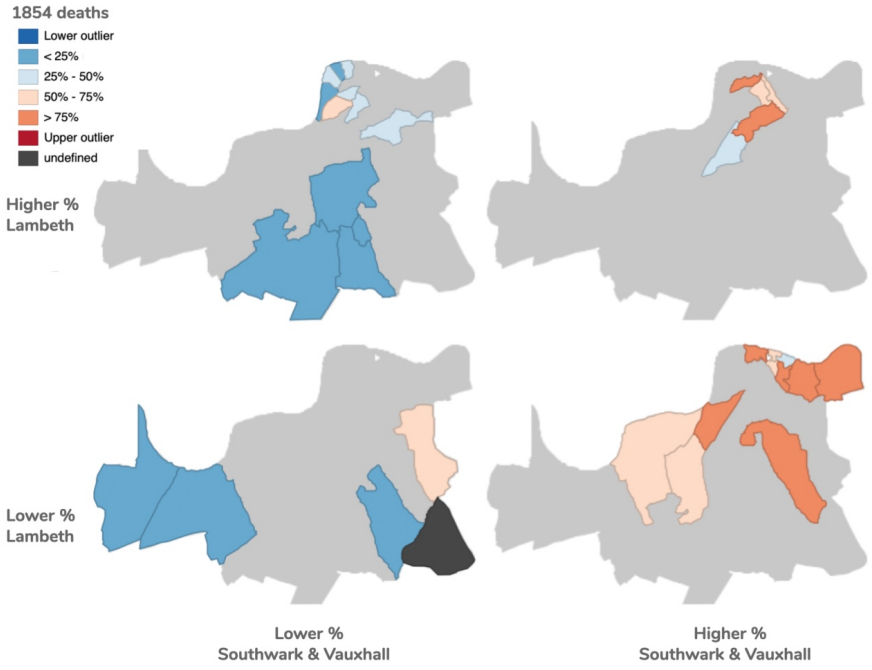
Using cluster analysis and other statistical techniques like conditional plots and averages charts, it is today possible to replicate and illustrate Snow's analyses with our GeoDa demo scripts.





The Insights

Snow used a natural experiment to find out that cholera cases concentrated in groups of people who relied on specific water supply mechanisms, whereas groups which relied on a different water supply were not affected even if they were located right next to clusters of infections.



More Information

Access our data and documentation to replicate these findings in GeoDa.

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

2.2 Sherlock Holmes and the Napoleon Busts

Finding a key common feature of the smashed Napoleon busts allowed the famous detective to solve the mystery of why they were smashed.

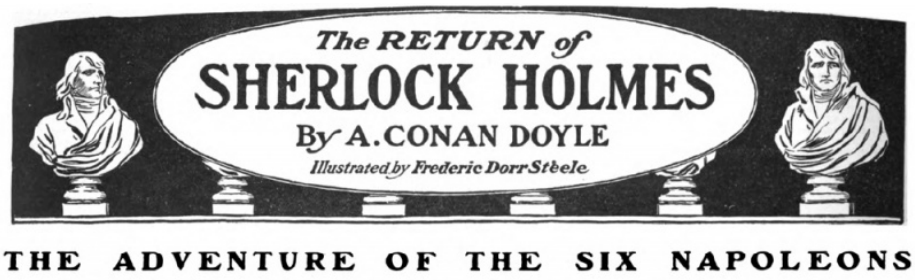
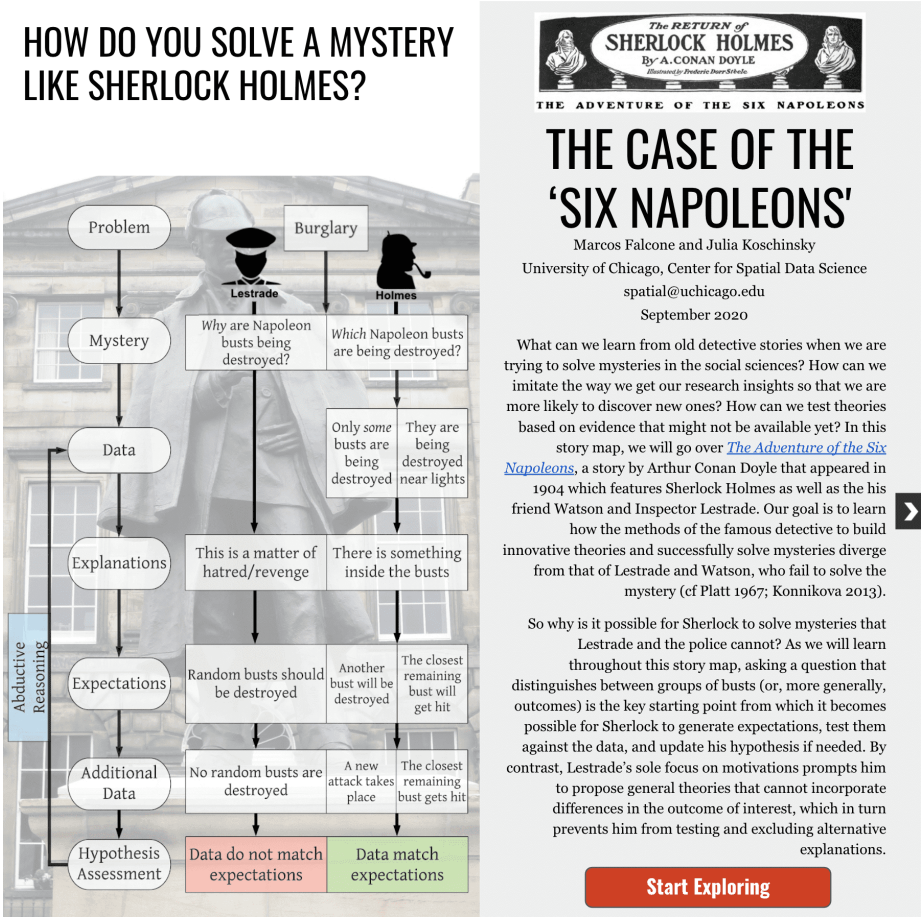


Figure 2.2: Source: The Arthur Conan Doyle Encyclopedia

The Puzzle

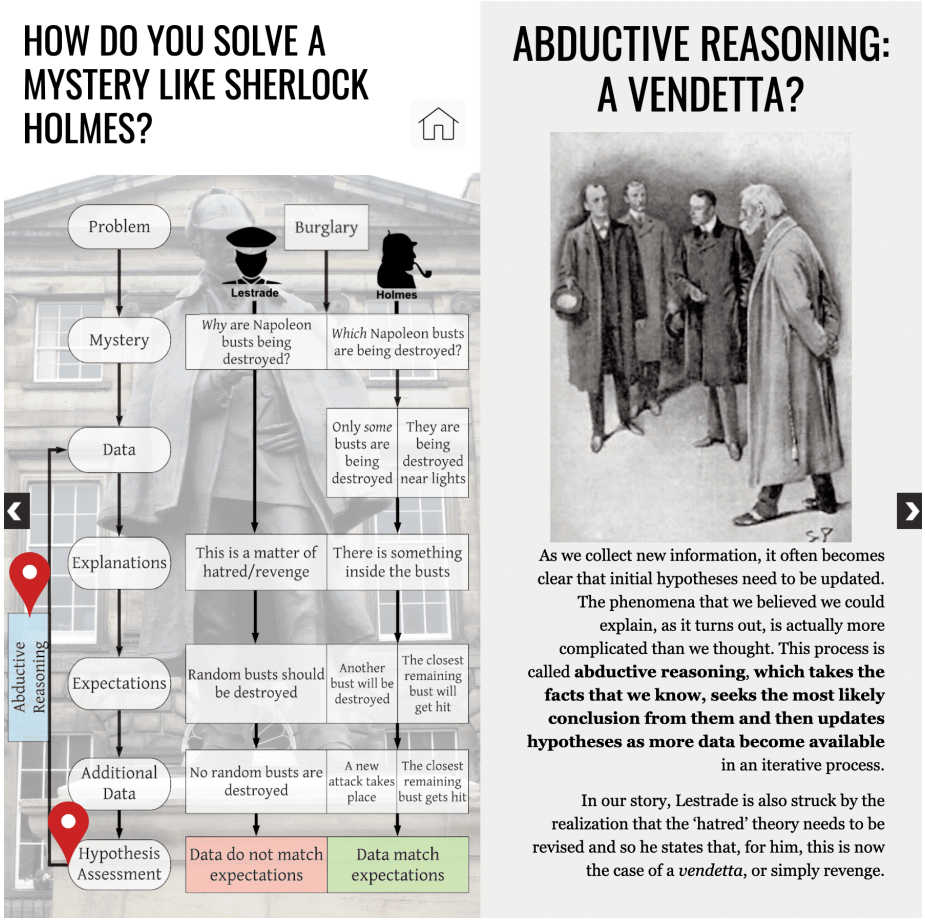
In The Adventures of the Six Napoleons, the famous detective Sherlock Holmes faced a mystery: Napoleon busts were being destroyed in private property across the city of London and nobody knew why.

For an overview of this case, see our story map and video.

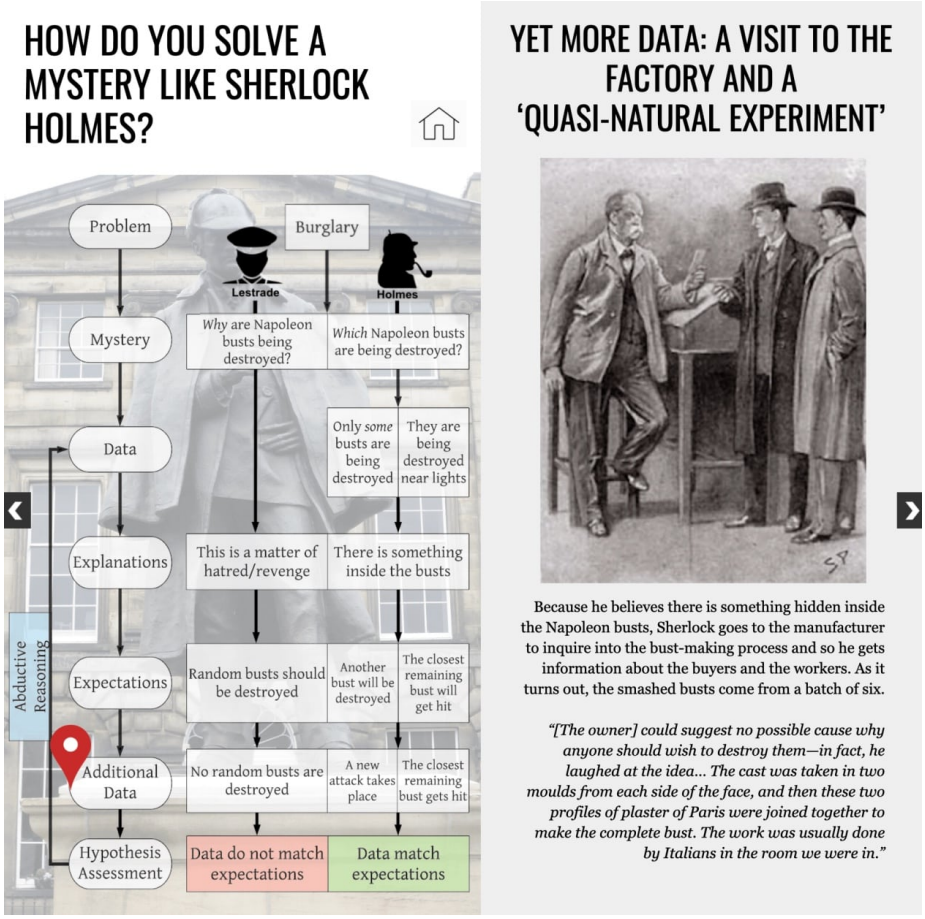


The Research Design

Policemen initially thought that hatred was causing someone to break the busts – or that they were dealing with a potential vendetta.



According to Sherlock, however, neither of these theories explained why only a specific subset of busts was being smashed. By observing the circumstances of the attacks and discovering that the busts came from the same manufacturer and the same batch, he came up with the theory that there was a hidden object inside one of these busts. He then predicted which bust would be the next to be destroyed, alerted the police and conducted a quasi-natural experiment at the location where he thought that the attack would take place.



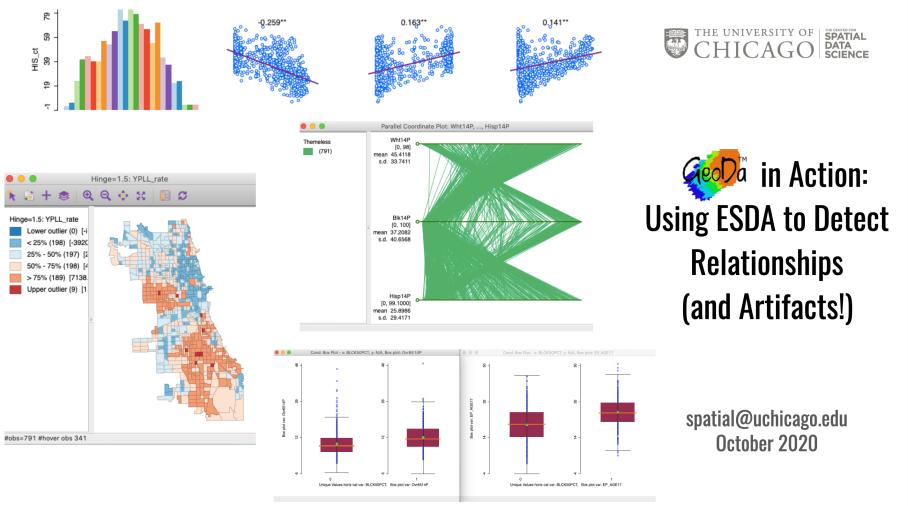
The Insights

As it turned out, Sherlock was right and the burglar was caught on site. Basing his approach on demarcation and a quasi-natural experimental setting was successful at solving the mystery.

2.3 The Immigrant Paradox

We demonstrate that health outcomes in poor immigrant neighborhoods that are better than those in nearby non-immigrant poor neighborhoods do not actually reflect the "Immigrant Paradox" theory but are likely an artifact of a younger immigrant population.

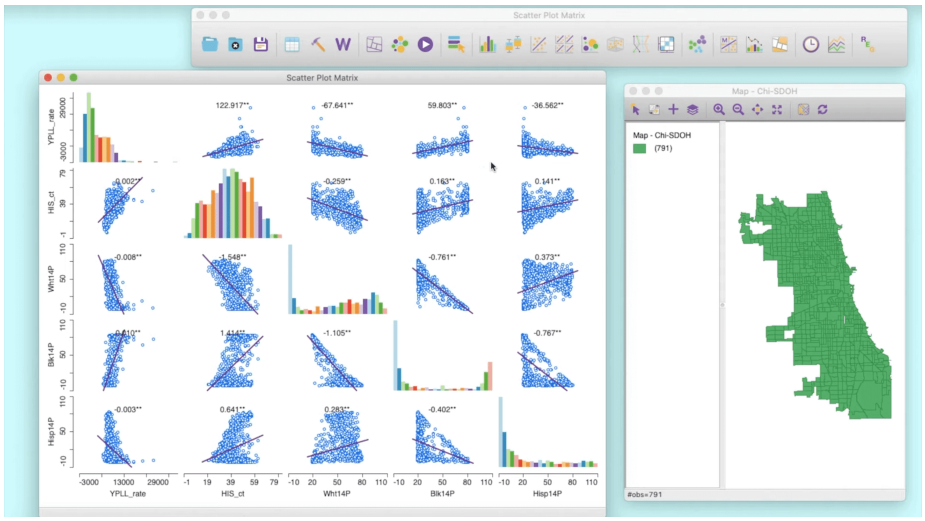
Figure 2.3: Source: Pixabay



The Puzzle

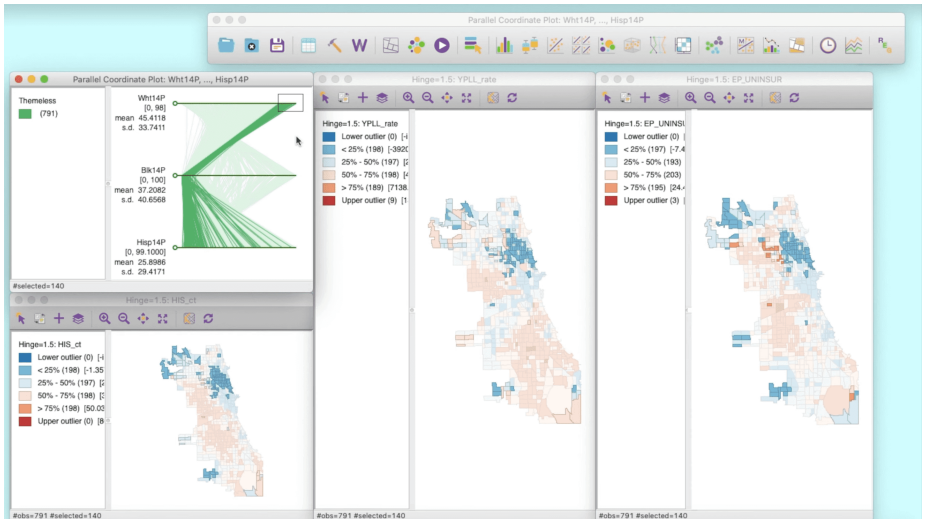
Economic hardship has long been associated with worse health outcomes, particularly with more life years lost prematurely. But, in the US, Hispanic immigrants often do better in terms of health even when they face similar socioeconomic issues. Why?

For an overview of this case, see our video here.



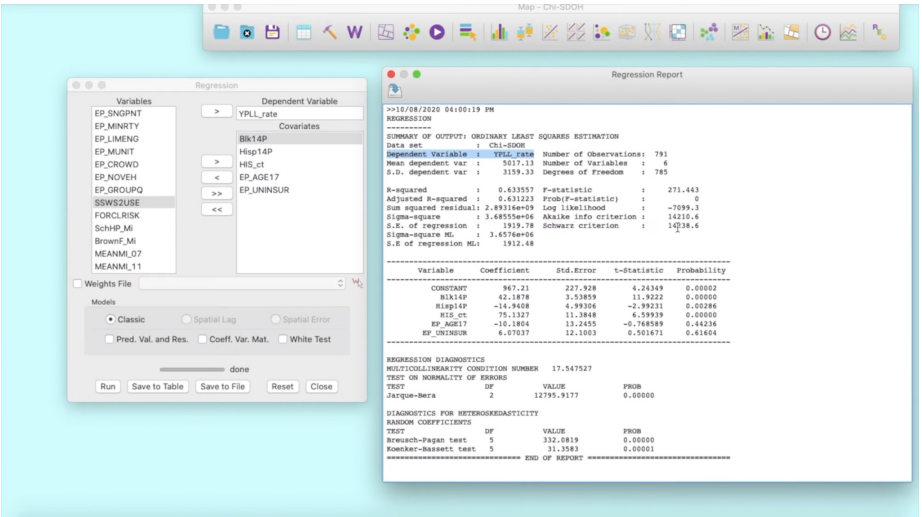
The Research Design

We compare premature mortality outcomes for different demographic groups and use a process of abductive reasoning to explore the plausibility of hypotheses that could explain these outcomes.



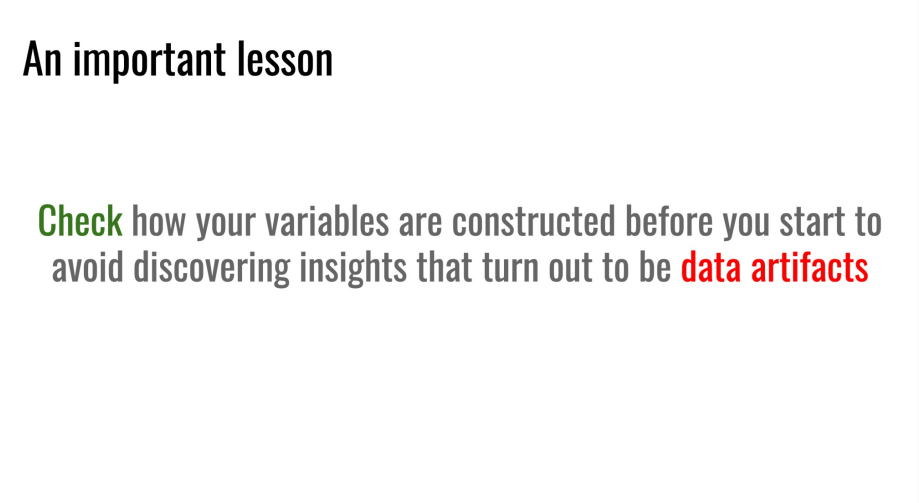
The Tools

We test these hypotheses with scatter plots, parallel coordinate plots, conditional box plots and various types of maps – and with regression analysis.



The Insights

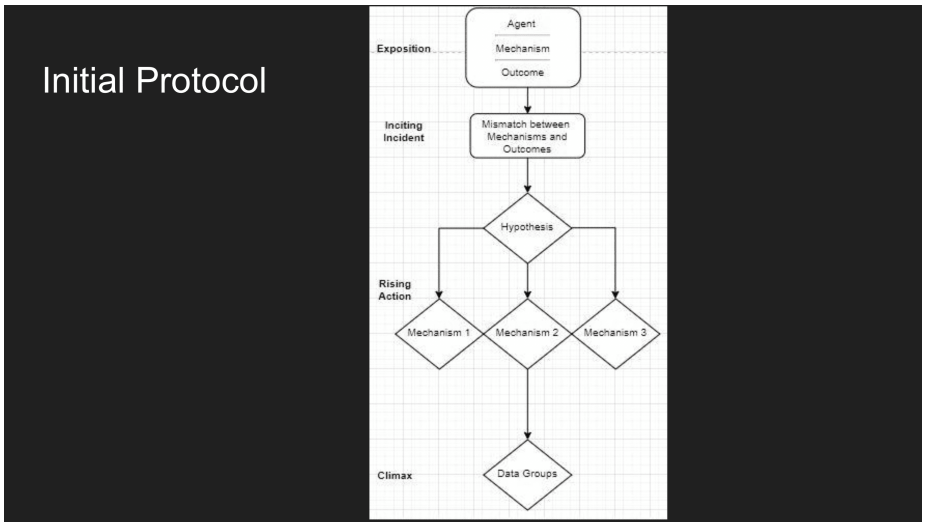
In the end, while insurance rates for different groups do not seem to explain the difference in health outcomes, younger ages in Hispanic neighborhoods do appear to be the driving factor behind lower premature mortality compared to predominantly White or African-American neighborhoods in Chicago.



2.4 Health and Race: A Preliminary Approach

Author: **Atman Mehta** (3rd year student in the College)

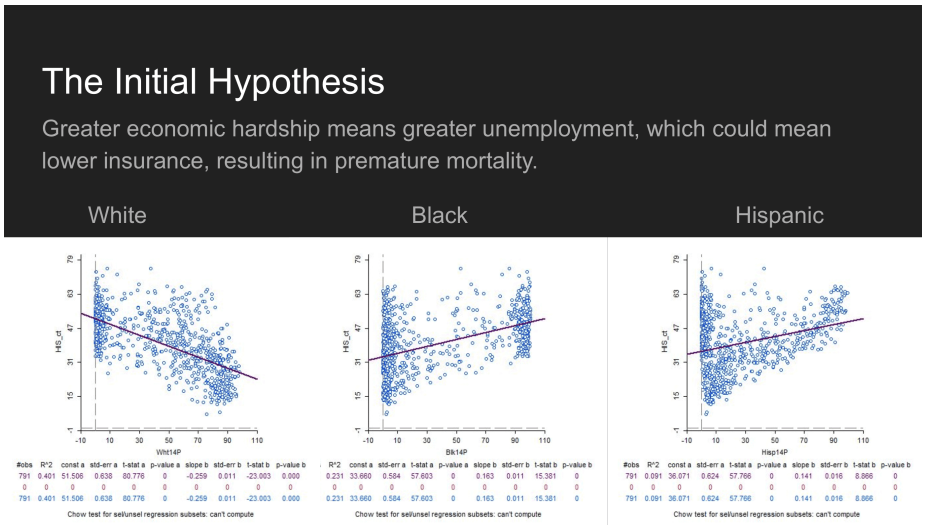
A protocol that differentiates groups based on racial majorities provides an initial assessment of potential determinants of health indicators



The Puzzle

How can we create groups to explain differences in outcomes in the social sciences? The question of the determinants of health in Chicago can be used as an example to illustrate this process.

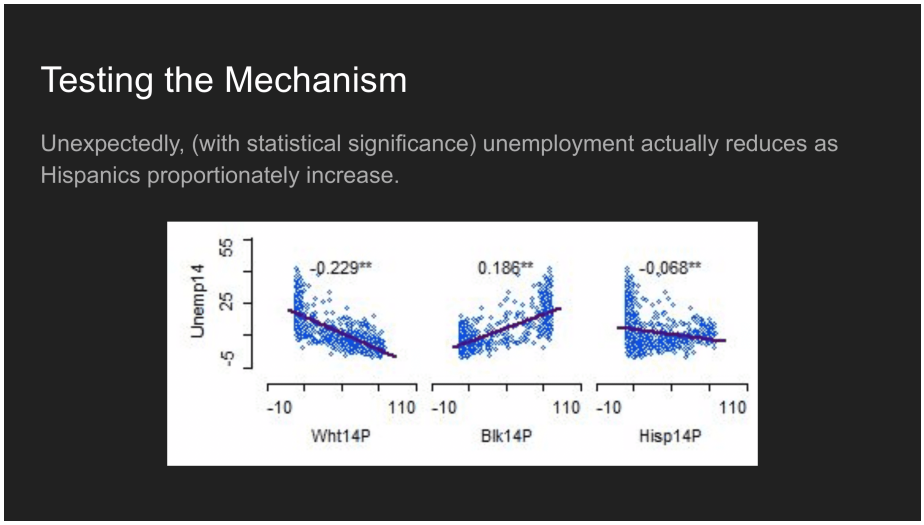
For an overview of this case, see more from our summer project here.



The Research Design

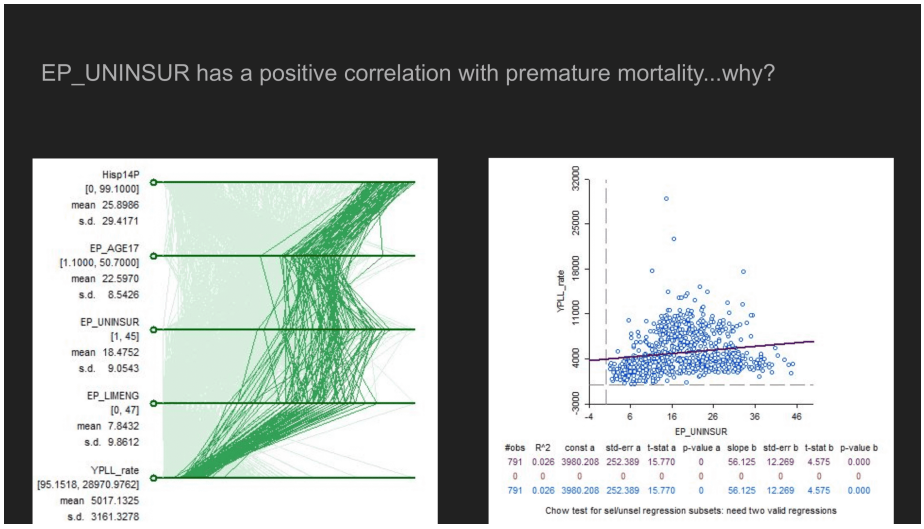
Atman first looked at potential plausible explanations and mechanisms, with the main idea being that greater economic hardship means greater unemploy-

ment, which means lower insurance and results in premature mortality. He then differentiated demographic groups according to race to test if the variables behaved in the expected ways across tracts with different racial majorities.



The Tools

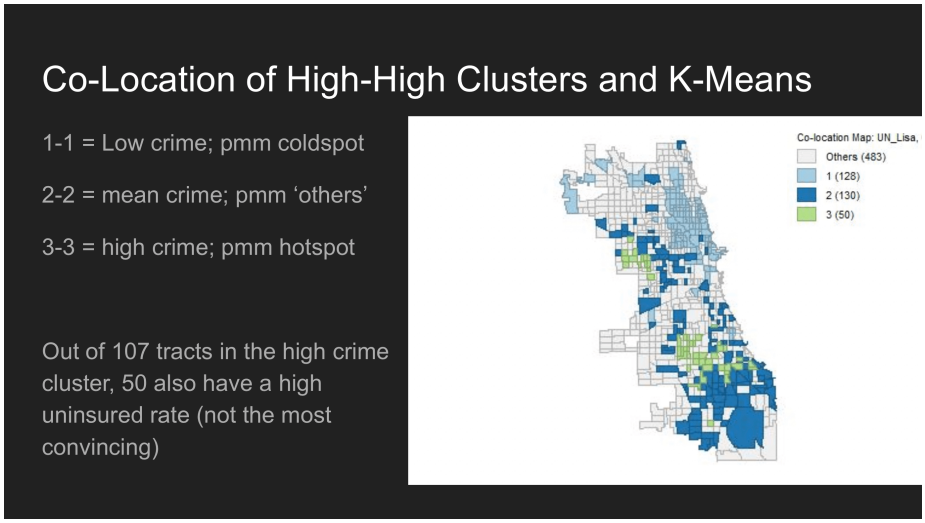
Several hypotheses were tested by using tools available in GeoDa such as parallel coordinate plots and scatter plots as well as co-location, cluster and LISA maps.



The Insights

The resulting protocol helps in structuring the problem at hand in a way that is easy to test. Substantially, dissimilarities in indicators such as premature mortality rates for different racial groups do not seem to be explained by un-

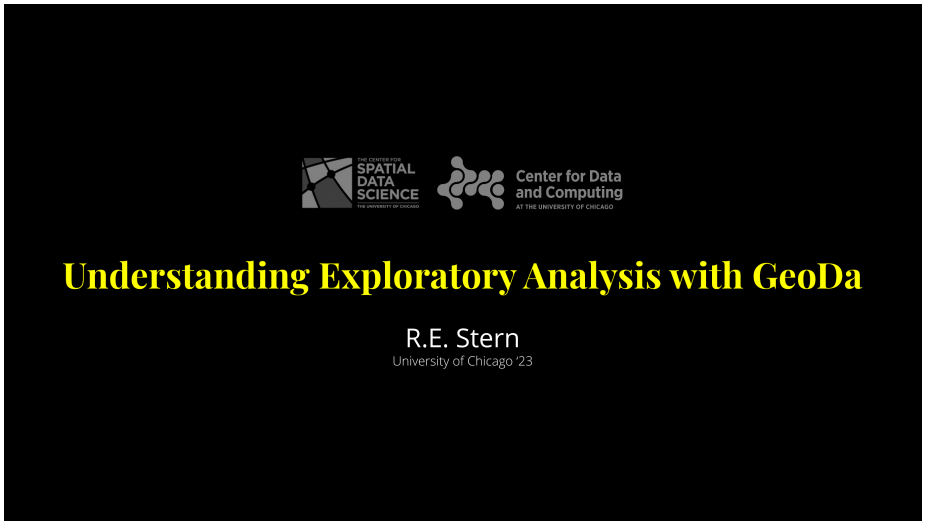
employment or violent crime. This puzzle is explored in more detail in the case of The Immigrant Paradox, also available on this website.



2.5 Turnout and Elections: A Spatial Perspective

Author: **R.E. Stern** (1st year student in the College)

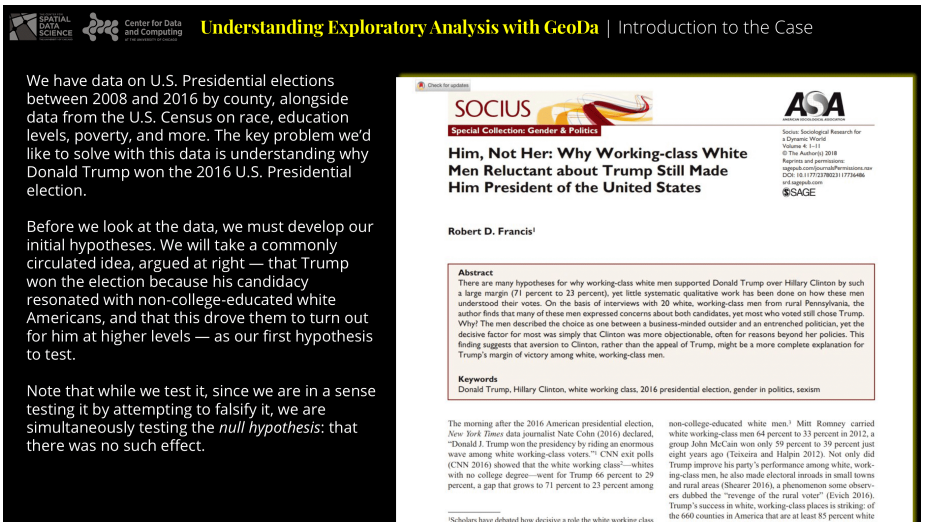
Changes in turnout in presidential elections from 2012 to 2016, which could have had an impact on their outcome, appear to be related to demographics.



The Puzzle

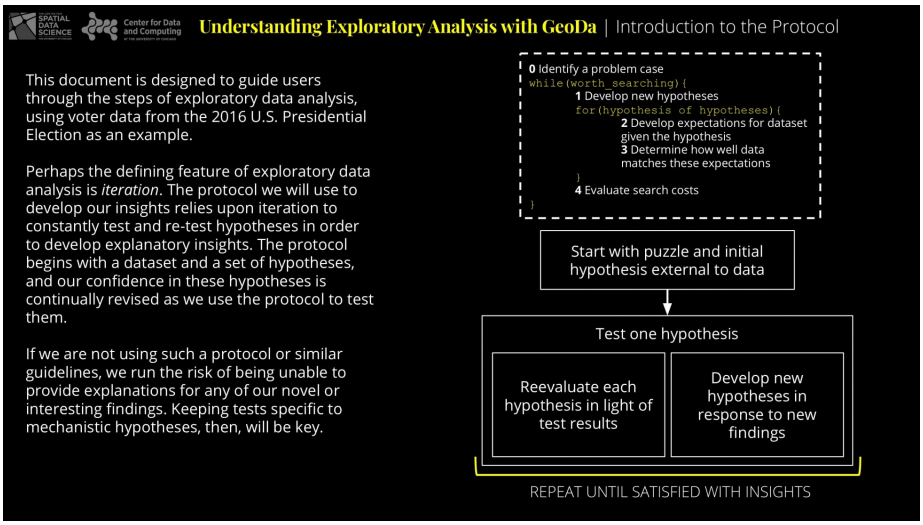
What are the factors that explain changes in turnout in the 2016 presidential election compared to previous elections?

For an overview of this case, see more from our summer project here.



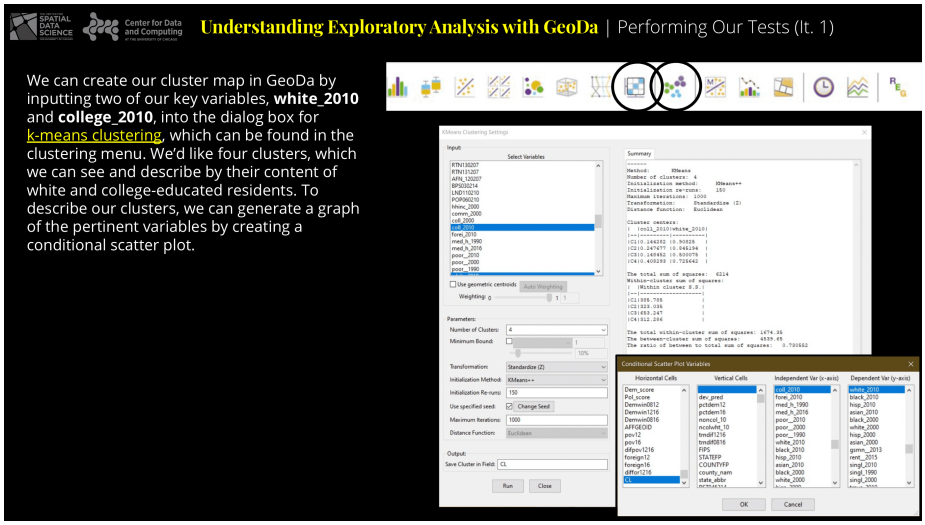
The Research Design

R.E.'s protocol highlights the importance of an iterative discovery process, i.e. the continuous testing and re-testing of hypotheses to develop explanatory insights. The hypothesis that non-college white voters turned out to vote in higher numbers in 2016 leads to expectations with an estimated probability that are then tested.



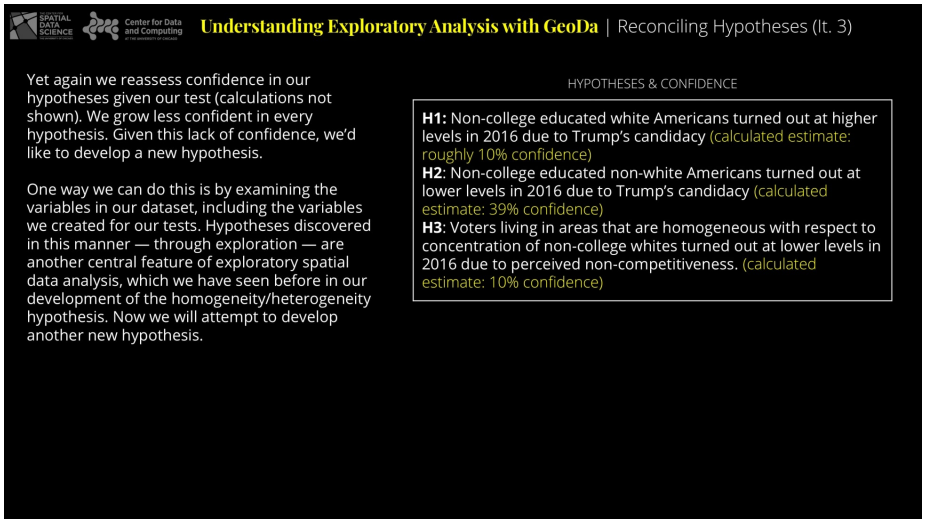
The Tools

First, exploratory spatial data analysis tools such as box maps give an overview of the data. Then, GeoDa's cluster maps such as K-Means or Local Moran's I allow users to re-assess the probabilities they assign to each hypothesis.



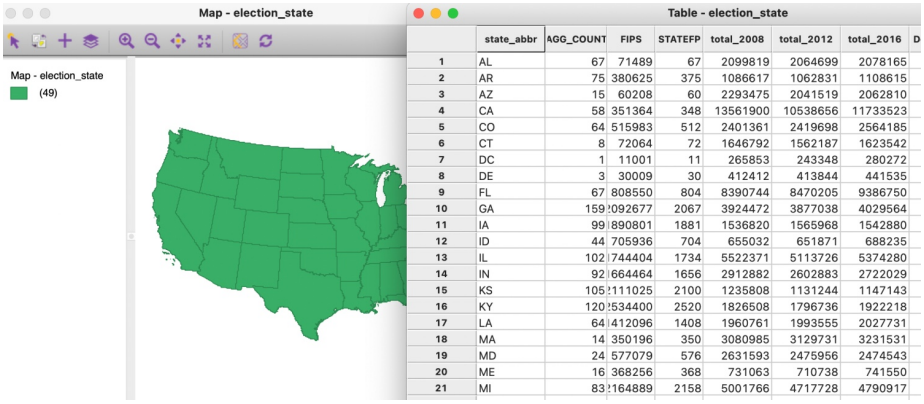
The Insights

Counties with more white residents without college degrees did see slightly higher turnout in 2016 compared to 2012 – more importantly, though, counties with more non-white residents and no college degrees saw less turnout.



More Information

Access our data to replicate these findings in GeoDa.

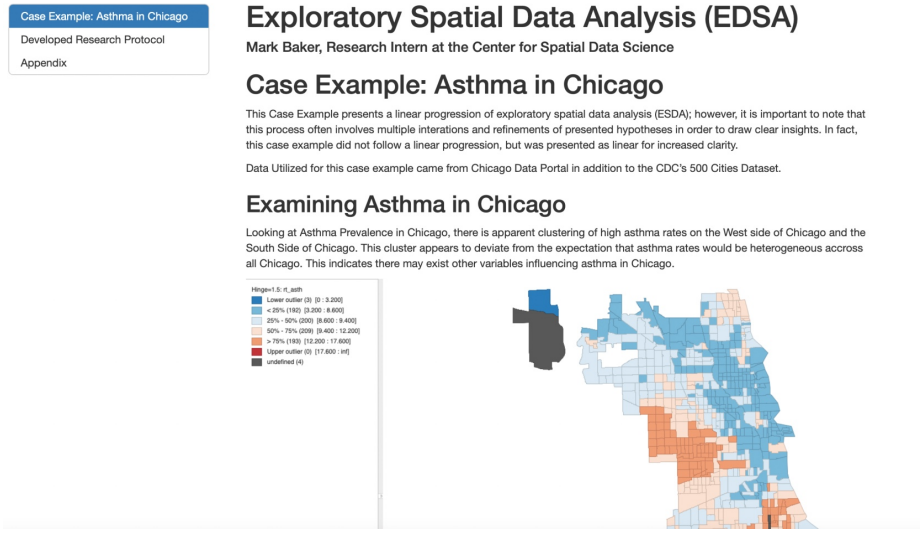


	state_abbr	AGG_COUNT	FIPS	STATEFP	total_2008	total_2012	total_2016	D
1	AL	67	71489	67	2099819	2064699	2078165	
2	AR	75	380625	375	1086617	1062831	1108615	
3	AZ	15	60208	60	2293475	2041519	2062810	
4	CA	58	351364	348	13561900	10538656	11733523	
5	CO	64	515983	512	2401361	2419698	2564185	
6	CT	8	72064	72	1646792	1562187	1623542	
7	DC	1	11001	11	265853	243348	280272	
8	DE	3	30009	30	412412	413844	441535	
9	FL	67	808550	804	8390744	8470205	9386750	
10	GA	159	092677	2067	3924472	3877038	4029564	
11	IA	99	890801	1881	1536820	1565968	1542880	
12	ID	44	705936	704	655032	651871	688235	
13	IL	102	744404	1734	5522371	5113726	5374280	
14	IN	92	664464	1656	2912882	2602883	2722029	
15	KS	105	111025	2100	1235808	1131244	1147143	
16	KY	120	534400	2520	1826508	1796736	1922218	
17	LA	64	412096	1408	1960761	1993555	2027731	
18	MA	14	350196	350	3080985	3129731	3231531	
19	MD	24	577079	576	2631593	2475956	2474543	
20	ME	16	368256	368	731063	710738	741550	
21	MI	83	164889	2158	5001766	4717728	4790917	

2.6 Asthma and Pollution

Authors: **Mark Baker** and **Jizhou Wang** (3rd year students in the College)

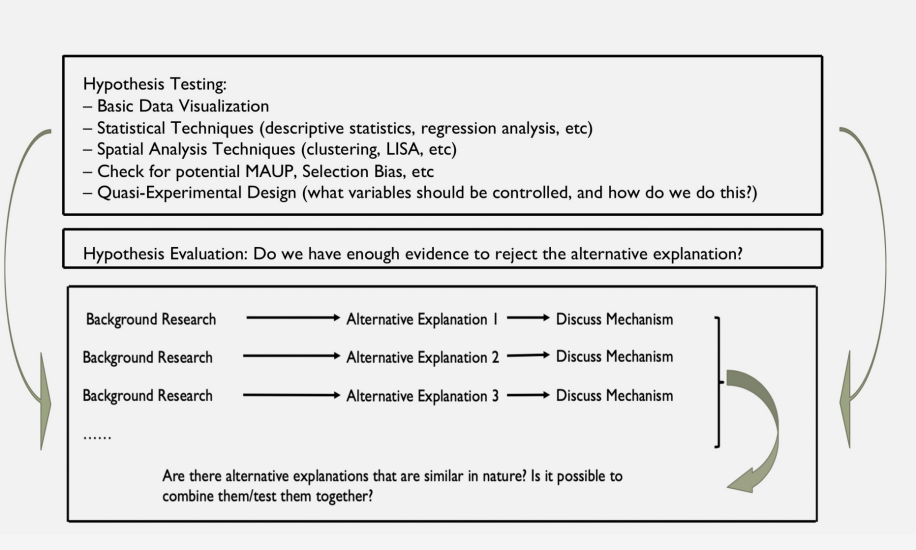
Proximity to potential bus pollution is higher in areas with more residents who are economically vulnerable and African-American, groups which are also at higher asthma risk.



The Puzzle

Are groups at higher risk of asthma also more likely to be exposed to bus pollution?

For an overview of this case, see insights and methods from our summer project.



The Research Design

Path diagrams outline what factors could be driving spatial concentrations of asthma, including pollution from traffic, lower housing values and older homes. Mechanisms for hypotheses are developed and tested to assess their strength based on a classification by Mark Baker.

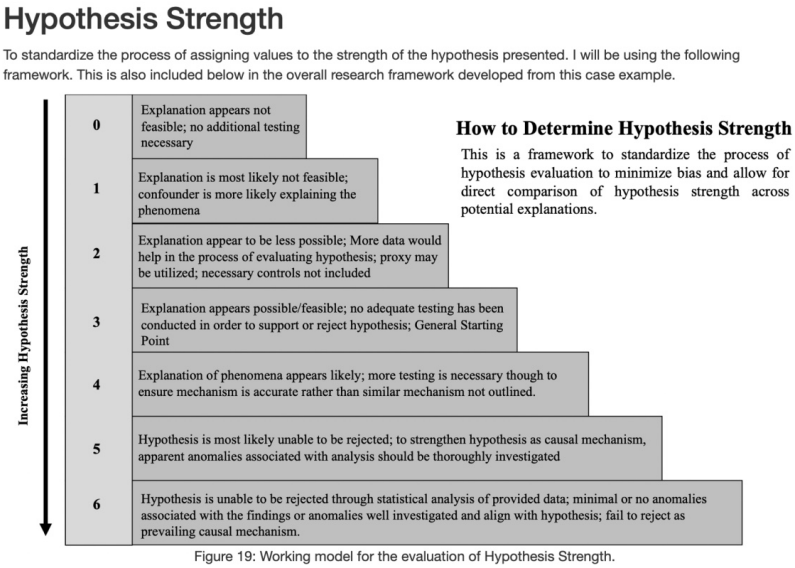
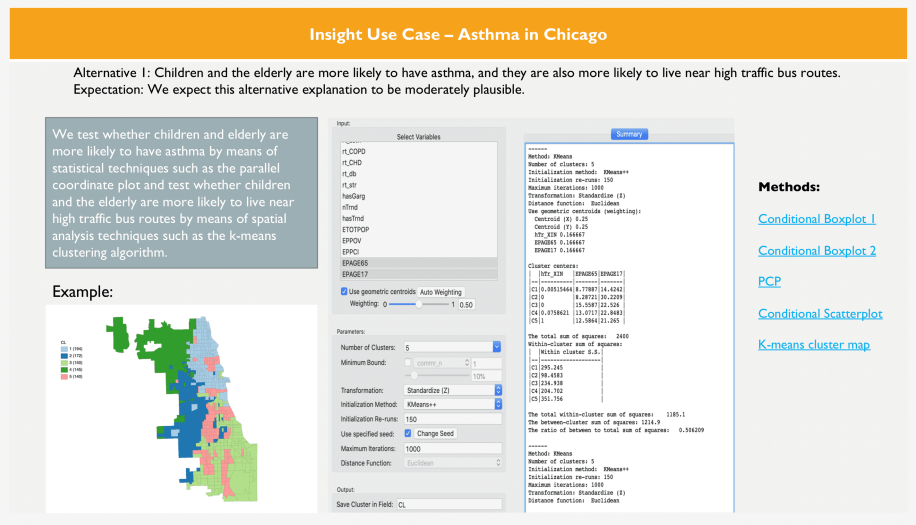


Figure 19: Working model for the evaluation of Hypothesis Strength.

The Tools

Mark Baker and Jizhou Wang assess the strength of different hypotheses with

boxplots, scatterplots, parallel coordinate plots, cluster analyses, averages charts and regressions in GeoDa.



The Insights

Economically vulnerable and African-American neighborhoods that have a higher asthma risk are also more likely to be exposed to traffic pollution.

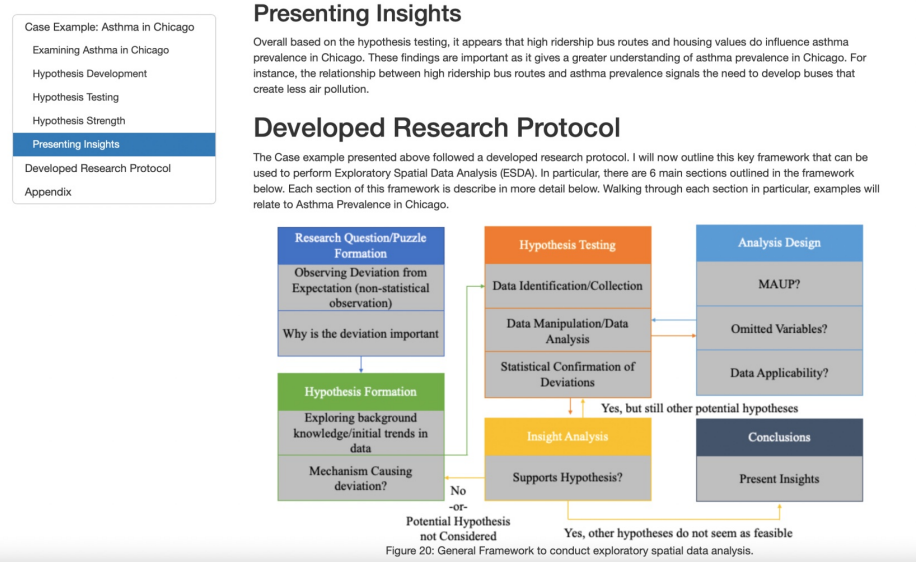
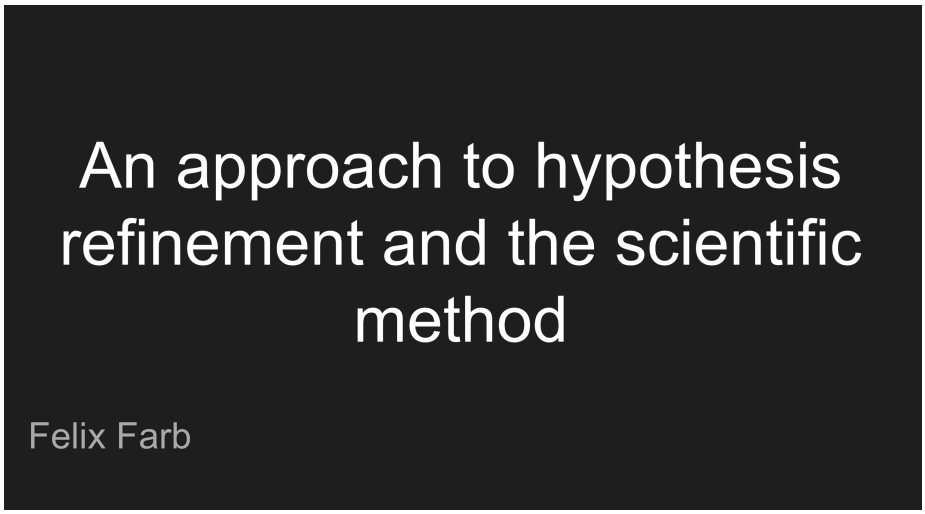


Figure 20: General Framework to conduct exploratory spatial data analysis.

2.7 Racial Diversity: What Built Environment Features Distinguish Racially Diverse from Non-Diverse Areas?

Author: **Felix Farb** (Junior High School student)

Planning-related factors like highway dividers and land use diversity, as well as out-migration of African-American residents, distinguish racially diverse from non-diverse areas in Chicago.



The Puzzle

Do racially diverse areas in the city of Chicago differ in terms of planning-related factors from non-diverse areas?

For an overview of this case, see insights and methods from our summer project.

Question: Why do racially diverse areas in the city exist where they do?

Comparing: Diverse areas to their non Diverse neighbors.

This question is interesting, but too broad to be answered on its own. Let's break this question down into multiple more specific possible causes.

The Research Design

An exploratory spatial data analysis identifies areas in Chicago that are racially diverse and compares potential planning-related drivers such as highway dividers and diverse land uses to gentrification. Felix's protocol demonstrates how hypotheses can be made more and more specific to make them testable.

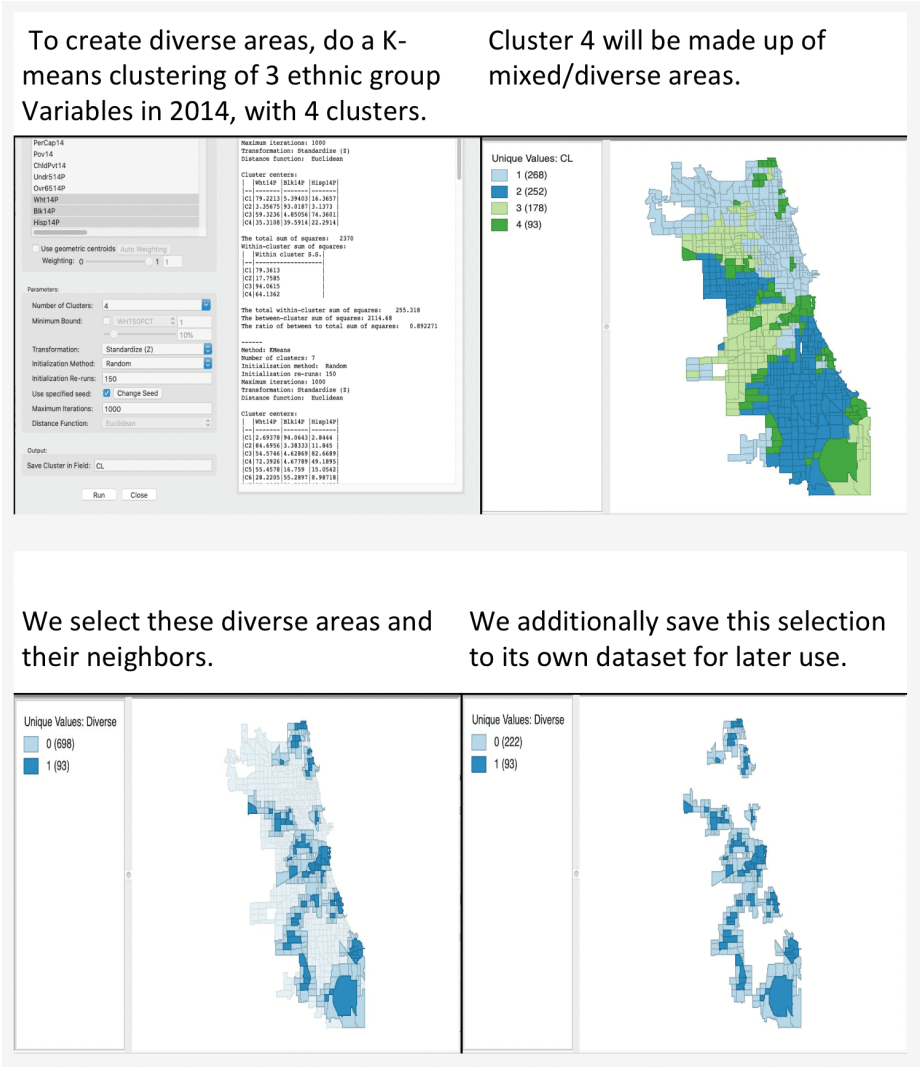
H1	**diverse (D) areas are more likely to form near non-diverse (ND) ones when there's a physical divider like a highway between them**
definition	
variable measurement	
data exploration test	
result	
result significance	
New H1	
H2	**Diverse areas are created by and correlated to, the level of diversity of land use in those areas**
definition	
variable measurement	
data exploration test	
result	
result significance	
New H2	
H3	**Diverse areas are created by change in the racial makeup of areas through gentrification and other processes.**
definition	
variable measurement	
data exploration	
result	
result significance	
New H3	
H4	**proximity of D vs ND areas is due to the diverse areas being transition zones with demographics made up of a combination of their neighbors**
definition	
variable measurement	
data exploration	
result	
result significance	
New H4	

These "causes" or hypotheses are direct, but we can still make them more precise, so from here we should follow a framework and do research in order to turn these into hypotheses that aim to answer a specific question.

Why should we do this?
It's important to be constantly questioning, testing, and refining your hypotheses during research. The research process should be constantly iterative to its core.

The Tools

Different maps, distance buffers, averages charts and K-Means clusters are used to assess the relevance of the results to each hypothesis.



The Insights

This initial exploratory analysis suggests that racially diverse neighborhoods in Chicago are more likely to be physically divided from other neighborhoods through highways and have more diverse land uses. Some diverse areas are also characterized by disproportionate out-migration of African-American residents.

New H1	**Highways and possibly public transport can create diversity when they converge many areas of the city upon one another.**

New H2	**Features such as Universities, Museums, and parks create diversity by attracting residents from backgrounds more reflective of the city as a whole.**

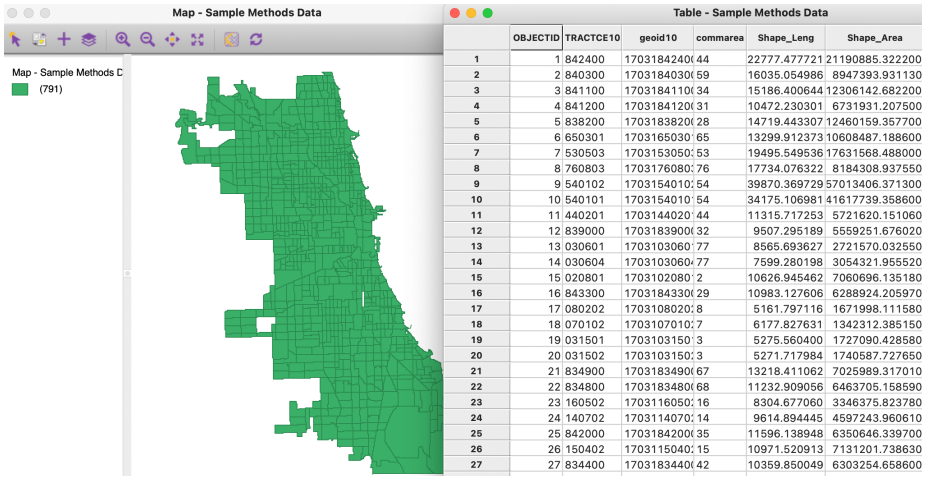
New H3	**Some areas that are diverse have become diverse due to an abnormal decrease in black population in these areas.**

New H4	**Diverse areas can be created at boundaries that are fuzzy and not physically separated, between different racially homogenous neighborhoods.**

Through this refinement process, we have created 4 supported, targeted theories. From here we can finalize our results through more rigorous data testing, or focus in on one of these theories to dig into more deeply, possibly coming up with new theories and results.

More Information

Access our data to replicate these findings in GeoDa.



	OBJECTID	TRACTCE10	geoid10	commarea	Shape_Leng	Shape_Area
1	1	842400	1703184240(	44	22777.477721	21190885.322200
2	2	840300	1703184030(	59	16035.054986	8947393.931130
3	3	841100	1703184110(	34	15186.400644	12306142.682200
4	4	841200	1703184120(	31	10472.230301	6731931.207500
5	5	838200	1703183820(	28	14719.443307	12460159.357700
6	6	650301	1703165030	65	13299.912373	10608487.188600
7	7	530503	1703153050	53	19495.549536	17631568.488000
8	8	760803	1703176080	76	17734.076322	8184308.937550
9	9	540102	1703154010	54	39870.369729	57013406.371300
10	10	540101	1703154010	54	34175.106981	41617739.358600
11	11	440201	1703144020	44	11315.717253	5721620.151060
12	12	839000	1703183900(	32	9507.295189	5559251.676020
13	13	030601	1703103060	77	8565.693627	2721570.032550
14	14	030604	1703103060	77	7599.280198	3054321.955520
15	15	020801	1703102080	2	10626.945462	7060696.135180
16	16	843300	1703184330(	29	10983.127606	6288924.205970
17	17	080202	1703108020	8	5161.797116	1671998.111580
18	18	070102	1703107010	7	6177.827631	1342312.385150
19	19	031501	1703103150	3	5275.560400	1727090.428580
20	20	031502	1703103150	3	5271.717984	1740587.727650
21	21	834900	1703183490(	67	13218.411062	7025989.317010
22	22	834800	1703183480(	68	11232.909056	6463705.158590
23	23	160502	1703116050	16	8304.677060	3346375.823780
24	24	140702	1703114070	14	9614.894445	4597243.960610
25	25	842000	1703184200(	35	11596.138948	6350646.339700
26	26	150402	1703115040	15	10971.520913	7131201.738630
27	27	834400	1703183440(	42	10359.850049	6303254.658600

References

”

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21.