Introduction to Spatial Data Science: Final Project

Jesus Cantu Jr.

March 20, 2020

"…by studying death we gain capacity to understand life as well as to protect it." (Noguchi & DiMona, 1985)

Introduction

In July of 1995, the midwestern United States experienced a severe heat wave, for which, over the course of just a single week, the city of Chicago, Illinois, documented about 514 heat-related deaths and 696 excess deaths [2], a baseline mortality rate of about 72 individuals per day—among the highest in the history of the nation. Following the legacy of Durkheim and Mauss, Eric Klinenberg (1999) carried what he called a "social autopsy" of this disaster by combining elements of urban sociology, which highlights the importance of political and economic power in a city's socio-spatial organization, with those of the Chicago School, whose approach tried to account for the geography of urban vulnerability. In doing so, his analysis of the heat wave helps to explain the mortality patterns observed across the city beyond what can be officially attributed to the weather. Patterns that clearly reflect the inequalities of Chicago's built infrastructure.

Throughout the 1995 heat wave "geography was linked to destiny" [3]. As *Figure 1* illustrates, the Chicago community areas with the highest death rates lie on the south and southwest sides of the city. Of the fifteen community areas with the highest mortality rates during the heat wave, eleven contained high proportions of people living below half of the federal poverty line[1] and ten are home to populations between 94-99% African-American (see *Table 1*); city regions which have been historically marginalized. Klinenberg (1999) notes a strong positive relationship between death rates among community areas and homicide rates. High mortality rates were also strongly associated with the number of seniors living alone, as deaths were more prevalent among low-income, elderly, African-Americans. Judith Helfand's (2019) film, *Cooked: Survival by Zip Code,* captures the life and death stories of these individuals, many of whom spent their last moments alone, confined to small, single-bedroom apartments, too afraid of neighborhood crime to even open a window and far-removed from the safety networks allocated to the rest of [WHITE] America in times of disaster to receive any immediate assistance.

The purpose of this research project is to survey the city's socio-economic environment, 25 years after the tragedy, for changes in structure against the backdrop of current population health statistics. Do current mortality rates follow the same patterns as observed by Klinenberg (1999)? If so, what can modern data tell us about people's living conditions and their capacity to deal with future disasters in the face of an ever-changing climate? In the wake of the 2020 COVID-19 (coronavirus) pandemic, and the inevitable spread of the virus across the U.S., it is of outmost importance to identify areas and individuals who are at increased risk of suffering a severe infection, particularly as a result of complications with other chronic illnesses or auto-immune disorders. These susceptible individuals should be prioritized when allocating medical/financial assistance by state and local governments. Ultimately, I hope the results of this preliminary spatial analysis can inspire further research into the transmission dynamics and control of COVID-19 across different geographic regions.

[1] About $15,150 for a four-person family in 1995 [4].

Data

The City of Chicago Data Portal (https://data.cityofchicago.org/) holds a rich repository of data regarding the city's infrastructure. Datasets can be downloaded for free and are available, at different organizational levels, in a variety of categories ranging from administration and finance information to sanitation and public safety requests. Here, we will be focusing on a dataset that contains information on selected underlying causes of death in Chicago.[2] The data includes the cumulative number of deaths, the average number of annual deaths, the average annual crude and adjusted rates with corresponding 95% confidence intervals, and the average annual years of potential life lost per 100,00 residents aged 75 and younger due to selected causes of death (YPLL)[3] by Chicago community area for the years 2006 to 2010. A ranking for each measure is also provided, with the highest value indicated with a ranking of 1.

Age-adjustment was done using a slight modification of the direct method to the 2000 US standard million population were the less than 1 year-old and 1 to 4 year-old groups are combined. Rates are expressed per 100,000 population. Age-specific population estimates by community area were obtained through a linear interpolation of counts from the 2000 and 2010 Census. Because the age-adjusted death rate is a mortality rate that controls for the effects of differences in population age distributions, it is useful for comparing data over time and by geographic area. For this study, we will be using this measure as our main dependent variable while paying particular attention to the following causes of death: coronary heart disease, diabetes-related, and stroke as they have been found to increase the risk of severe complications from coronavirus infection among individuals [6-9]. Data from deaths resulting from unintentional injury was also included as a possible control; however, this assumes that most of these types of deaths are a result of chance events. Spatially, its distribution should be random (i.e., null hypothesis). We should also expect there to be little association between this mortality rate and our independent variables.

To explore the city's socio-economic environment, I am using a dataset that contains a selection of six socioeconomic indicators of public health significance and a "hardship index" by Chicago community area for the years 2008 to 2012. The variables include: the percent of occupied housing units with more than one person per room (i.e., crowded housing), the percent of households living below the federal poverty line, the percent of persons aged 16 years and older in the labor force that are unemployed, the percent of people aged 25 years or older without a high school diploma, the percent of the population under 18 or over 64 years of age (i.e., dependency), and per capita income. The Chicago Department of Public Health (CDPH) calculated the indicators using census tract-level estimates obtained from the 2008-2012 American Community Survey 5-year estimates.[4] A community area's per capita income was estimated by dividing aggregate income of the census tracts within community area by the number of residents. The hardship index is a score that incorporates each of the six socio-economic variables according to the methods described in [10].

---

[2] An "underlying cause of death" is defined as the main disease, accident, or injury that caused the death.

[3] The years of potential life lost rate (YPLL) is a summary of years lost due to premature death per 100,000 population at or below 75. In contrast to mortality rates, YPLL emphasizes the effect of premature mortality on a population. YPLL is the sum of the differences between a predetermined end point (average life expectancy) and the ages of death for those who died before that end point, divided by the total population at or below that end point, and multiplied by 100,000.

[4] According to U.S. Census Bureau 2008-2012 American Community Survey 5-year estimates, 3.2% of occupied housing units in the U.S. had more than one person per room; 10.9% of households were living below the federal poverty level; 9.3% of persons aged 16 years or older in the labor force were unemployed; 14.2% of persons aged 25 years or older did not have a high school diploma; 37.2% of the population was under 18 or over 64 years of age; and per capita income was $28,051.

Scores on the index can range from 1 to 100, with a higher index number representing a greater level of hardship. Note, that these scores are standardized for the 77 community areas, and cannot be compared to scores generated from data of other jurisdictions.

As a proxy to survey individual's access to healthcare, I calculated the relative distance in kilometers from each community area's centroid to the nearest hospital (see *Figure 2*). A list of hospitals with their corresponding addresses was obtained from CPH; the list was last updated on August 28, 2011. Precaution was taken to remove institutions that would not likely treat individuals suffering from COVID-19 infections (e.g., rehabilitation centers). This process generated a list of 39 institutions, for a complete list see the *Appendix*. Addresses were geocoded using the 'ggmap' package in R. Distances to the nearest hospital were calculated using the Haversine formula, which determines the great-circle-distance between two points on a sphere given their longitudes and latitudes. It should be noted that this method, as implemented using the 'geosphere' package, assumes a spherical earth, ignoring ellipsoidal effects. Unless specified, all other data wrangling and analysis was carried out using R (version 3.6.1).

Exploratory Data Analysis

Descriptive statistics for our selected underlying causes of death are shown in *Table 2*. For 2006 to 2010, coronary heart disease appears to have the highest average annual, age-adjusted, death rate amongst community areas in Chicago.[5] It is then followed by stroke, unintentional injury, and diabetes-related deaths. For the time period, no outliers were identified among diabetes-related mortality rates, but were present for all other underlying causes (see *Figure 3*). *Figure 4* shows a scatter plot of mortality rates vs. YPLL. In the plot, every dot represents a community area while each color represents a different underlying cause of death. Because of the way in which YPLL is calculated, this measure gives more weight to a death the earlier it occurs. It appears, then, that unintentional injury resulting in death is occurring at a higher proportion within younger individuals compared to the other causes, which tend to afflict older populations. From this plot, we can also observe more of a cloud-like pattern for deaths from coronary heart disease; there is more variability in the average annual mortality rates for this underlying cause of death between community areas compared to the others.

Summary statistics for our independent variables are shown in *Table 3*. In order to reduce the number of independent variables, and to minimize issues of multicollinearity later in regression analysis, correlation between independent variables was tested using the Pearson correlation method. The results are visualized in a correlogram (see *Figure 5*). Most variables, with the exception of per capita income, are positively correlated with each other. Of significance, is the hardship index which appears to share a moderate to strong uphill linear relationship with most socioeconomic variables, and per capita income which shares the opposite. From here on out, the analysis focuses on these two variables and our proxy for access to healthcare services (i.e., kilometer distance to nearest hospital

[5] From here on out, any reference to mortality rates will always be average annual, age-adjusted, rates.

Geo-visualization

Mapping was used to describe the spatial aspects of the data and to assess potential spatial heterogeneity in our variables. Box maps are shown in *Figures 6-7*, a box map is an augmented quartile map with an additional lower and upper category. When there are lower outliers, for example, the starting point for the breaks will be the minimum value for that variable, and the second its lower fence. In contrast, when there are no outliers, then the starting point will be lower fence, and the second break its minimum value. The same method is applied at the upper end of the distribution.

Visually, we can detect some heterogeneity in mortality rates across community areas in Chicago for each of our underlying causes of death (see *Figure 6*). Highest mortality rates are concentrated within the south and southwest regions of the city, particularly so for stroke and diabetes-related deaths. Surprisingly, these regions also have high indices of death as a result of unintentional, injury which we expected to have a more random distribution across the city. The appearance of spatial heterogeneity is more pronounced among our main independent variables, where there is also more evidence of clustering of like values (see *Figure 7*). For per capita income, there is almost an incline seen across the city going from the north to the south side— rich to poor. The opposite occurs for the hardship index which increases for community areas as you travel southwards. For these regions in the south, the larger presence of hospitals in the upper side of Chicago makes most distances to the nearest hospital seem almost like outliers.

Results from multivariable regression analysis are shown in *Tables 4-5*. Because of issues of collinearity, per capita income and the hardship index had to be fitted separately. Both of these variables are significantly associated with mortality rates ($p < 0.01$), across each of the underlying causes of death. As they are, however, the models can at best explain about 37% of the variance in mortality rates for some of the datasets, with those that include the hardship index showing more explanatory power. Further analysis will require more comprehensive variables at the community level.

Spatial Autocorrelation Analysis

Tests for spatial autocorrelation were carried out in GeoDa (version 1.14.0), as this program provided an easier way to implement such analysis using Moran's I. For each underlying cause of death, global spatial autocorrelation was examined across our variables by creating a univariate, Moran scatter plot (e.g., *Figure 8*) with significance being assessed by means of a permutation test. Queen-based contiguity was chosen as a weights measure between the polygon figures (i.e., community areas) after sensitivity analysis. This measure resulted in a minimum and maximum of 7 and 9 neighbors, respectively (with a mean of 5). Among our global search, results show the appearance of significant clustering for all our dependent and independent measures (pseudo p-value < 0.001).

Identification of clusters and the assessment of their significance was carried out by computing a local, univariate Moran statistic and analyzing the associated significance map and cluster map (e.g., *Figure 9 & 10*), with queen-based contiguity being kept as the choice of weights. Results for the analysis of our three socioeconomic variables show positive spatial autocorrelation or clusters of like values. In their Moran scatter plots (not shown), clusters were found to lie within the upper right and lower left quadrants, indicating low-low (L-L) and high-high (H-H) values relative to the mean. Brushing of the scatter plot helped identify clusters shared between our socioeconomic variables, with the following community areas sharing the same qualities among the hardship index and the measure of per capita income (pseudo p-values < 0.001): Lincoln Park (H-H), Near North Side (H-H), New City (L-L), Englewood (L-L). This process was repeated, for each underlying cause

of death, to identify patterns of mortality rates across the city. The following clusters were observed among all causes of death (pseudo p-values < 0.01): Greater Grand Crossing (H-H), Pullman (H-H), Fuller Park (H-H), Englewood (H-H), and Lincoln Square (L-L). Of particular significance are Fuller Park, Grand Crossing, and Englewood which were noted by Klinenberg (1999) as community areas to have had experienced some of the highest heat-related deaths during the 1995 Chicago heat wave. It appears that these areas continue to be afflicted by poor socioeconomic conditions (i.e., lower per capita income and a higher hardship index). This study reveals that these areas may also lack proper healthcare infrastructure, in terms of access to nearby hospitals.

Conclusion

The indication of clustering does not provide an explanation for why the clustering occurs. Different processes can result in the same spatial patterns (e.g., true contagion vs. apparent contagion). However, the implementation of regression analysis alongside novel geo-visualization techniques allows us to start building a better picture of the city's environment, its resources, and their effect on individuals health and well-being. Of relevance to our epidemiological question is *Figure 11* which shows a linear association between the percent of individuals aged 25 and older without a high-school diploma and the percent of households living in crowded conditions across community areas in Chicago— a clear sign of hard living conditions, as enumerated by the hardship index. Evaluating *Figure 12* in light of this information tells us that the community areas facing the hardest living conditions as a result of lower educational attainment, particularly in terms of income and mortality rates, also tend to be the most crowded and with the highest number of dependents (either children or elderly). These living conditions could result in catastrophic events as COVID-19 spreads across Chicago. Disaster preparedness requires us to identify vulnerable populations early and to plan methods for which we can allocate essential resources to them expeditiously. It is unwise to do so while fighting a global pandemic— it is almost impossible; thus, only time will tell if we are able to avoid yet another tragedy, much like the 1995 heath wave.

Bibliography

1. Noguchi, Thomas T., and Joseph DiMona. *Coroner at Large*. Pochet Books, 1985.
2. Whitman, S., Good, G., Donoghue, E. R., Benbow, N., Shou, W., & Mou, S. (1997). Mortality in Chicago attributed to the July 1995 heat wave. *American journal of public health*, *87*(9), 1515–1518. https://doi.org/10.2105/ajph.87.9.1515
3. Klinenberg, E. (1999). Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave. *Theory and Society, 28*(2), 239-295. Retrieved March 18, 2020, from www.jstor.org/stable/3108472
4. Prior HHS Poverty Guidelines and Federal Register References. (2020, January 21). Retrieved from https://aspe.hhs.gov/prior-hhs-poverty-guidelines-and-federal-register-references
5. *COOKED: Survival by Zip Code*. 2019. [film] Directed by J. Helfand. Chicago, Illinois, USA: ITVS, Kartemquin Films.
6. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., … Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, *395*(10223), 507–513. doi: 10.1016/s0140-6736(20)30211-7
7. Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., … Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, *395*(10223), 514–523. doi: 10.1016/s0140-6736(20)30154-9
8. Guan, W.-jie, Doremalen, N. van, Cao, B., Lu, X., & China Medical Treatment Expert Group. (2020).Clinical Characteristics of Coronavirus Disease 2019 in China. *NEJM*, doi:10.1056/NEJMoa2002032
9. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., … Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, *395*(10229), 1054–1062. doi: 10.1016/s0140-6736(20)30566-3
10. Montiel, M. Lisa, Nathan, R.P., Wright, D.J. (2004). An Update on Urban Hardship. The *Nelson A. Rockefeller Institute of Government*. http://www.phasocal.org/wp-content/uploads/2014/03/EH-urban-hardship-Rockefeller-Institute-2004.pdf

Figures & Tables

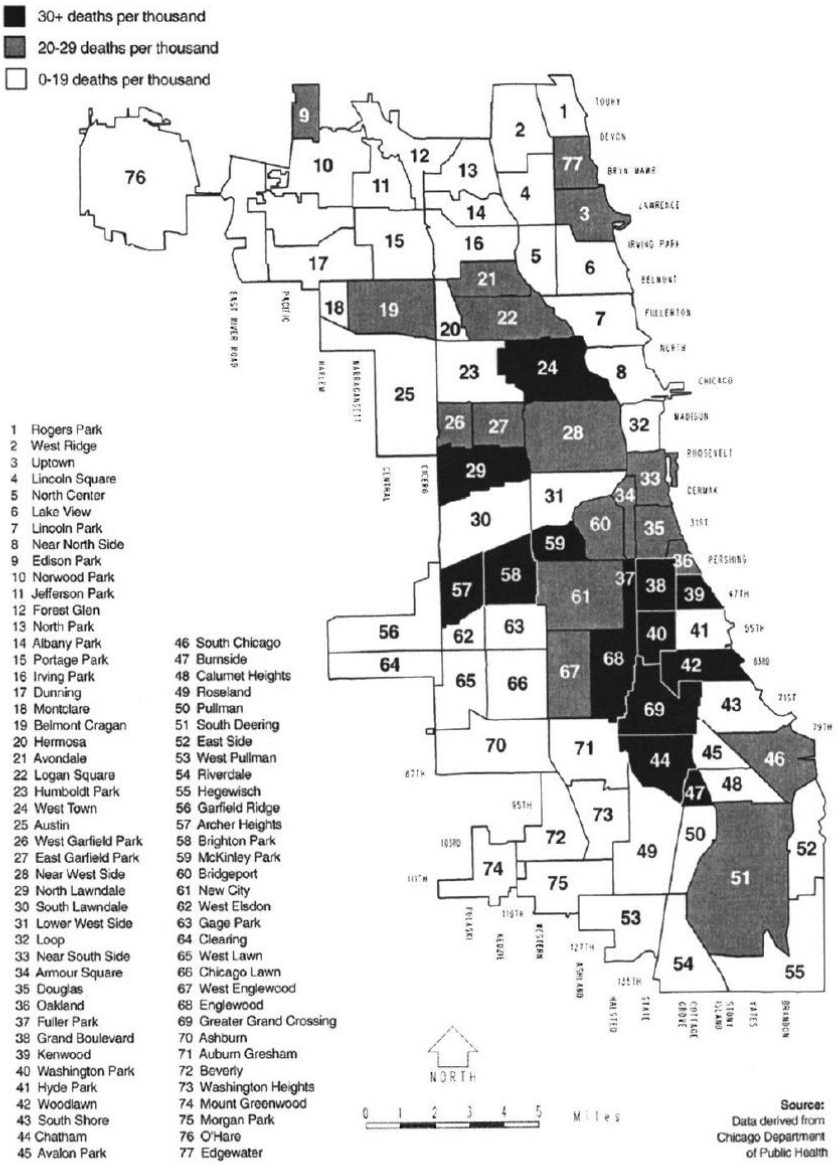

Figure 3. Chicago community areas with highest heat-related death rates.

Figure 1. Chicago community areas with the highest heat-related death rates during 1995 heat wave. Obtained from "Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave," by Klinenberg, E., 1999, *Theory and Society, 28*(2), p. 251.

Table 1. Chicago community areas with highest heat-related death rates

Community area	Heat-related deaths/100,000	Percent of population black	Median family income ($)
Fuller Park	92	99	8,371
Woodlawn	73	96	17,714
Archer Heights	54	0	37,744
Grand Crossing	52	99	22,931
Washington Park	51	99	9,657
Grand Boulevard	47	99	8,371
McKinley Park	45	0	31,597
North Lawndale	40	96	14,209
Chatham	35	99	29,258
Kenwood	33	77	31,954
Englewood	33	99	22,931
West Town	32	10	20,532
Brighton Park	31	0	30,677
Burnside	30	98	30,179
Near South Side	29	94	7,576
Chicago	7	39	30,707

Data based on 514 heat-related deaths located by the Illinois Department of Public Health.

Table 1. Chicago community areas with the highest heat-related death rates during 1995 heat wave. Obtained from "Denaturalizing Disaster: A Social Autopsy of the 1995 Chicago Heat Wave," by Klinenberg, E., 1999, *Theory and Society, 28*(2), p. 254.

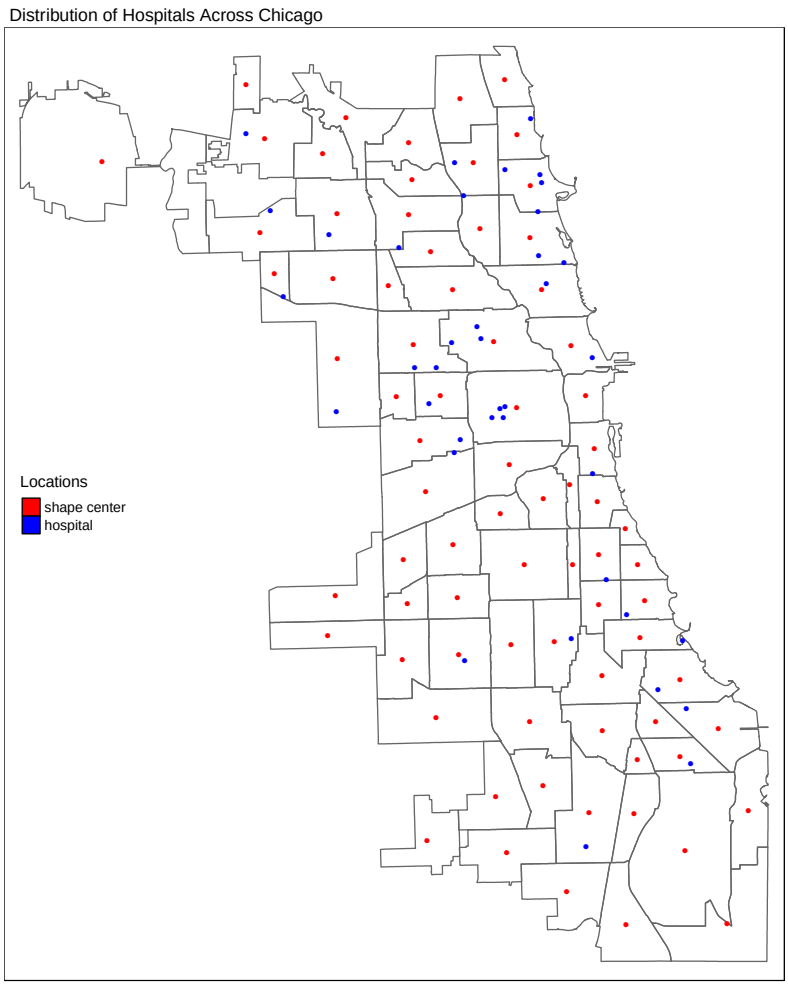

Figure 2. Map showing Chicago community areas with centroid (marked in red) and the locations of hospitals throughout the city (marked in blue); hospital data as of August 28, 2011.

Descriptive statistics for dependent variables

Cause.of.Death	*mean*	*SD*	*median*	*min*	*max*
coronary heart disease	144	30	140	83	248
diabetes-related	28	10	27	5	49
injury, unintentional	32	12	29	14	78
stroke (cerebrovascular disease)	45	14	41	21	88

Table 2. Summary statistics of dependent variables; data for Chicago community areas (2006-2010).

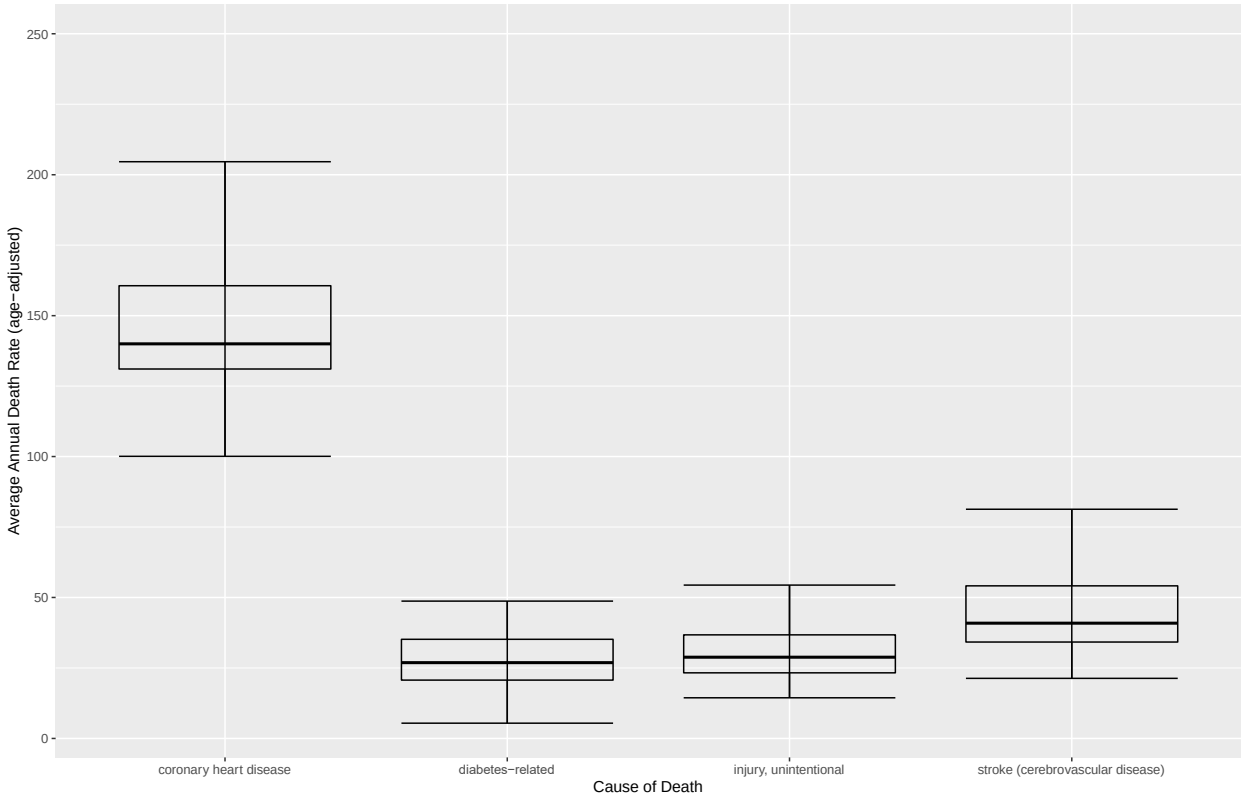

Figure 3. Box and whisker plot of dependent variables. For each underlying cause of death, outliers (red) and observations (black) are shown as points; data for Chicago community areas (2006-2010).

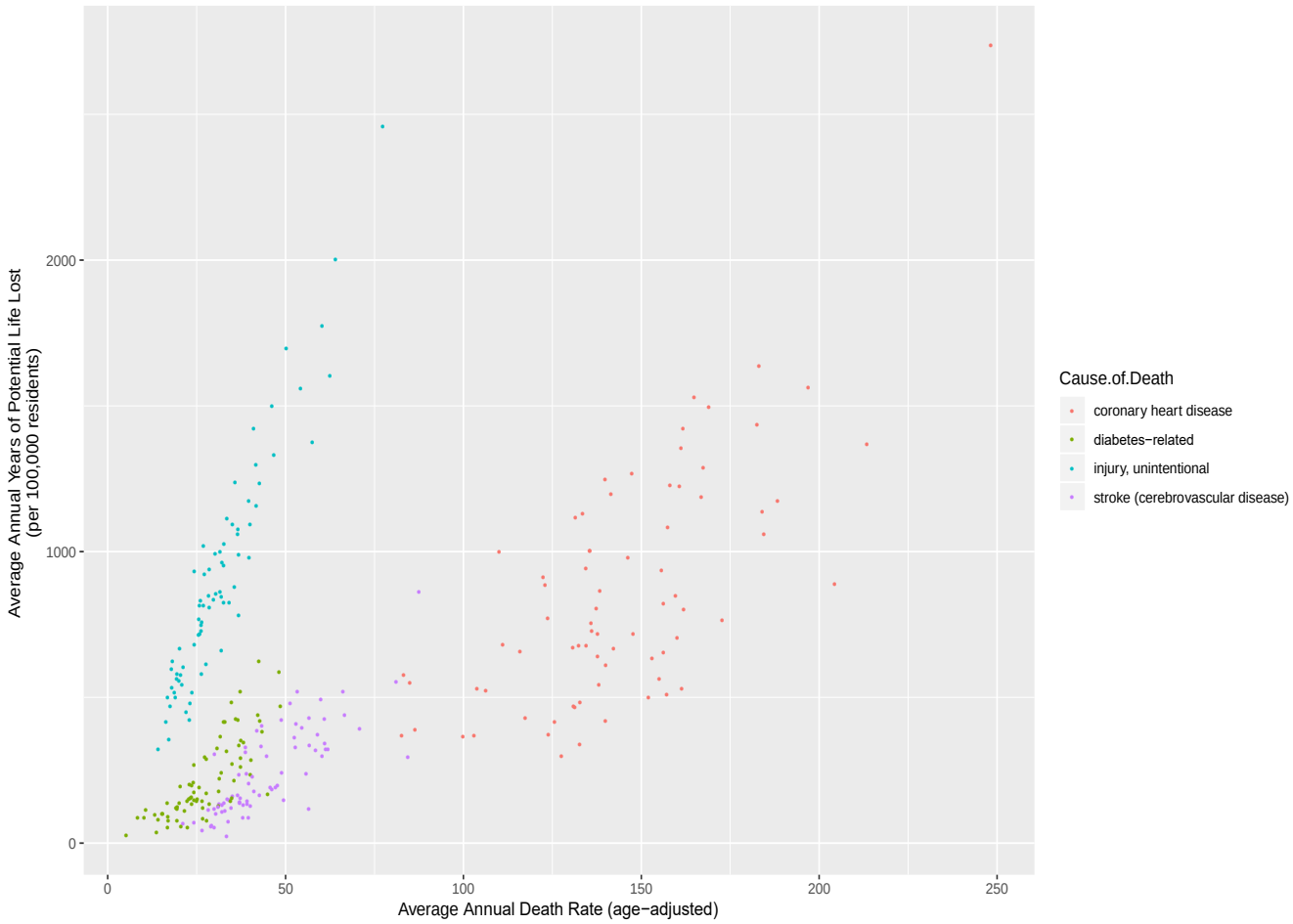

Figure 4. Bivariate analysis of average annual death rate vs. average annual years of potential life lost (YPLL) colored by underlying cause of death; data for Chicago community areas (2006-2010).

Descriptive statistics for independent variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PERCENT.OF.HOUSING.CROWDED	77	5	4	0	2	7	16
PERCENT.HOUSEHOLDS.BELOW.POVERTY	77	22	12	3	13	29	56
PERCENT.AGED.16..UNEMPLOYED	77	15	8	5	9	20	36
PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA	77	20	12	2	12	27	55
PERCENT.AGED.UNDER.18.OR.OVER.64	77	36	7	14	32	40	52
PER.CAPITA.INCOME	77	25,563	15,293	8,201	15,754	28,887	88,669
HARDSHIP.INDEX	77	50	29	1	25	74	98
DISTANCE.NEAREST.HOSPITAL	74	2	2	0	1	3	8

Table 3. Summary statistics of independent variables; data for Chicago community areas (2008-2012).

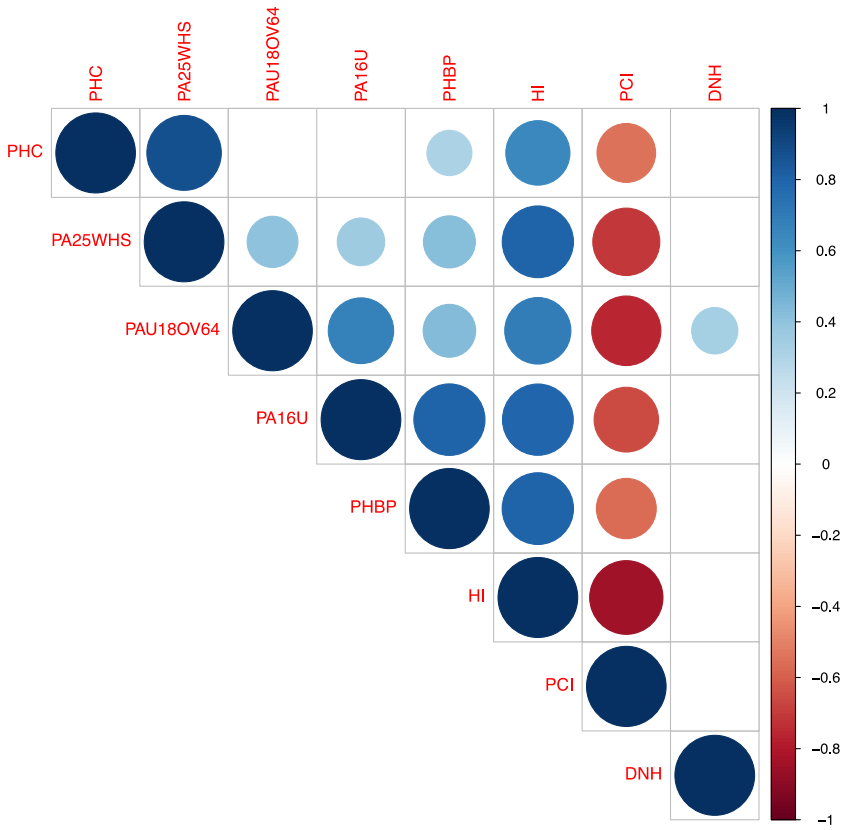

Figure 5. Correlogram for independent variables; data for Chicago community areas (2008-2012). Variable names were shortened to fit plot and are as follows: percent of occupied housing units with more than one person per room (PHC), percent of individuals aged 25 years or older without a high school diploma (PA25WHS), percent of the population under 18 or over 64 years of age (PAU18OV64), percent of persons aged 16 years and older in the labor force that are unemployed (PA16U), percent of households living below the federal poverty line (PHBP), hardship index (HI), per capita income (PCI), kilometer distance to the nearest hospital (DNH). Note, correlation coefficients with corresponding significance levels larger than 0.01 were ignored.

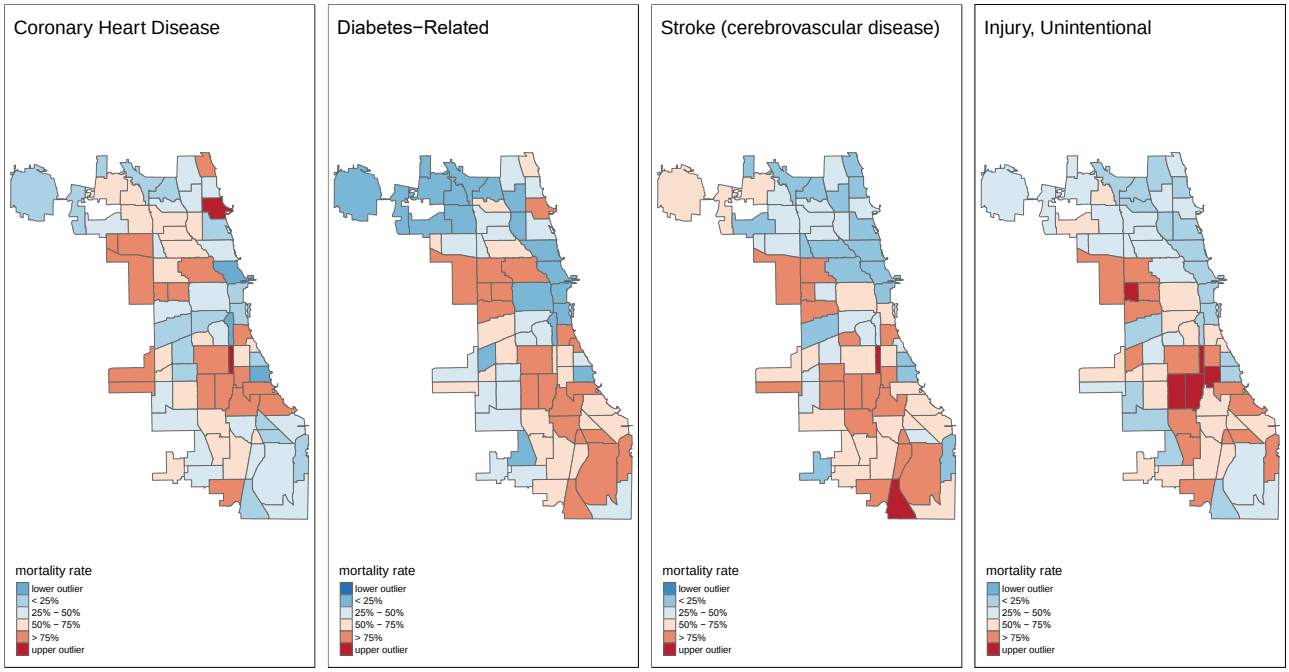

Figure 6. Box map showing average, age-adjusted, annual mortality rates across the city of Chicago for each underlying cause of death; data by Chicago community area (2006-2010).

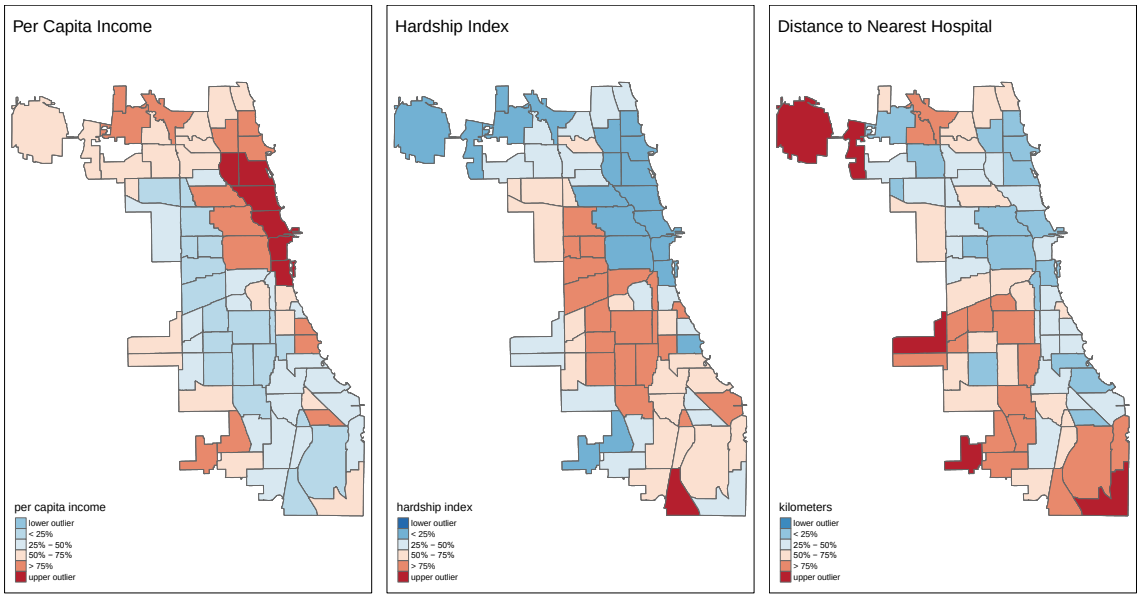

Figure 7. Box map showing data for each socioeconomic variable by Chicago community area (2008-2012).

Dependent Variable	Mortality Rate			
	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Independent Variable				
Hardship Index	0.331***	0.199***	0.298***	0.259***
	(0.114)	(0.032)	(0.044)	(0.040)
Distance from Nearest Hospital (km)	-0.786	-0.384	0.732	-0.548
	(1.986)	(0.556)	(0.767)	(0.690)
Constant	129.631***	18.740***	28.419***	20.019***
	(7.988)	(2.239)	(3.084)	(2.775)
Observations	77	77	77	77
R^2	0.104	0.345	0.390	0.369
Adjusted R^2	0.079	0.327	0.374	0.352
Residual Std. Error (df = 74)	28.446	7.972	10.983	9.883
F Statistic (df = 2; 74)	4.275**	19.483***	23.669***	21.658***
Note:				*p**p***p<0.01

Table 4. Results from multivariable regression analysis. Models 1-4 vary in underling cause of death; data is as follows: for model (1) coronary heart disease, (2) diabetes-related, (3) stroke (cerebrovascular disease), and (4) injury, unintentional.

Dependent Variable	Mortality Rates			
Independent Variable	(Model 1)	(Model 2)	(Model 3)	(Model 4)
Per Capita Income	-0.001***	-0.0004***	-0.0005***	-0.0004***
	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Distance from Nearest Hospital (km)	-1.908	-0.882	0.112	-1.107
	(1.942)	(0.566)	(0.839)	(0.744)
Constant	169.343***	39.416***	57.147***	45.381***
	(8.349)	(2.433)	(3.606)	(3.196)
Observations	77	77	77	77
R^2	0.168	0.342	0.292	0.289
Adjusted R^2	0.146	0.325	0.273	0.270
Residual Std. Error (df = 74)	27.403	7.988	11.835	10.491
F Statistic (df = 2; 74)	7.477***	19.265***	15.250***	15.052***
Note:				*p**p***p<0.01

Table 5. Results from multivariable regression analysis. Models 1-4 vary in underling cause of death; data is as follows: model (1) coronary heart disease, (2) diabetes-related, (3) stroke (cerebrovascular disease), and (4) injury, unintentional.

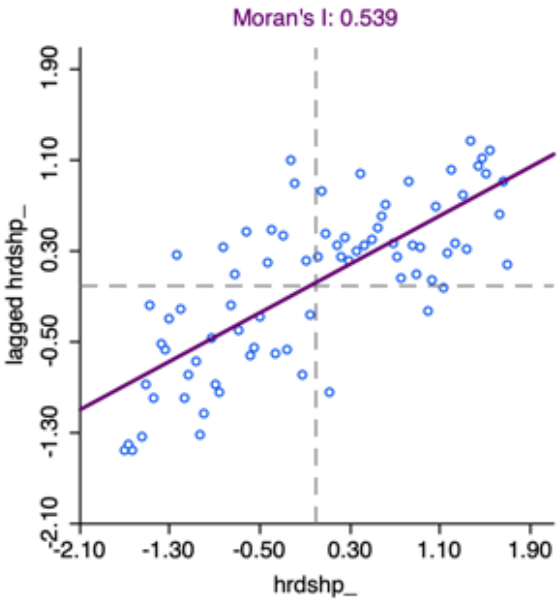

Figure 8. Moran scatter plot for the description of univariate global spatial autocorrelation. Significance was examined by means of a permutation test (~999 permutations). Here, we can observed that values for the hardship index most generally fall within the low-low (L-L) and high-high (H-H) categories, the lower left and upper right quadrants respectively. Data for Chicago community areas (2008-2012).

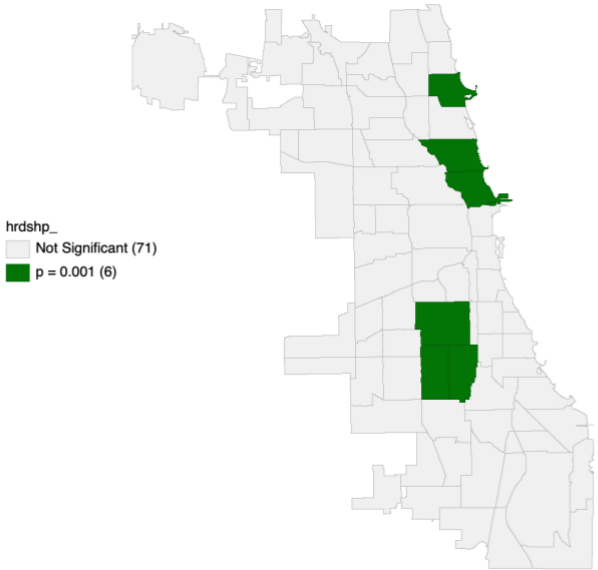

Figure 9. Associated significance map for local, univariate Moran statistic. Significance was examined by means of a permutation test (~999 permutations). Here, we can detect specific clusters of like values for our hardship index (pseudo p value < 0.001). Data for Chicago community areas (2008-2012).

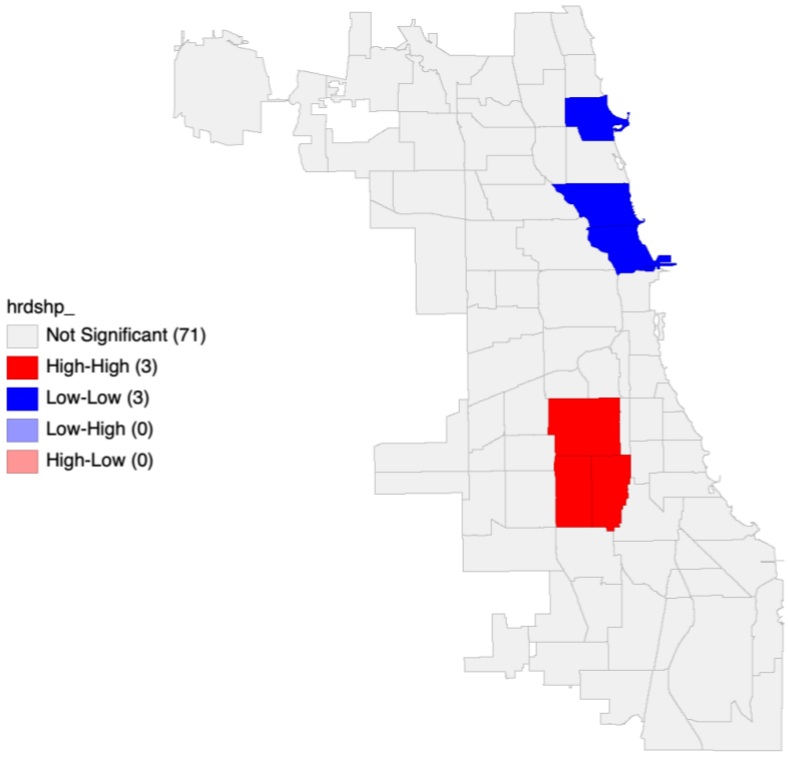

Figure 10. Associated cluster map for local, univariate Moran statistic. Significance was examined by means of a permutation test (~999 permutations). Here, we can detect specific clusters of like values, H-H and L-L, for our hardship index. Data for Chicago community areas (2008-2012).

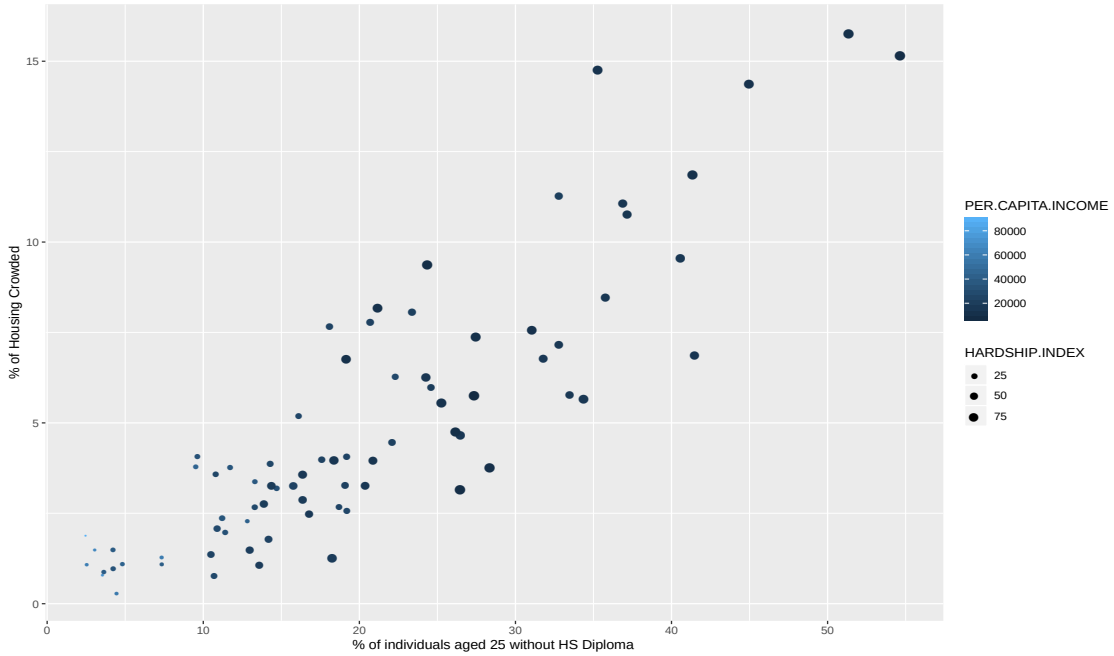

Figure 11. Bubble chart showing the association between some of our socioeconomic variables; data for Chicago community areas (2008-2012).

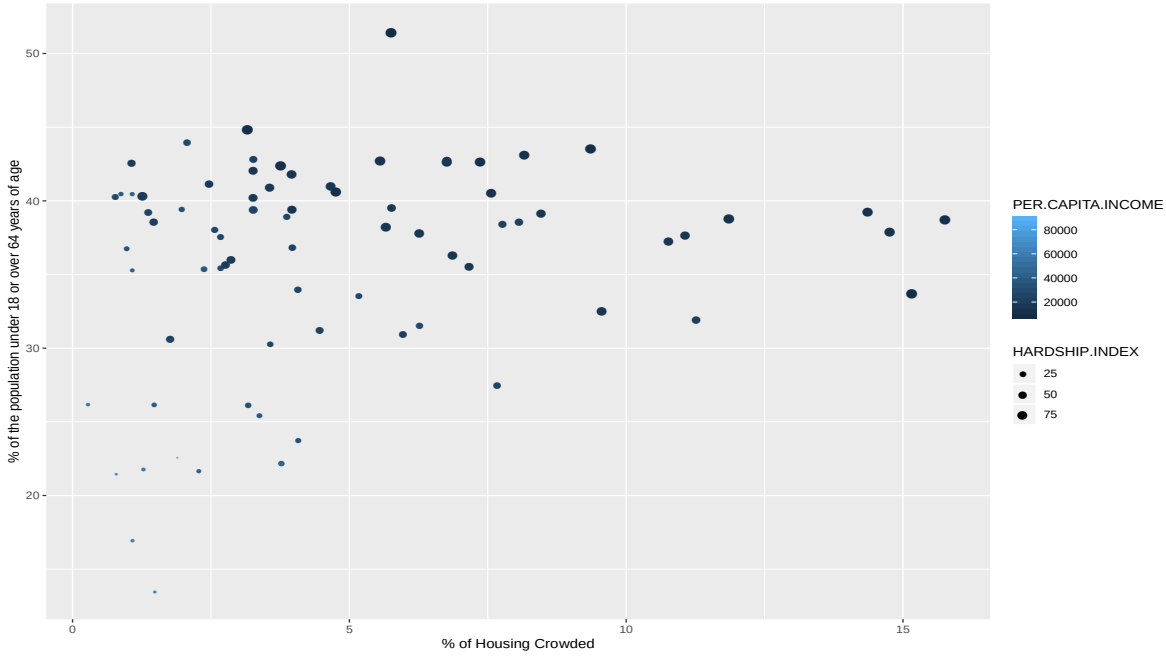

Figure 12. Bubble chart showing the association between some of our socioeconomic variables; data for Chicago community areas (2008-2012).