IDS 702 Final Project: Crime and Coffee in Chicago

Sarwari Das

Summary

This report investigates the relationship between gentrification and neighborhood crime rates in Chicago by using growth in coffee shops as a proxy for gentrification. The data analyzed contained information about the number of crimes committed per year, as well as growth of coffee shops and other demographic variables about the neighborhood. The final valid model was a hierarchical linear regression with neighborhood treated as the random intercept, and percentage of White population in the neighborhood treated as random slope. The model provides interesting insights into the the factors that affect the number of crimes in Chicago, and shows the role that different socio-economic factors can play in this relationship.

Introduction

According to a study by the Urban Displacement Project, as of 2017, more than 200,000 low-income Chicago households were at risk of, or already experiencing, some form of gentrification. One dimension of change often associated with gentrification is the change in crime rates. While policymakers contend that the in-migration of wealthier residents[1] allows communities to benefit from an income influx, empirical studies have produced contradictory results on the same. (Brown-Saracino 2010). One on hand, increased tax revenue in a community can lead to higher investments in services like after-dark police presence and public education, which are linked to lower crime. (Freeman, 2006). On the other, the displacement of families and disruption of social networks can lead to social friction and create a perfect breeding ground for criminal activity. Overall, literature indicates that while the benefits and detriments of gentrification varies across social groups (Lloyd 2005), in most cases, the gentrifiers benefit while the gentrified suffer.

As a process, gentrification is multifaceted and unfolds at an uneven pace, such that census data to capture its effect often feels inadequate. For example, increased disposable income associated with gentrification leads to economic change within a neighborhood, but changing social networks leads to a social change as well. To capture these changes, literature (Brown-Saracino 2010) has often resorted to a non census-based measure that maps onto the nonlinear tendencies of gentrification —the number of coffee shops in a neighborhood. Through supplying a commodity in a previously disinvested area in response to demand (economic effects), while allowing space for people to socialize and exchange ideas (social effects), coffee shops serve a critical function in the community and can be considered an important indicator of gentrification.

In this study, I use a hierarchical model with neighborhoods as a random intercept to investigate the relationship between gentrification (proxied by coffee shops) and crime rates across neighborhoods in Chicago. Further, I examine the other sociological factors that play a role in this relationship, and attempt to find if the effect of gentrification varies with the racial composition of a neighborhood.

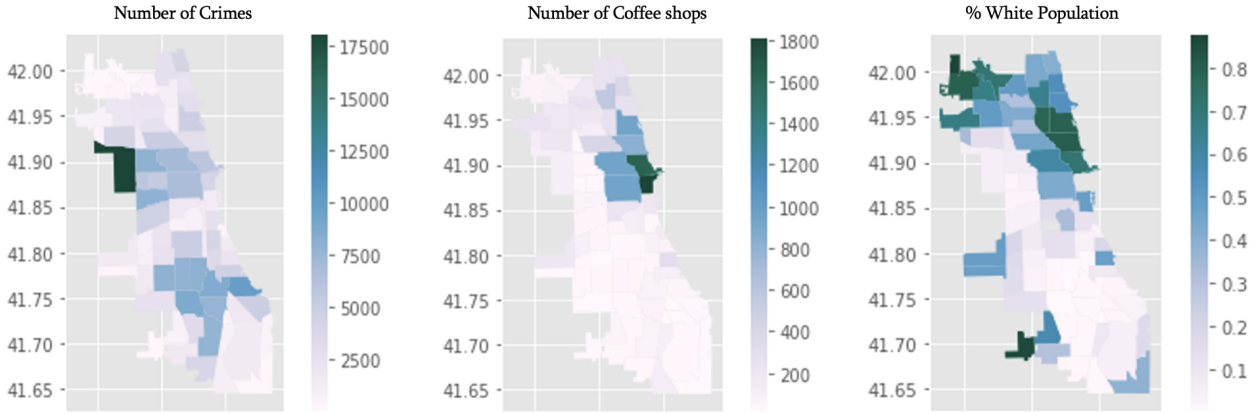


Figure 1: Distributions for variables of interest across neighborhoods in Chicago (2005-2019)

Data Pre-processing

The dataset for this project was built from three sources. An illustration of the data merging process is shown in the Appendix.

- Crime data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system and reflects individual incidents of crime that occurred in Chicago from 2001 to the present. I define crime by subsetting for homicides and robberies since the former is less susceptible to definitional variation by police while being less likely to be misreported, and the latter has far-reaching effects on urban life through its influence on choices regarding where to live and

[1] Although it exists in many forms, this report specifically conceptualizes gentrification as a process that involves the in-migration of wealth and the out-migration of poverty.

work (Papachristos, et al, 2011). I further filtered the data for the years between 2005 and 2019 to match the dimensions of my other datasets, leaving me with 202328 rows. These crime were mapped to the 77 neighborhoods in Chicago, and I dropped 3 observations that had an invalid values of '0'. I had no missing data, so I grouped my dataset by year and neighborhood, taking a sum of all crimes for every year-neighborhood level, resulting in 1155 rows of data.

- Data for coffee shops comes from Business licenses issued by the Department of Business Affairs, Chicago from 2002 to the present. To find coffee shops, I subset for all licenses with 'coffee' or 'café' in their titles, which gives me 15499 coffee shops. Like with crimes, I subset the data for 2005-2019. Instead of a neighborhood, this data was mapped to the latitude and longitude of each coffee shop. I used a shapefile of the boundaries of neighborhoods in Chicago to map each point location to a neighborhood, which failed to map two neighborhoods - Loop and Mckinley Park, leaving me with data on coffee shops mapped to the remaining 75 neighborhoods. With no other missing data, I grouped my dataset by year and neighborhood, taking a sum of coffee shops for each year-neighborhood level. This leaves me with 1125 rows of data.
- Additionally, I use demographic data from the Chicago Metropolitan Agency for Planning to collect the following counts about a neighborhood's population: *% of White (Non-Hispanic), Black, Latino, and Asian people, % of people with a College and High School education, Median Family Income (Real), Median Age, Total Population, % of unemployed individuals, and % of people in income bands (less than $25k, $25k-$50k, $50k-$75k, $75k-$100k, $100k-$150k, greater than $150k).* This data was only available for 5-year estimates by neighborhood, since Census surveys like the ACS do not record yearly data for geographies with population <65k. I group my data by neighborhoods across three 5-year time periods (2005-2009, 2010-2014 and 2015-2019), leaving me with 231 rows.

I had to similarly group my data on coffee and crime for the three time-periods, leaving me with 231 rows and 17 variables (Code book in Appendix) in total. I also divide my response variable 'crimes' by the total neighborhood population to get 'crime rates', which is now my main response. I center all my control variables to avoid potential problems with multicollinearity. Variables that represent a percent, I multiply by 100 for easier interpretation. All variables (except time-period and the neighborhood name) are numeric. Figure 1 shows a distribution of these datapoints across neighborhoods in Chicago, and summary statistics can be found in the Appendix.

Exploratory Data Analysis

Since *crimerate* can only take positive values, its distribution is highly right-skewed. A log transformation results in a fairly normal (Figure 2) distribution which is unaffected by outliers, and so I will use *log(crimerate)* as my response. Next, to discern if a multi-level model is appropriate, I explore the distribution of *log(crimerate)* across a random sample of eight neighborhoods in Chicago (Fig 3). Differences are seen in median crime rates across neighborhoods, which suggests that a random intercepts hierarchical model could be used to borrow information across these groups while accounting for varied baseline crime rates across neighborhoods. A similar analysis across time-periods (Fig 4) reveals that crime rates vary across time, which is expected. The choice to include time as an additional varying intercept or a potential interaction term with coffeeshops can be made during the modeling process.

Figure 2: Distribution of log (crimerate)

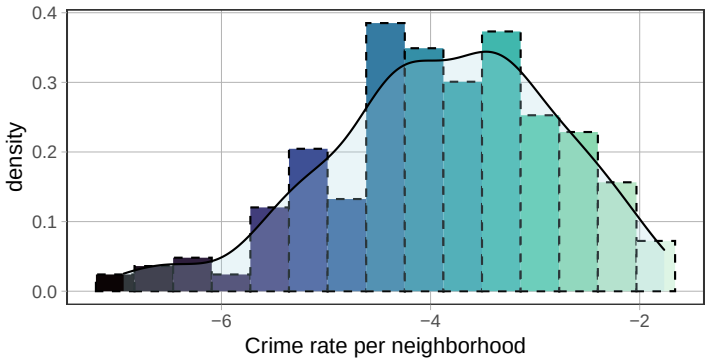
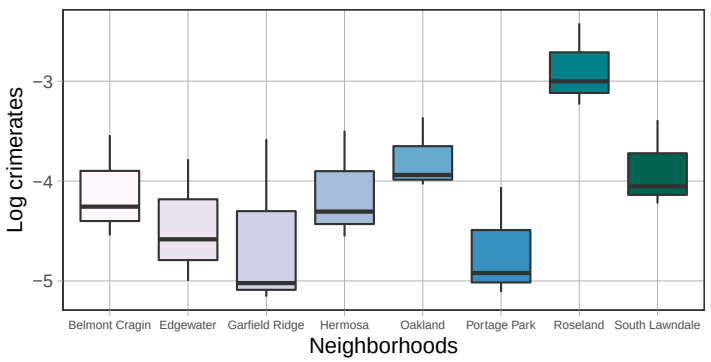


Figure 3: Crime levels by neighborhoods in Chicago



On plotting continuous control variables against *log(crimerate)*, I see linear relationships for all. Here, I log-transform the *coffeeshops* variable to visualize a slightly decreasing linear trend better (graph in Appendix), and retain this transformation throughout my analysis. Although this trend is decreasing across all neighborhoods, one interesting finding is that this relationship changes with the racial composition of a neighborhood. For example in Fig 5, we can see that the relationship between *% of white people* and *log(crimerate)* differs by neighborhoods. In Fig 1, we can also see that while coffeeshops tend to be situated in areas with higher *% of white* population, crimes in those areas are lower. This makes a case for a potential random slopes model of non-white population by neighborhood.

I conduct a similar results with median income and unemployment rate quartiles as well. I find that compared to the lower quartiles (1st and 2nd quartile), i.e, when unemployment is low, there exists a positive trend between *coffeeshops* and *crimerate*, although number of crimes are a lot lower. This trend flattens out for high unemployment quartiles. The same result applies for median income quartiles as well. This insight can support the view of policymakers since it indicates that neighborhoods with lower median incomes or high unemployment can benefit with reduced crime from coffee shop growth/ gentrification.

I also find potential interactions for proportions of each race with income bands. Coffeeshops showed a variation with timeperiods and income bands as well. Overall, my potential interactions are: *WhitePerc:MedianAge, WhitePerc:BlackPerc, MedianAge:IncomeLessThan25k, coffeeshops:IncomeLessThan25k, BlackPerc:MedianAge, coffeeshops:Income75k_100k, MedianAge:Income50k_75k, BlackPerc:IncomeLessThan25k, timeperiod:Income50k_75k, coffeeshops:Income25k_50k, BlackPerc:Income75k_100k, MedianAge:Income25k_50k, and timeperiod:MedianAge.* These will be explored further in my model building process.

Fig 4: Crime rate vs Coffee shops across timeperiods

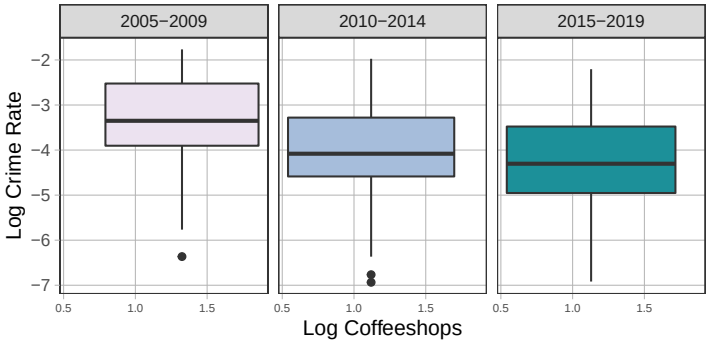
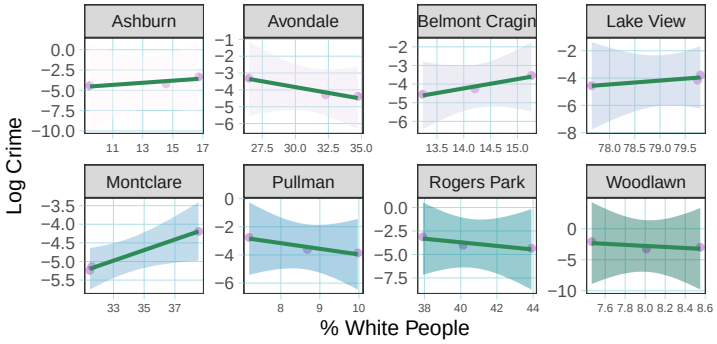


Fig 5: Crime rate vs White % by neighborhood



Model Selection

I start modeling by using a regular linear regression by defining a null model and full model to use in a stepwise regression. The null model contains an intercept, whereas the full model contains the main effects for coffeeshops and all control variables, as well the aforementioned potential interactions. I use BIC as a decision criterion although using AIC yields the same results. To build a parsimonious model. I conduct an ANOVA F-test and check AIC with each term from the suggested stepwise model to check it significantly improves the model. After elimination of variables that don't improve performance, I am left with the following variables: *coffeeshops, WhitePerc, BlackPerc, timeperiod, MedianAge, IncomeLessThan25k, Income50k-75k, Income50k-75k, WhitePerc:MedianAge, coffeeshops:IncomeLessThan25k, coffeeshops:Income50k-75k* and *coffeeshops:Income25k-50k.* Although we haven't accounted for neighborhoods, this model explains 90% of the variation in *log(crimerates)* in the overall data. The *coffeeshops* variable is highly significant (full results in Appendix). On checking model assumptions, I see that the assumption of normality of residuals is violated, while linearity and independence are met. On checking for multicollinearity (table and correlation plot in Appendix), two variables *WhitePerc* and *WhitePerc:MedianAge* show VIF greater than 5. As they are already mean centered, I drop *WhitePerc:MedianAge* from my model, as a result, the VIF of *WhitePerc* drops below 5. This model is not influenced by outliers or leverage points.

Table 1: Comparison of AIC for different hierarchical models

Model	***AIC***
Neighborhood (random intercept)	63.59
Time-Period (random intercept)	272.02
Neighborhood + Time-Period (random intercept)	151.82
Neighborhood (random intercept) + % of White population (random slope)	60.12

Next, I fit a random intercept hierarchical linear regression models including my selected variables from the step model with neighborhood as a random intercept. I am curious to know if time-period can serve as a random intercept, so I fit two additional models, with time-period as an intercept individually and in combination with neighborhoods as well. In Table 1 above, we can see that these two intercepts do not improve our model; hence I proceed with neighborhood as the only random-intercept. Further, from the results of EDA, I test out random slopes for *% of White population.* An ANOVA test finds significant improvement (p<0.05) by adding this random slope, so I decide to add it to my model.[2]

My final model can be represented as:

$$\log(\text{crimerates})_i = (\beta_0+\gamma_{0j})+\beta_1\cdot\log(\text{coffeeshops})_{ij}+(\beta_2+\gamma_{1j})\cdot\text{WhitePerc}_{ij}+\beta_3\cdot\text{BlackPerc}_{ij}+\sum_{m=2}^{3}+\beta_{4m}\cdot\mathbb{I}[\text{timeperiod}_{ij}=m]+$$

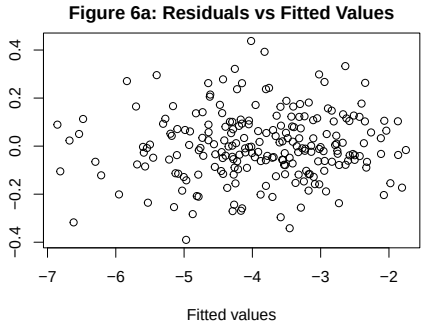
$$\beta_5\cdot\text{MedianAge}_{ij}+\beta_6\cdot\text{IncomeLessThan25k}_{ij}+\beta_7\cdot\text{Income50k}-75\text{k}_{ij}+\beta_8\cdot\text{Income25k}-50\text{k}_{ij}+\beta_9\cdot\text{coffeeshops}:\text{IncomeLessThan25k}_{ij}$$

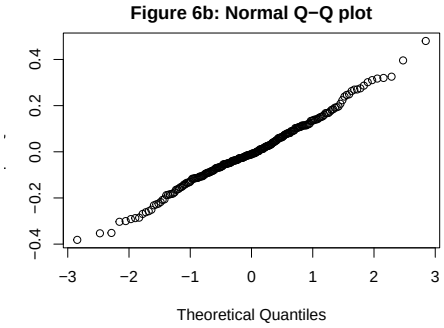
$$+\beta_{10}\cdot\text{coffeeshops}:\text{Income50k}-75\text{k}_{ij}+\beta_{11}\cdot\text{coffeeshops}:\text{Income25k}-50\text{k}_{ij}+\varepsilon_{ij}$$

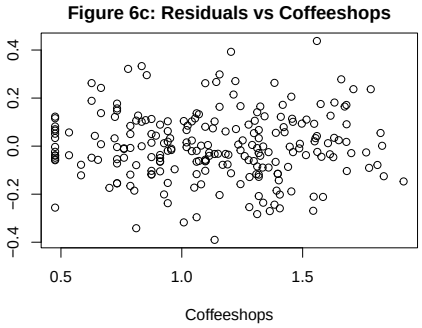
$$\varepsilon_{ij}\sim\mathcal{N}(0,\sigma^2), i=1,...,n; j=1,...,J; (\gamma_{0j},\gamma_{1j})\sim\mathcal{N}(0,\tau_0^2), i=1,...,n; j=1,...,J$$

[2]It's worth mentioning that I initially fit this data using a Poisson Regression with counts of crimes (offset by population to model rates) as my response variable. On comparing the RMSE for Poisson hierarchal model vs my current model, my current model performs better. Since no linear assumptions are violated in my current model, I choose to ahead with it.

The final model satisfies the linear regression assumptions. Residual vs fitted value plots (Fig 6a), show no discernible trends; points are scattered randomly across the Y axis and form a band around zero, and so the independence and constant variance assumption is met. The same applies to plotting residuals against each predictor (Fig 6c); no trends are found, and so linearity is met. The QQ-plot is fairly fit to the 45 degree line (Fig 6b), and therefore normality is not violated as well.







Model Interpretation

Looking at Table 2, with the exception of the income band *Income25k-50k* and *MedianAge*, all fixed effects in my model are significant. On exponentiating the coefficients, we can see that the baseline crime rate is around 3% when control variables are at their baseline. Further, on average, a 1% increase in number of coffeeshops is associated with a 0.26% increase in the crime rate of that area (p<0.01), holding all else constant. A 1% increase in the % of white population in an area from at its mean (~27%), is associated with a 2% decrease in crime (p<0.001), on average, while a percent increase in the % of black population in an area from at its mean (39.3%) is associated with a 1% increase in crime (p<0.001). Compared to the baseline time-period of 2005-2009, average crime rates in 2010-2014 fell by about 50% in both, 2010-2014 (p<0.001) and 2015-2019 (p<0.001). Further, as the percent of people earning less than 25k in an area increase 1% from its mean (11.45%), crime rate increases by 7% on average (p<0.001), and as the percent of people earning between 50k-75k increase by 1% (from its mean of 5%), the crime rate decreases by 12% on average (p<0.01).

Looking at our random effects (Table 3), the estimated τ_0^2 (0.28) describes the across neighborhood variation attributed to the random/varying intercept, whereas the estimated τ_1^2 (0.008) describes the across neighborhood variation attributed to the random slope of WhitePerc. Hence τ_0^2 tells us how neighborhoods vary in log(crime rates), while τ_1^2 explains less than 1% of that variation within neighborhoods. Further, the estimated standard error σ^2 (0.17) describes the remaining unexplained variation within neighborhoods. Overall, it seems that WhitePerc does not contribute much to explain the variation in crimes within neighborhoods, while neighborhoods vary measurably in baseline crime rates. A dotplot of the model in the appendix can show that although intercepts across neighborhoods are seen to vary, results are not significant while looking at the WhitePerc in a neighborhood.

Table 2: Fixed effects of the hierarchical linear regression model

	Estimate (Exp.)	Std. Error	t value	Lower Bound	Upper Bound	p value
Intercept	0.0244	0.1266	-29.3455	-3.9847	-3.4451	0.0000
coffeeshops	1.3086	0.0932	2.8866	0.0742	0.4657	0.0043
WhitePerc	0.9826	0.0026	-6.6476	-0.0227	-0.0122	0.0000
BlackPerc	1.0101	0.0015	6.5843	0.0071	0.0130	0.0000
timeperiod_2010-2014	0.5191	0.0356	-18.3971	-0.7282	-0.5810	0.0000
timeperiod_2015-2019	0.4524	0.0386	-20.5249	-0.8734	-0.7102	0.0000
MedianAge	0.9917	0.0076	-1.0913	-0.0246	0.0075	0.2768
IncomeLessThan25k	1.0799	0.0159	4.8217	0.0462	0.1078	0.0000
Income50k-75k	0.8830	0.0450	-2.7626	-0.2155	-0.0370	0.0064
Income25k-50k	1.0599	0.0328	1.7719	-0.0075	0.1256	0.0782
coffeeshops:IncomeLessThan25k	0.9548	0.0127	-3.6380	-0.0712	-0.0218	0.0004
coffeeshops:Income50k-75k	1.1360	0.0369	3.4536	0.0533	0.2056	0.0007
coffeeshops:Income25k-50k	0.9376	0.0294	-2.1878	-0.1259	-0.0045	0.0300

Table 3: Variance of the random effects

	Groups	Name	Variance	Std.Dev.
1	Neighborhood	(Intercept)	0.0824	0.2871
2	Neighborhood	WhitePerc	0.0001	0.0080
4	Residual		0.0294	0.1715

Limitations

I believe that there are several ways to improve this analysis. Firstly, we should be wary of bias in the data itself. I find coffeeshops by subsetting for business licenses with 'coffee' or 'café' in their titles, but its possible that some of these are diners or general stores with these words in their title, such that they don't meet up to the gentrification proxy requirement that has been studied extensively by researchers. An alternative to this method would be to use APIs like GoogleMaps or Yelp that specifically identify coffeeshops as a point of interest (POI). Further, I think my analysis is constrained by lack of data. Since the American Community Survey does not publish demographic details for neighborhoods on an annual level, I had to aggregate my data to 5-year time periods which made it significantly smaller. Additionaly, I believe that it is hard to draw out estimates for 5-year time-periods since changes in crimerates could be the result of exogenous factors (like changes in government policies, for example). The way around this would be to query the ACS 1-year experimental estimates for the city of Chicago, and then use shapefiles for neighborhood boundaries to estimate population sizes for each neighborhood. By studying changes across years, we should be able to see the effect of gentrification on a more granular level. We could also try to include lagged crime rates as a predictor in our model, since change in crime rates makes for a more interesting analysis than just absolute rates. Lastly, we have missing data for two neighborhoods. This could be fixed by further investigating the boundaries of these neighborhoods.

In the model itself, I had several candidates for varying slopes. I was specifically interested in whether race would play a role in the relationship between coffeeshops and crime, but its possible that something like unemployment rate would make for a more accurate model. Although I did explore several of these, I think there is scope for more research here. Next, my varying slopes for WhitePerc explained a very low variation in crimerate within neighborhood. Since ANOVA and AIC suggests that I stick to my random slope-random intercept model, I have done so, but its possible that results could be improved by including additional varying slopes, different functional forms for predictors, and so on.

Conclusion

Overall, the model was valid and addressed questions about the factors that affected the crime rates across neighborhoods in Chicago. Specifically, it was discovered that higher numbers of coffeeshops were associated with higher crime rates, and that higher populations of people in higher income bands are associated to a decrease in crimerates. Further, we see that crimerates vary across neighborhoods, and within neighborhoods, there is some variation by the percentage of White population in that neighborhood. This analysis can provide a foundation for future research about how growth of coffee shops in predominantly White neighborhoods can affect crime when compared to growth of coffee shops in predominantly Black neighborhoods (for example, Table 4 in the Appendix). Controlling for all else, if growth of coffee shops are truly a proxy for gentrification, then gentrification could be linked to higher crimes, but of course the caveat is that we can't establish a direct causal relationship here. Ideally, we could try to incorporate a time-series model too see what other demographic variables are changing over time as well.

Citations

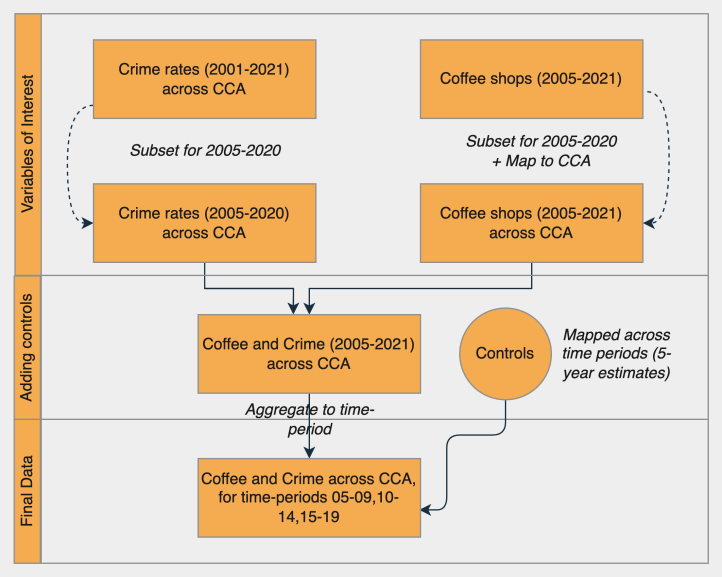
1. Brown-Saracino, Japonica ed. 2010. *The Gentrification Debates.* New York: Routledge.
2. Freeman, Lance. 2006. "Displacement or Succession?: Residential Mobility in Gentrifying Neighborhoods." *Urban Affairs Review* 40(4):463–91.
3. Lloyd, Richard. 2005. *Neo-Bohemia: Art and Commerce in the Post-Industrial City.* New York: Routledge.

Appendix

1. Data dictionary

Variable	Description
Neighborhood	*denotes the 77 neighborhoods of Chicago*
crimes	*count of homicides and robberies committed in neighborhoods in Chicago per time-period*
crimerate	*crimes by total population of a neighborhood*
coffeeshops	*count of coffeeshops in a neighborhood in Chicago per time-period*
TotalPopulation	*Total population in a neighborhood*
MedianIncome	*Median Family Income (Real) of a neighborhood*
MedianAge	*Median Age of a neighborhood*
IncomeLessThan25k	*% of people earning less than $25,000 in a neighborhood*
Income_25k_to_50k	*% of people earning between $25,000 and $50,0000 in a neighborhood*
Income_50k_to_75k	*% of people earning between $50,000 and $75,0000 in a neighborhood*
Income_75k_to_100k	*% of people earning between $75,000 and $100,0000 in a neighborhood*
Income_100k_to_150k	*% of people earning between $100,000 and $150,0000 in a neighborhood*
IncomeGreaterThan150k	*% of people earning greater than $150,0000 in a neighborhood*
UnemployedPerc	*% of unemployed people in a neighborhood*
timeperiod	*time periods (2005-2009), (2010-2014), (2015-2019)*
WhitePerc	*% of White (Non - hispanic) population in a neighborhood*
AsianPerc	*% of Asian population in a neighborhood*
BlackPerc	*% of Black population in a neighborhood*
HispPerc	*% of Hispanic population in a neighborhood*
HSperc	*% of population with high school education in a neighborhood*
BACHperc	*% of population with college education in a neighborhood*

2. Data Merging Process

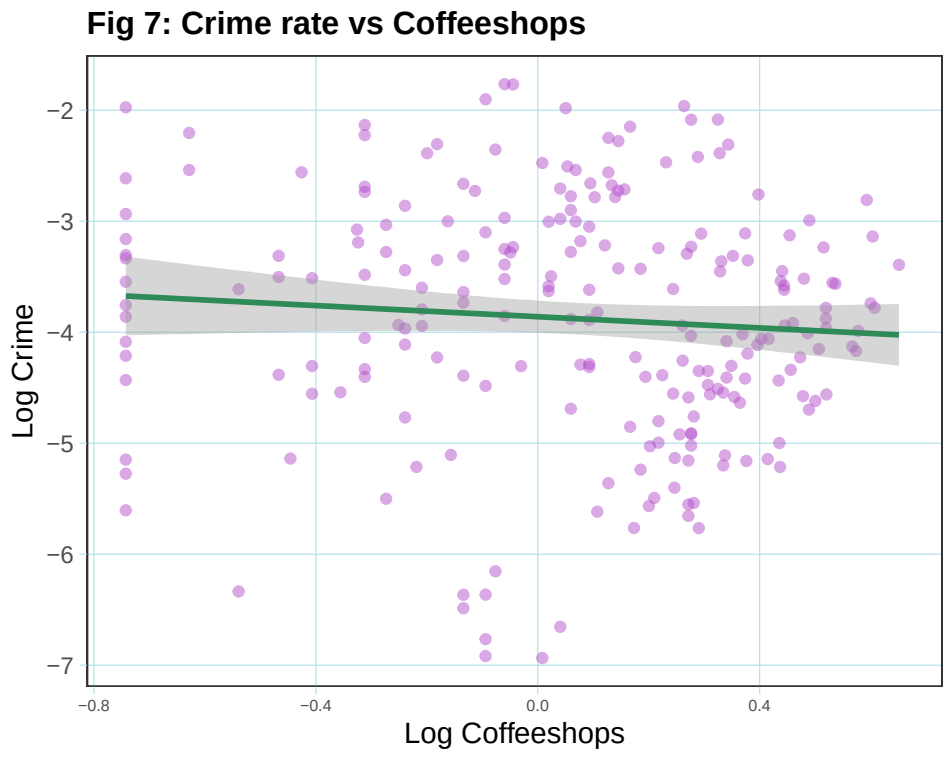


3. Summary Statistics

Table 5: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
crimerate	225	0.03	0.03	0.001	0.01	0.04	0.17
coffeeshops	231	64.29	123.15	5	10.4	56.5	1,005
MedianAge	225	35.03	4.86	21.52	31.49	38.00	48.50
MedianIncome	225	49,208.83	22,508.96	14,204.39	32,530.77	60,880.00	125,033.00
AsianPerc	225	0.05	0.10	0.00	0.003	0.07	0.73
BlackPerc	225	0.39	0.40	0.00	0.03	0.88	1.00
HispPerc	225	0.26	0.28	0.00	0.04	0.45	0.92
WhitePerc	225	0.28	0.27	0.00	0.03	0.49	0.92
UnemployedPerc	225	0.07	0.03	0.003	0.04	0.09	0.15
IncomeLessThan25k	225	0.11	0.05	0.03	0.07	0.15	0.27
Income_25k_to_50k	225	0.09	0.02	0.04	0.07	0.10	0.13
Income_50k_to_75k	225	0.06	0.01	0.01	0.05	0.07	0.10
Income_75k_to_100k	225	0.04	0.02	0.01	0.03	0.05	0.09
Income_100k_to_150k	225	0.04	0.02	0.00	0.02	0.06	0.10
IncomeGreaterThan150k	225	0.04	0.04	0.00	0.01	0.05	0.22
HSperc	225	0.41	0.20	0.04	0.21	0.57	0.81
BACHperc	225	0.17	0.14	0.02	0.07	0.22	0.65

4. Crime Rate vs Coffeeshops



5. Linear Model Summary

Table 6: Results

	Dependent variable:
	logcrimerate
coffeeshops	0.29^{***} (0.09)
WhitePerc	-0.02^{***} (0.002)
BlackPerc	0.01^{***} (0.001)
timeperiod2010-2014	-0.63^{***} (0.05)
timeperiod2015-2019	-0.73^{***} (0.06)
MedianAge	-0.04^{***} (0.01)
IncomeLessThan25k	0.10^{***} (0.02)
Income50k_75k	-0.28^{***} (0.05)
Income25k_50k	0.10^{*} (0.05)
WhitePerc:MedianAge	-0.001^{***} (0.0002)
coffeeshops:IncomeLessThan25k	-0.06^{***} (0.01)
coffeeshops:Income50k_75k	0.25^{***} (0.04)
BlackPerc:IncomeLessThan25k	-0.001^{***} (0.0002)
timeperiod2010-2014:Income25k_50k	-0.04 (0.03)
timeperiod2015-2019:Income25k_50k	-0.06^{**} (0.03)
coffeeshops:Income25k_50k	-0.10^{***} (0.04)
BlackPerc:Income50k_75k	-0.001 (0.0005)
Constant	-3.69^{***} (0.12)
Observations	225
R^2	0.93
Adjusted R^2	0.92
Residual Std. Error	0.31 (df = 207)
F Statistic	150.32^{***} (df = 17; 207)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Random Coefficients for historically black vs historically white neighborhoods

Neighborhood	*Random Intercept*	*Random Slope: WhitePerc*
Englewood (Historically black)	1.12	0.997
Chatham (Historically black)	1.08	0.998
Uptown (Historically white)	1.12	1.002
Logan Square (Historically white)	1.39	1.003

6. Variation in log(crimerates) across Neighborhoods

$GEOG

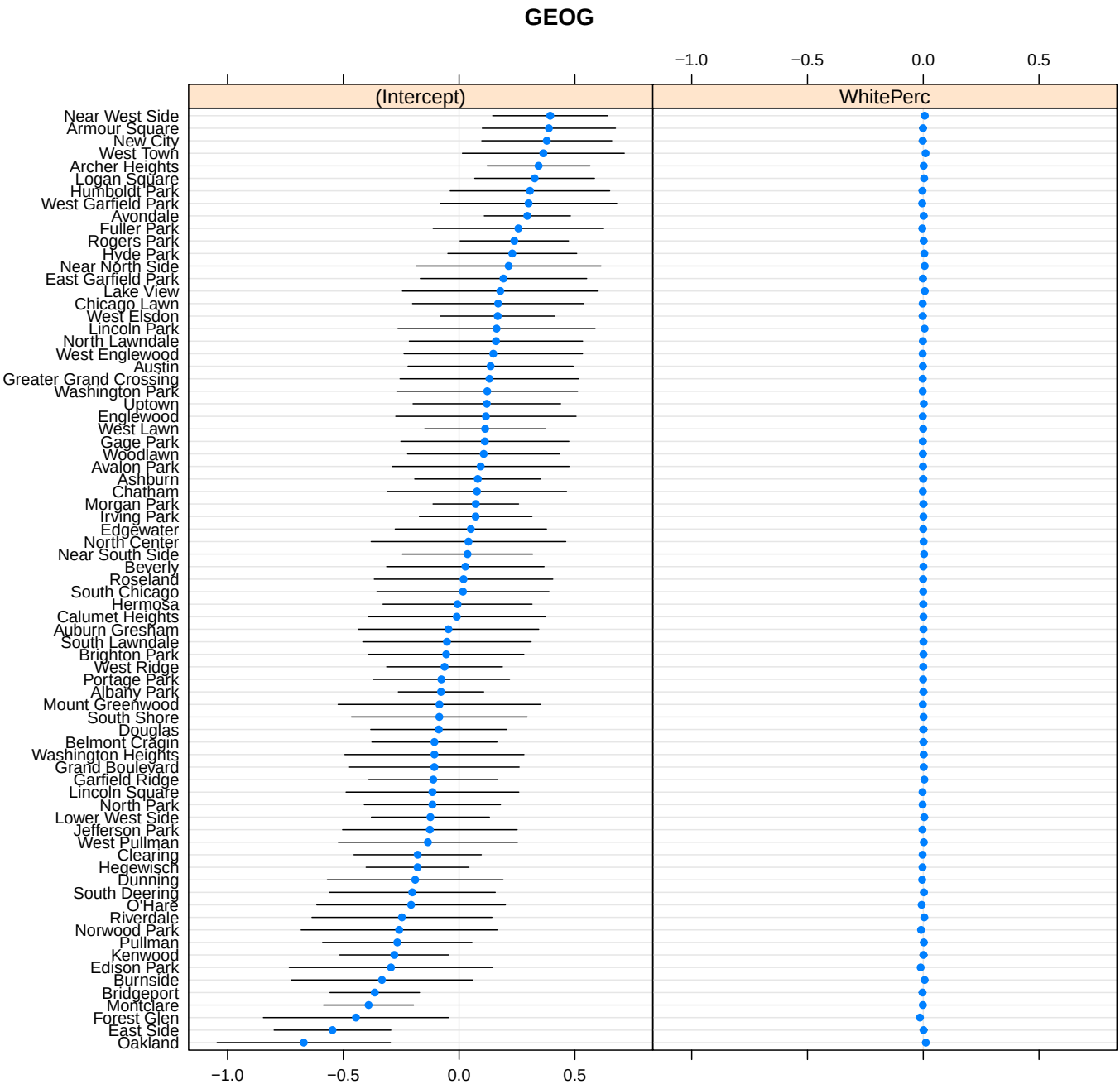


Figure 2: Dotplot for hierarchal model