Access to Food in Chicago: a Hierarchical Perspective

Daniel Berry

November 22, 2016

Abstract

Easy access to healthy, nutrient rich foods is a current public health crisis in the United States. Areas that have limited access to healthy foods and easy access to unhealthy foods are called "food deserts". In the United States today, X million people are living in a food desert. Access to food can be measured at the city-block level and can have substantial variation within a single neighborhood. However, much publicly available city wide health data is available at the neighborhood level. This gives the data a natural hierarchical structure as city blocks are nested within neighborhoods. In this paper we model presence/absence of deserts using a combination of block-level and neighborhood-level data from the city of Chicago. We apply construct a hierarchical logistic regression model and demonstrate its superiority.

Introduction

Access to readily available and inexpensive healthy food options is important for personal health. Unfortunately, for some in society today, access is not equitable. Some towns and neighborhoods have excellent healthy food options. Others do not. In some locations it is difficult to impossible for many residents to find healthy food options. In these areas, known as food deserts, there is readily available processed, calorie dense but nutrient poor foods.

A person's access to food is determined by the city block in which they live. Any increase in precision beyond the city block level is meaningless. However, rarely is data reported at the city block level. Sampling variability at that level would make drawing conclusions difficult. For that and other reasons, often city data is reported at the neighborhood, zip code, or telephone area code level. This gives the data a natural hierarchical structure where city blocks are nested within neighborhoods.

Food Deserts

There is no single agreed upon definition of a food desert. Some common metrics are based on distances to nearest supermarket with cutoffs at 1 mile for urban areas and 10 miles for rural areas. 1 mile may seem like a short distance to travel, but in urban areas with low car ownership rates where residents rely on public transportation, a 1 mile journey can take a substantial amount of time. For this work we used the definition from of defining a city block as being in a food desert if the city block is more than 1 mile from a supermarket. Supermarket in this context is a grocery store that is larger than 10000 square feet that is not primarily a liquor store. Distance is defined as the great circle distance from the center of mass of the city block to the center of mass of the grocery store.

We can see the location of food deserts as computed using our metric visualized in figure 1.

Chicago is a very racially segregated city. As shown in figures 2 and 3, many neighborhoods in Chicago are >75% a single race.

In fact, it appears that many food deserts are located in majority black neighborhoods. We will explore this relationship later on during the model building phase. Chicago also has very strong class divisions. For example the (near) north side is very wealthy while the south side and suburbs are less so, see 4.

Perhaps instead, food deserts are associated with urban decay. We might hypothesize that

Methods

In the following section we describe the procedures used. We begin by describing the data used and including data sources. Then we move on to describing the types of models used.

Data Gathering and Manipulation

All data was sourced from the generous Open Data Portal operated by the city of Chicago. The portal can be found at data.cityofchicago.org. We utilized the following data files:

Block level data

Crimes 2001 - present This file contains a record for crimes in Chicago since 2001 with information about the type of crime as well as its location. The location is pseudo-anonymized to be random but within the same city block. For each city block we counted the total number of crimes committed within 1 mile in 2009. Our hypothesis *a priori* was that food deserts were often located in high-crime areas.

311 Service Requests: Vacant Buildings This file contains a record for every 311 service call about an abandoned/unoccupied/unlawfully occupied building. For each city block we counted the number of calls for vacant buildings where the building was located within 1 mile of the city block. Our hypothesis was that food deserts were located in areas with higher levels of vacant buildings.

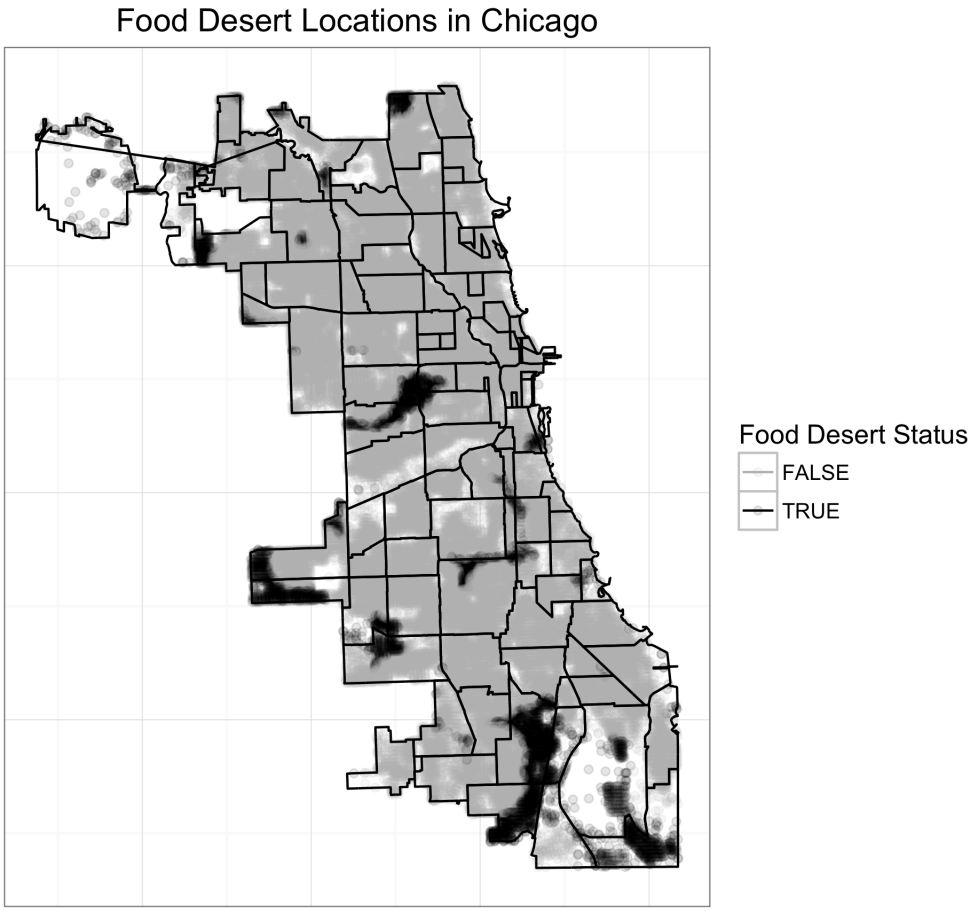

Figure 1: Food Desert Locations in Chicago

CTA Ridership: Avg. Weekly Boardings during October 2010 This file contains average weekly boardings for the month of October 2010 for every CTA (Chicago Transit Authority) bus stop in Chicago. We counted the total boardings for all stops within a mile of each city block. Our hypothesis was that residents of food deserts would tend to have a higher reliance on public transportation than residents in other areas.

Census Block Population This file contains the population of each city block.

Neighborhood level data

Public Health Statistics: selected public health indicators by Chicago community area This file contains a record for every neighborhood in Chicago with selected public health information. Examples include teenage births per 100,000 residents and Gonorrhea prevelence (cases per 100,000). Our hypothesis was that food deserts were likely to be underserved in a more general public health sense than just lacking access to food.

Census Data: Selected socioeconomic indicators This file contains a record for every neighborhood in Chicago with information about the socioeconomic status of that neighborhood. Examples include percent of residents below the povery level and percent without a high school diploma.

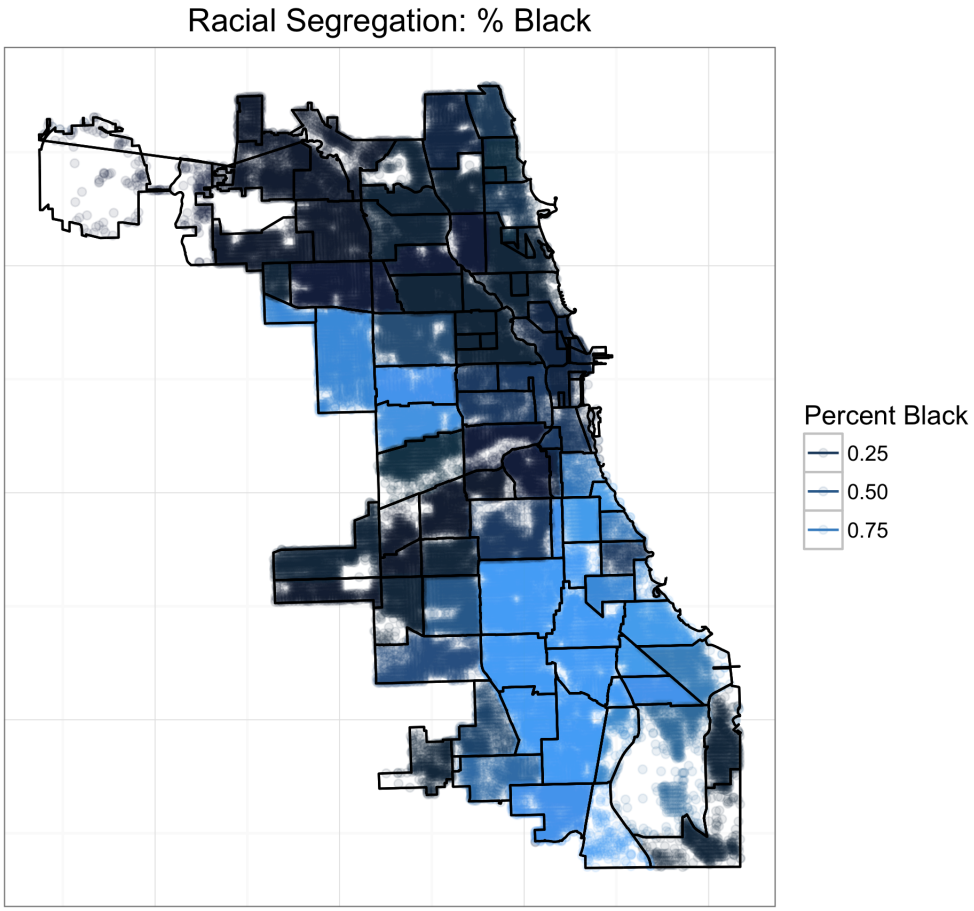

Figure 2: Neighborhoods by Black Percentage

Race by Community Area This file contains a record for every neighborhood in Chicago with the number of residents of each race who reside in that neighborhood.

We tried to gather data on crimes and use that information in the model, however the available dataset for crimes in chicago is rather large (>1 GB) and we didn't have time to finish extracting features from that model. We hypothesized that food deserts were more likely to be in high crime areas.

Generalized Linear Models

In ordinary linear regression (or ordinary least squares, OLS) we assume that our observations y are some linear combination of the covariates, x_i's that we have plus random error. In matrix notation this can be expressed as:

$$y = X\beta + \epsilon$$

where y is the vector of responses, X is the (model) matrix of data, β is the vector of regression coefficients, and $\epsilon \sim N(0, \sigma^2 I)$ is the random error.

Generalized linear models extends this to non-normally distributed responses through the use of a link function g:

$$Y = g^{-1}(X\beta)$$

. For 0,1 or binomially distributed responses, there are a variety of link functions available. One of the most common is the logit link: $g(x) = \log(\frac{x}{1-x})$. For simplicity, we chose the logit link although there are certainly other options available (probit, t, etc.).

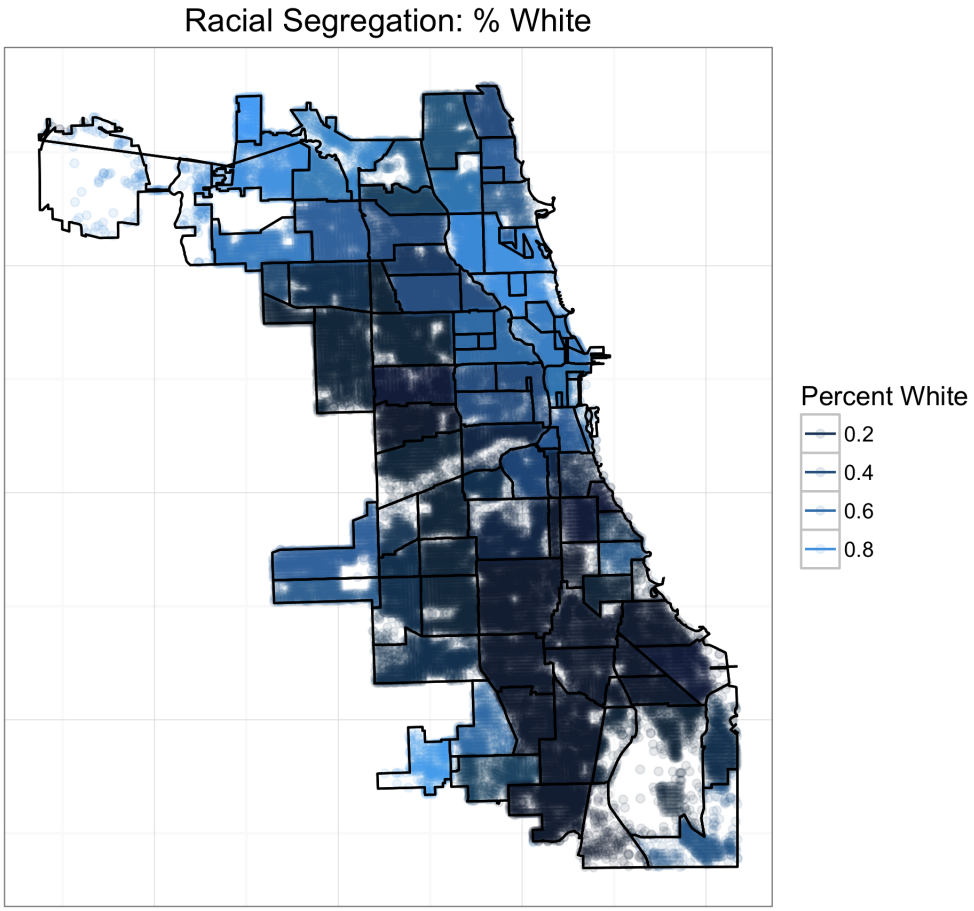

Figure 3: Neighborhoods by White Percentage

Mixed Models

So far we have discussed models which have only "fixed effects", that is, effects that are nonrandom and do not vary across covariates. We may believe for example that each neighborhood is different enough to merit its own intercept but that each neighborhood is sampled from a large population of candidate neighborhoods. In that situation we're not interested necessarily in each intercept, but more the variability between neighborhoods. In this situation we would fit a random intercept model where each intercept is (without loss of generality) distributed normally with mean zero and variance σ_α^2. To further generalize, we could allow the fixed effect coefficients to vary in a similar manner. This is called a random slope model. In our case we fit random intercept models with fixed effects for each slope. This reflects our belief that the neighborhoods are different from each other and should have different baseline probabilities of being a food desert (due to potentially unmeasured covariates) but that the effect of CTA ridership and number of vacant buildings nearby should be the same across all neighborhoods.

Hierarchical Models

Hierarchical models are another extension to the family of linear models which allow for hierarchical structured data. Recalling the random intercept model, since we have neighborhood level covariates, we would like to use the information contained within to predict the random intercept for each neighborhood.

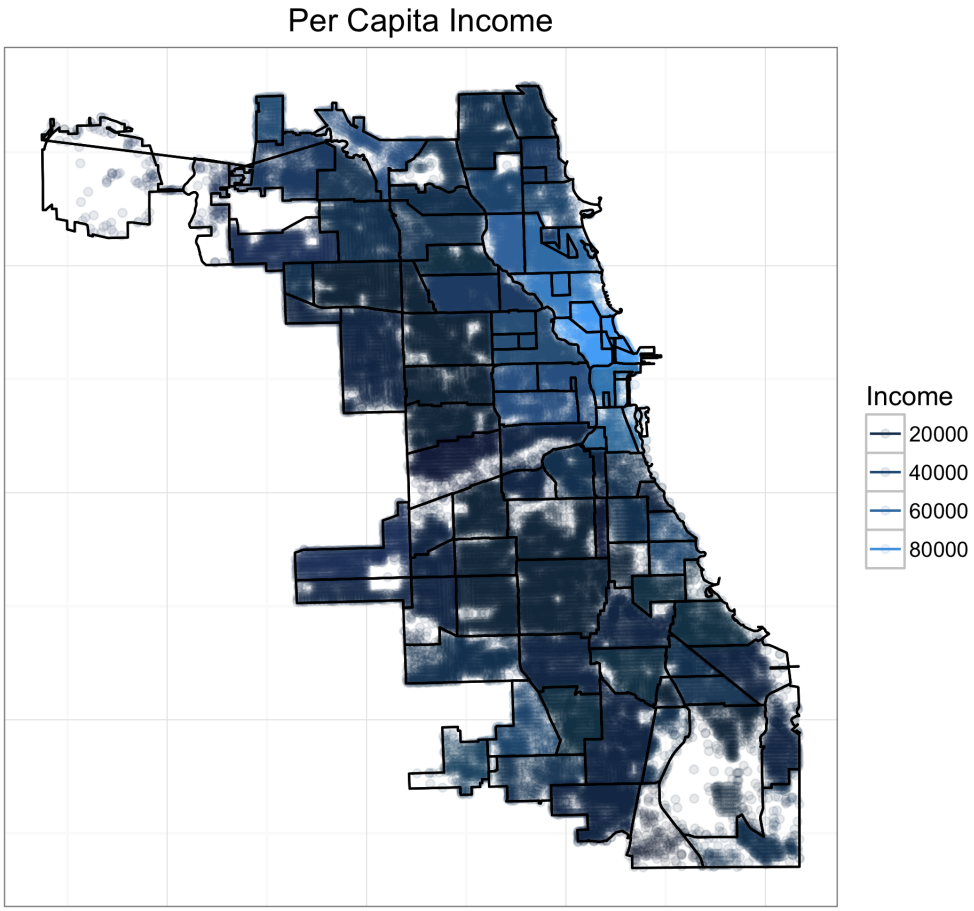

Figure 4: Neighborhoods by Per Capita Income

Models Fit

Complete Pooling

To begin we have the simplest model: ordinary regression using only the block-level variables. This model pools together every neighborhood as if the neighborhood distinctions don't matter.

$$y_{ij} = \text{logit}^{-1}\left(\alpha + X_B\beta_B + \epsilon_{ij}\right)$$

Where $\epsilon_{ij} \sim N(0, \sigma^2)$

No Pooling

The next model has a different but nonrandom intercept for each neighborhood, a fixed effect for that neighborhood. This would correspond to our belief that the neighborhoods are each different from the others.

$$y_i = \text{logit}^{-1}\left(\alpha + X_B\beta_B + \gamma_j + \epsilon_{ij}\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$

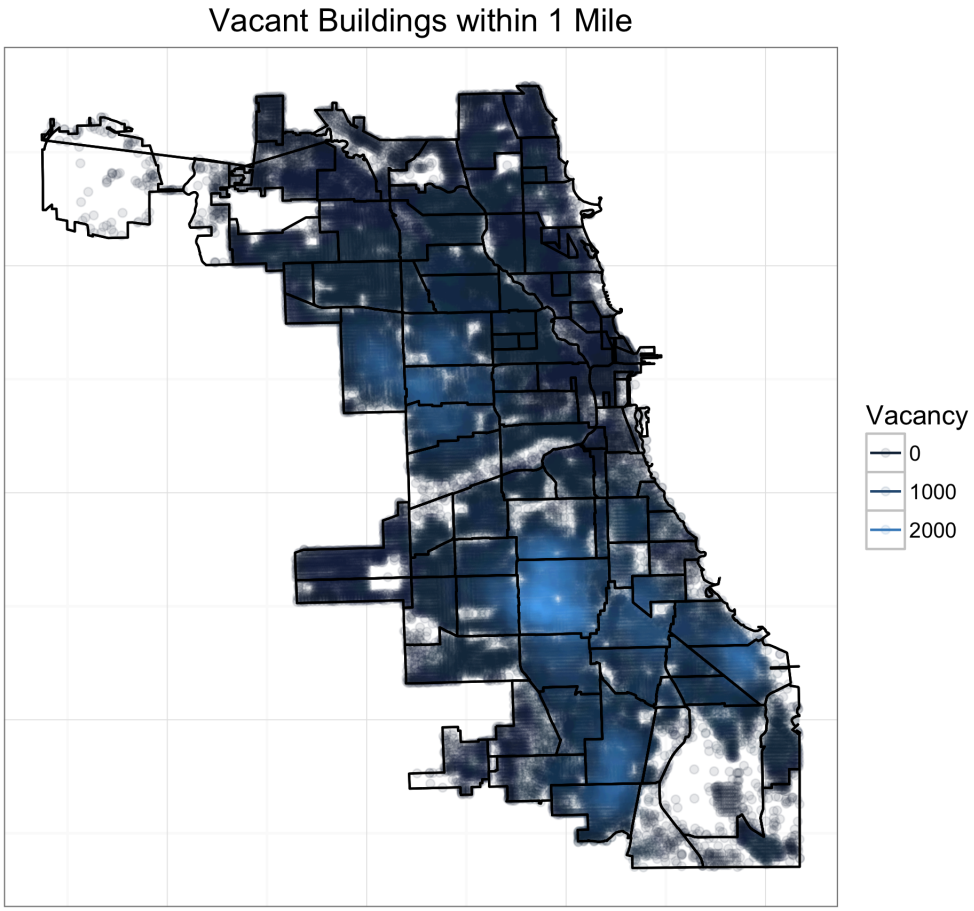

Figure 5: Neighborhoods by Number Vacant Buildings with 1 Mile Radius

Partial pooling

The next model has a random intercept for each neighborhood which corresponds to partially pooling the data together. For every neighborhood we use some of the information in other neighborhoods to estimate its intercept. That is, the intercepts in the previous model are shrunk toward the common mean.

$$y_i = \text{logit}^{-1}\left(\alpha_{j[i]} + X_B\beta_B + \epsilon_i\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$ and $\alpha_j \sim N(0, \sigma_\alpha^2)$

Hierarchical

The final and most complicated model that was fit was a hierarchical model including the neighborhood level predictors in estimating the random intercept for each neighborhood.

$$y_i = \text{logit}^{-1}\left(\alpha_{j[i]} + X_B\beta_B + \epsilon_i\right)$$

Where $\epsilon_i \sim N(0, \sigma^2)$ and $\alpha_j \sim N(X_N\beta_N, \sigma_\alpha^2)$

Model Comparison

Models were compared using AIC and cross validated Breir Score (Mean Square Error in the case of 2-class logistic regression). Lower values of AIC indicate better fitting models and thus can be used to compare the

performance of models to each other. The cross validation was performed 10 fold where the model was fit on 80% of the data (sampled in a stratified manner based on Neighborhood) and evaluated on the remaining 20%. This gives a way of quantifying the predictive ability of the model on new, unseen city blocks. .

Results

Model Parameters

Complete Pooling

Table 1: Complete pooling model summary

	Dependent variable:
	desert
CTA_counts	-1.588^{***} (0.040)
vacant_counts	0.410^{***} (0.018)
crime	0.056^{***} (0.019)
Constant	-2.849^{***} (0.029)
Observations	36,870
Log Likelihood	−9,909.356
Akaike Inf. Crit.	19,826.710
Note:	*p<0.1; **p<0.05; ***p<0.01

No Pooling

Table 2: No pooling model summary

	Dependent variable:
	desert
CTA_counts	-0.783^{***} (0.061)
vacant_counts	-0.408^{***} (0.038)
crime	0.101^{***} (0.023)
Neighborhood intercepts ommitted due to space	
Constant	−20.678 (757.423)
Observations	36,870
Log Likelihood	−6,407.134
Akaike Inf. Crit.	12,970.270
Note:	*p<0.1; **p<0.05; ***p<0.01

Partial Pooling

Hierarchical

Model Comparison

We can see from table 5 that the No Pooling model has the lowest AIC which is to be expected as in a certain sense this model has the most flexibility. The intercept term for each neighborhood is the average of only

Table 3: Partial pooling model summary

	Dependent variable:
	desert
CTA_counts	-0.794^{***} (0.061)
vacant_counts	-0.404^{***} (0.038)
crime	0.101^{***} (0.023)
Constant	-6.307^{***} (0.657)
Observations	36,870
Log Likelihood	−6,564.705
Akaike Inf. Crit.	13,139.410
Bayesian Inf. Crit.	13,181.990
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Hierarchical Model Summary

	Dependent variable:
	desert
CTA_counts	-0.779^{***} (0.061)
crime	0.102^{***} (0.023)
vacant_counts	-0.413^{***} (0.038)
Cancer..All.Sites.	3.289^{***} (0.852)
Diabetes.related	-1.954^{**} (0.801)
Dependency	1.158^{*} (0.695)
TOTAL.POPULATION	-0.147^{***} (0.048)
Constant	-6.221^{***} (0.522)
Observations	36,870
Log Likelihood	−6,540.603
Akaike Inf. Crit.	13,099.210
Bayesian Inf. Crit.	13,175.840
Note:	*p<0.1; **p<0.05; ***p<0.01

the observations in that neighborhood and is not "shrunk" to any sort of common mean. The intercepts for the hierarchical model are shrunk due to the effect of the neighborhood level regressors.

As we can see from table 6, the Hierarchical model outperforms all 4 types of models in terms of held out predictive ability.

Table 5: Model AICs

Model	AIC
Complete Pooling	19826.7
No Pooling	12970.3
Partial Pooling	13139.4
Hierarchical	13099.2

Table 6: Model Cross Validated MSEs

Model	MSE
Complete Pooling	0.07382216
No Pooling	0.05328587
Partial Pooling	0.05329956
Hierarchical	0.05323632

Conclusions

In terms of cross validated accuracy: the hierarchical model was more accurate on average on new city blocks than the other 3 models indicating support for the hierarchical structure of the data. However, the evidence was not as strong as the author would have liked. Consider the model summary in table 4. We see that food deserts tend to be located in neighborhoods with higher incidences of all site cancer. Perhaps surprisingly, in the prescence of the other information, a block in a neighborhood with higher incidences of diabetes was less likely to be in a food desert. City blocks in neighborhoods that are more populous (TOTAL.POPULATION) are less likely to be food deserts. Finally, blocks in neighborhoods with higher rates of dependency (% of the population younger than 18 or older than 64) are more likely to be in food deserts.

While we have evidence for the utility of multi-level data for modeling food desert prescence in Chicago, the knowledge in this report is likely common knowledge for anyone working in this field.

Public health variables tended to be more predictive than purely racial variables, although there is a strong correlation between race and health in Chicago.

Future Work

Some issues due to not having data from grocery stores outside the city limits, could affect food desert status of city blocks near the borders.