Predicting Life Expectancy in Baltimore at the block-level Resolution

Jacob Fiksel

October 28, 2016

1 Introduction

In 2008, the city of Baltimore released a detailed report on various health outcomes for the 55 Community Statistical Areas (CSAs) within city limits [1]. Each CSA is defined as a "cluster of neighborhoods developed by the City's Planning Department" [1]. Amongst various health inequalities between CSAs, one outcome that stood out in particular was life expectancy [1], which differed by 20 years between the CSAs with the longest and shortest life expectancies. This inequality in life expectancies within Baltimore prompted city leaders and researchers to investigate how non-health public policies might effect health outcomes [3].

Because of budgetary concerns, city leaders need to not only know which policies and programs lead to positive health outcomes, but also where to focus these efforts. The natural answer is to enact policies that will effect those in CSAs with lower life expectancies, but this assumes that there is no life expectancy inequality within a given CSA. One way of revealing whether health inequalities exist at the CSA level is to measure variables associated with life expectancy at a finer geographic scale. Two such scales are Census blocks and block groups. Census blocks "are formed by streets, roads, railroads, streams and other bodies of water, other visible physical and cultural features" [4]. A Census block group is a collection of Census blocks and is "the smallest geographic entity for which the decennial census tabulates and publishes sample data" [4]. For the remainder of the report, it will be assumed that blocks and block groups refer specifically to Census blocks and block groups.

Recent data at the block group level show that median household income, which is positively correlated with life expectancy at the CSA level [1], differs by over $85,000 within the Patterson Park CSA. Figure 1 shows a clear geographic divide between the block groups in the Patterson Park CSA based on median household income. Policies that increase life expectancy may not be targeted towards residents in the Patterson Park CSA who live in the block groups with lower median household income, due to the average CSA life expectancy being raised by residents who live south of this geographic and economic divide.

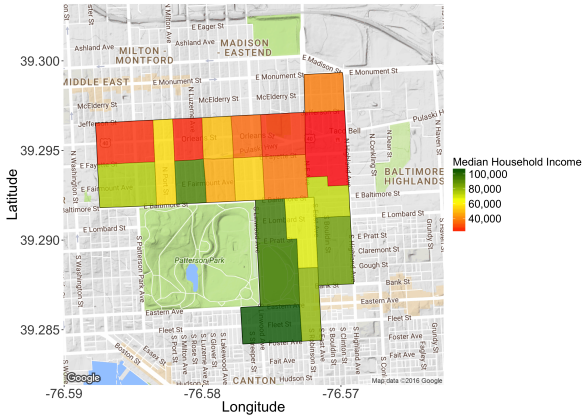

Figure 1: Inequality of block group median household incomes within CSA. Block groups within the Patterson Park CSA, outlined in black, are colored by their median household incomes. Lower median household income is in red, while higher household income is in green. Median household income begins to fall in block groups North of E. Baltimore St, with the divide becoming more clear north of Fayette St.

[1] Life expectancy is defined "as the average number of years a person born today would live if he/she experienced the mortality rates observed in [a CSA] over the course of his/her life" and the details of the calculation can be found in the technical notes section of [2]

To provide public health experts precise information on where health inequalities exist within CSAs, this report seeks to estimate life expectancy at the block level in Baltimore. We use publicly available data that can be measured at the block level and that provides surrogate measures of variables that have been shown to be associated with life expectancy. A particular challenge with predicting life expectancy at the block level is that life expectancy is only reported at the CSA level, while the variables we use to predict life expectancy are measured at a finer geographic scale. We resolve this issue by aggregating block measurements to the CSA level to model the relationship of various predictor variables to life expectancy. This model can then be used to predict life expectancy with block level measurements.

2 Methods

2.1 Literature of Determinants of Life Expectancy for Pre-Analysis Variable Selection

We first conducted a literature review to guide the selection of variables that have previously been found to be correlated with life expectancy or important health outcomes. Social factors, such as income, employment, and education, have been estimated to contribute to 85%-90% of preventable mortality in the US, and are more easily quantifiable at the block level than availability and quality of medical care [5]. Wilson and Daly corroborated these findings for income, showing that life expectancy in Chicago neighborhoods was positively correlated with median household income [6]. In addition to socioeconomic factors, racial segregation is believed to have a negative impact on health through "psychopathologic pathways" [5]. Other potential determinants of health outcomes at the neighborhood level are crime, the density of vacant buildings, alcohol availability, and concentration of fast-food restaurants [7] [8] [5].

2.2 Data Sources

The data used in this study is publicly available for download and the social determinants of life expectancy discussed in Section 2.1 were included if they were measured at the block group level or at any finer geographic resolution. Block group level data collected from 2010 through 2014 was downloaded through the *acs* package [9] in R. Measurements on all other variables were downloaded through the OpenBaltimore [10] API in R.

To assist in visualization and assignment of blocks and block groups to CSAs, shapefiles were downloaded for all three geographical units. We downloaded the shapefile for Baltimore City CSAs from the Baltimore Neighborhood Indicators Alliance [11]. Shapefiles for the Census blocks and block groups were downloaded directly into R using the *tigris* package [12]. Any block groups or blocks for which at least one of the variables of interest was not available were not included in the analysis.

2.3 Variable Definition

At the block group level, we collected surrogate variables for income, employment, education, and segregation, in addition the total population. We measured the first three variables, respectively, using median household income, the percent of residents between 20-64 years of age that have worked in the past 12 months, and the percent of residents over 25 years of age that did not receive a high school diploma or receive an equivalent degree. Because the Census data collected on race shows that 93% of Baltimore citizens are either black or white, we chose to use the percentage of

citizens that are black in a block group as measure of racial segregation. We assign measurements made at the block group level to all blocks within that block group under the assumption that blocks are relatively homogeneous within block groups.

We measured levels of crime, liquor availability, the number of vacant houses, and concentration of fast-food restaurants by first downloading the location (in latitude and longitude) of all reported shootings and burglaries from 2014-2015, liquor stores, vacant houses, and fast food restaurants. For fast food restaurants, we extracted the locations of all McDonalds, Burger King, KFC, Taco Bell, Popeyes, and Wendys from a list of restaurants in Baltimore City. We quantify this location data by estimating the density of each of these variables across Baltimore City and interpolating this density to the center of each block. See Supplementary Information for more details.

Blocks are defined as falling in a particular block group based on tract and block group number provided in the block and block group shapefiles. Blocks and block groups are defined as falling in a particular CSA based on whether their geographic center, calculated from the shapefiles, lies within a given CSA's boundary. This information is used to aggregate measured variables to the CSA level, as described in Section 2.4. Life expectancies for each CSA are from 2014.

2.4 Aggregating Measurements to the CSA Level

All block level measurements are first averaged the block group level. These averages are then aggregated to the CSA level using a weighted average across block groups within a given CSA, with each block group's weight being the Census estimate of the population within that block group. We use this two-step weighted average procedure so that blocks in block groups that are not as densely populated do not bias CSA level aggregates to estimates not representative of the people living within a CSA.

2.5 Statistical Model

For statistical modeling, we chose to use a standard multivariate least squares regression model [13]. All variables measured at the block level were aggregated to the CSA level. Therefore, using i as an index for CSAs, f as the log function, and g as the logit function our final model was:

$$\begin{aligned} LE_i = \beta_0 &+ \beta_1 f(\text{income}_i) + \beta_2 g(\text{segregation}_i) + \beta_3 g(\text{employment}_i) \\ &+ \beta_4 g(\text{education}_i) + \beta_5 f(\text{vacancies}_i) + \beta_6 f(\text{shootings}_i) \\ &+ \beta_7 f(\text{burglaries}_i) + \beta_8 f(\text{fastfood}_i) + \beta_9 f(\text{liquor}_i) + \epsilon_i \end{aligned}$$

where LE is the CSA specific life expectancy, β_0 is an intercept term and income is the aggregated median household income. Segregation, employment, and education are, respectively, the aggregated percent of residents that are black, percent of residents between 20-64 years of age that have worked in the past 12 months, and the percent of residents over 25 years of age that had not graduated from high school or received an equivalent degree. Vacanies, shootings, burglaries, fastfood, and liquor are, respectively, the aggregated kernel density estimates for the following variables: locations of vacancies, shootings and robberies between 2014 and 2015, fast food restaurants, and liquor stores. The ϵ term represents random noise and we assume $\epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ across CSAs.

The log transformation was used to reduce scale effects, while the logit transformation was used transform the variables measured in percentages to take values across the real line. These same transformations were done for block level measurements. However, note that the final CSA level measurements are transformations of aggregated non-transformed data.

Standard multivariate weighted least squares and asymptotic assumptions were used to obtain the coefficients and their associated standard errors [14]. These coefficients were used to predict

life expectancy at the block level, using the measurements at a given block for all variables in the model.

3 Results

Coefficient estimates and their associated standard errors are given in Supplementary Table 1. We find the percentage of residents over 25 who have not received a high school degree or its equivalent to be the only statistically significant predictor variable, and the linear relationship between this variable and life expectancy at the CSA level is demonstrated in Supplementary Figure S2. However, Supplementary Table 2 reveals high correlation between many of our predictor variables, which may inflate standard error estimates and make inferences on individual predictors invalid. However, this should not reduce the predictive value of our model.

Our predictions reveal large potential health disparities within CSAs, with an over 30 year predicted life expectancy difference between blocks in the North Baltimore/Guilford/Homeland CSA. Figure 2 shows predicted life expectancy at the block level throughout all of Baltimore City, revealing potential health disparities in CSAs in North and Northeast Baltimore.

We evaluated the uncertainty in our model using the following cross-validation procedure. For a given CSA, we built our model at the CSA level using the variables from the other 54 CSAs. We predicted life expectancy at all blocks within that given CSA using this model and aggregated this predictions to the CSA level using the same procedure described in Section 2.4. After doing this for all 55 CSAs, we found our average median difference between predicted and actual CSA life expectancy to be 1.74 years. Thus, assuming our model holds for blocks in all cities, if we were to be given a collection of blocks that encompass a geographic area approximately the size of an average CSA in Baltimore, we would expect the aggregate of our predictions of life expectancy for all blocks to differ from the true life expectancy of that geographic area by 1.74 years.

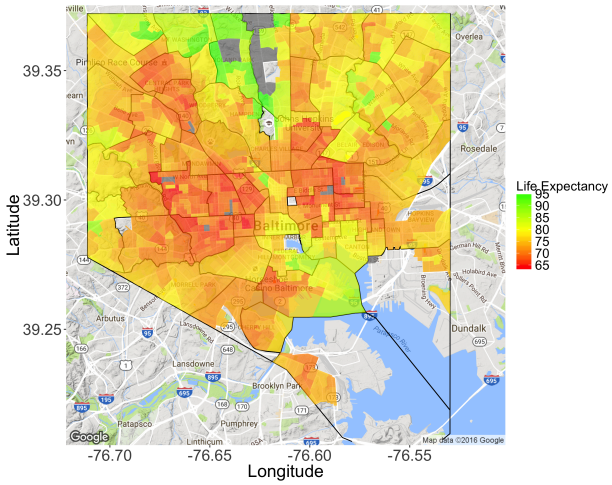

Figure 2: Predicted life expectancy at the block level in Baltimore. Each block is colored by its predicted life expectancy, with red being the lowest predicted life expectancy, and green being the highest. Any areas that are not colored in either do not have a defined block, or any of the variables included in our model is missing for that block. CSA borders are drawn in black. Predicted life expectancies are not constant within CSAs, especially in Northeast Baltimore.

4 Discussion

Our results demonstrate the potential of health disparities existing within Baltimore City CSAs. However, there are several issues to further explore in future analyses. First, our model cannot be used to ascribe the potential life expectancy disparities to any specific factor. This is not only due to the fact that our predictor variables are highly collinear, but also because this is an ecologic study using aggregated variables. A large cohort study would be better suited to determining individual factors that cause differences in health outcomes.

For our statistical analysis, we assumed independence between CSAs. However, CSAs are likely to be deeply interconnected due to the movement of people. A future analysis could attempt to model this dependence in a random effects model. However, it is not clear how CSAs are connected. One way could be geographically, with dependence being determined by the distance between any two CSAs. A more nuanced method of evaluating dependence would be to use social media to determine levels of communication between people in any two CSAs.

We also assumed that our model holds for all CSAs and blocks within Baltimore City. However, the variables included in our model may not affect life expectancy the same for blocks in the heart of Baltimore City as they do for blocks towards the outskirts of the city. In our procedure to estimate the uncertainty in our model, we find that three of the top four CSAs for which our aggregated predicted life expectancies differed the most from the reported life expectancies are located on the Northern edge of the city [1]. These CSAs may be more reflective of Baltimore County than they are of Baltimore City. Supplementary Figure S2 also reveals that CSAs that were highly influential in our model, using the Cook's Distance metric, are all located on the edge of the city.

To aggregate block level measurements to the CSA level, we used block group level population estimates. However, this ignores differences in population density that may exist within block groups. The Downtown/Seton Hill CSA demonstrates the flaw in ignoring the smaller scale population estimates. Using our cross-validation procedure, our aggregated life expectancy for this CSA differs from the true life expectancy by over 6 years. On inspection of Downtown/Seton Hill CSA, it contains office buildings, tourist attractions on the Inner Habor, as well as the University of Maryland Medical Center [1], which are not likely to have high residential population density. A method to obtain population estimates at the block level could be to look at the number of houses or apartment buildings on a given block.

In evaluating the uncertainty in our model, we are only able to compare our aggregated block predictions in a given CSA with the actual life expectancy for that CSA. This does not tell us how accurate each of our individual predictions are–if we randomly permute the block predictions for a given CSA for which the block groups have approximately equal populations, we will still arrive at the same aggregate value. A study that measured true life expectancy at a smaller geographic scale within a CSA such as Patterson Park could help validate this analysis and lead to model improvements.

Because of the issues discussed, this analysis is meant to jump-start more detailed analyses of the health disparities within Baltimore City. Using the results of this analysis, city planners and public health researchers can better target small geographic areas with low predicted life expectancy to study and implement the policies best suited to improving health outcomes.

References

[1] Baltimore City of Health. Baltimore city neighborhood health profiles: briefing to the baltimore city council, 2008.

[2] Baltimore City of Health. 2011 baltimore city neighborhood health profile, 2011.

[3] Rachel L Johnson Thornton, Amelia Greiner, Caroline M Fichtenberg, Beth J Feingold, Jonathan M Ellen, and Jacky M Jennings. Achieving a healthy zoning policy in baltimore: results of a health impact assessment of the transform baltimore zoning code rewrite. *Public Health Reports*, 128(Suppl 3):87, 2013.

[4] United states census bureau. `http://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf`. Accessed: October 28, 2016.

[5] Paula Braveman and Laura Gottlieb. The social determinants of health: it's time to consider the causes of the causes. *Public Health Reports*, 129, 2014.

[6] Margo Wilson and Martin Daly. Life expectancy, economic inequality, homicide, and reproductive timing in chicago neighbourhoods. *BMJ: British Medical Journal*, 314(7089):1271, 1997.

[7] David R Williams, Manuela V Costa, Adebola O Odunlami, and Selina A Mohammed. Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities. *Journal of public health management and practice: JPHMP*, 14(Suppl):S8, 2008.

[8] National Vacant Properties Campaign. Vacant properties: The true costs to communities, 2005.

[9] Ezra Haber Glenn. *acs: Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census*, 2016. R package version 2.0.

[10] Open baltimore. `https://data.baltimorecity.gov`. Accessed: October 28, 2016.

[11] Baltmiore neighborhood indicators alliance. `http://bniajfi.org`. Accessed: October 28, 2016.

[12] Kyle Walker. *tigris: Load Census TIGER/Line Shapefiles into R*, 2016. R package version 0.3.3.

[13] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[14] Thomas Shelburne Ferguson. *A course in large sample theory*, volume 49. Chapman & Hall London, 1996.

[15] MP Wand. Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, 3(4):433–445, 1994.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[17] Matt Wand. *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2015. R package version 2.23-15.

[18] Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. fields: Tools for spatial data, 2015. R package version 8.4-1.

[19] Roger Bivand, Tim Keitt, and Barry Rowlingson. *rgdal: Bindings for the Geospatial Data Abstraction Library*, 2016. R package version 1.1-10.

[20] Roger Bivand, Colin Rundel, Edzer Pebesma, Rainer Stuetz, and Karl Ove Hufthammer. *rgeos: Interface to Geometry Engine - Open Source (GEOS)*, 2016. R package version 0.3-21.

[21] U.s. geological survey. `https://www2.usgs.gov/faq/categories/9794/3022`. Accessed: October 28, 2016.