Life Expectancy For Each Street Block in Baltimore City

Kayode Sosina

October 28, 2016

Introduction

Life expectancy is a measure of how long an individual is expected to live on average and is commonly used in designing policy or even as a social indicator to evaluate the quality of life for any given region.[1,4]

The goal of this project is to develop a model for predicting life expectancy in Baltimore City for each street block. The aim is that with this new information we would be able to better examine what factors contribute to the life expectancy for any given block in any given neighborhood in Baltimore city and so aid decision making when policy changes are being implemented.

We have data gotten from the city of Baltimore which gives estimates of life expectancy at a Community statistical area (CSA) level. Briefly, a CSA is made up of several neighborhoods and these neighborhoods may belong to more than CSA. I.e., the boundaries of a CSA may go through a neighborhood so the CSA provides a consistent way to characterize a particular region of the city. This was done since the boundaries, and the names of the 270+ neighborhoods in Baltimore may change over time.

Since our outcome, life expectancy, is gotten at an aggregate level, the goal of this paper is to provide a street block prediction of the expected life expectancy by downscaling the expected life expectancy values at the CSA.

Data

Data was downloaded on the 14^{th} of September 2016 from the city of Baltimore[1], from the Baltimore Neighborhood Indicators Alliance–Jacob France Institute at the university of Baltimore[2], from the Maryland department of planning, and from the Census Bureau. In addition, all the block group data was downloaded from the census bureau using the *acs*[2] package in R while the *tigris*[6] package was used to download shapefiles along with the ones that were gotten from the Maryland department of planning . The datasets[3] downloaded include; real estate data which has information about ownership, street block divisions and valuation; child and family wellbeing data which has the life expectancy values, percent of babies born with a satisfactory birth weight, teen birth rate per 1,000 females (aged 15-19) along with other adult, juvenile, and infant health indicators for Baltimore City calculated for the 55 CSAs in Baltimore city.

The data downloaded generally fall in one of three categories, viz: Street level,

[1] https://data.baltimorecity.gov

[2] http://bniajfi.org

[3] For more information on these datasets along with details on what was downloaded,and what links that were used please check the data and datasets section in the supplement

Block group level and CSA level, where a block group consists of clusters of blocks within the same census tract that have the same first digit of their four-digit census block number[4]. Since goal of this paper was to downscale the life expectancy values gotten at the CSA level to the street block level, we were interested in finding what variables at the street block level which when aggregated to the CSA best explained the variation in the life expectancy values observed for the 55 CSAs. To this end, for variables that weren't observed at the street block level, we had to make a couple of assumptions[5]. 1. If the variable was observed at the street level, that is per household in a given street, then the street block value would be either the mean or the median of all households in that block. 2. If the variable was observed at the block group level, then we either interpolate new values for the street blocks by using the kth nearest neighbor approach which is based on both the spatial location of the street blocks and that of the block groups-this was taken as the median longitude and median latitude of all street blocks that made up a block group or we assume that street blocks in that block group have similar values as that of the block group to which they belong. 3. If the variable was observed at the CSA level we assume that the street blocks in that CSA have the same value for that variable.

Methods

The Model

Since our data is spatial in form, we checked to see if we had spatial autocorrelation and if there was a correlation between different variables due to their spatial location[6]. We discovered that there was a significant effect of both types of correlation at the 5% significance level, the pvalues were 0.02 and < 0.001 respectively. To account for this, we decided to fit a geographically weight regression model (GWR[5]). This model fits localized regression models through the use of a weight matrix and is defined as follows, let $Y = (y_1, \ldots, y_n)^T$ be the vector of the outcome variables, and X be the model matrix which contains the independent variables, then we have

$$Y = X\beta + \epsilon$$

where β is estimated using the generalized least squares approach. That is,

$$\hat{\beta} = (X^T W(u_i, v_i)X)^{-1} X^T W(u_i, v_i)Y$$

[4]Block Groups

[5]For more information, please check the data cleaning section of the supplement

[6]See Spatial correlation section in supplement

(u_i, v_i) is the i^{th} location and $W(u_i, v_i) = \text{Diag}[w_j(u_i, v_i)]$, $j = 1, \ldots, 54$ where $w(u_i, v_i)$ is a function which decays as the distance between location i and j increases. So let $x_i^T = (x_{i1}, \ldots, x_{ip})$ be the i^{th} row of X. Then the fitted value of y at location i is

$$\hat{y}_i = x_i^T \hat{\beta}$$

Let $M = X(X^T W(u_i, v_i)X)^{-1} X^T W(u_i, v_i)$ then M can be written as

$$\begin{pmatrix} x_1^T (X^T W(u_1, v_1)X)^{-1} X^T W(u_1, v_1) \\ x_2^T (X^T W(u_2, v_2)X)^{-1} X^T W(u_2, v_2) \\ \vdots \\ x_n^T (X^T W(u_n, v_n)X)^{-1} X^T W(u_n, v_n) \end{pmatrix}$$

and

$$Var(\hat{\beta}) = (X^T W(u_i, v_i)X)^{-1} X^T W(u_i, v_i) Var(Y) W(u_i, v_i) X (X^T W(u_i, v_i)X)^{-1}$$

where Var(Y) $= \sigma^2 I$, I is a $n \times n$ identity matrix. To fit the model, using the *GWmodel*[3] package in R, we used the following variables; proportion of households in a street block headed by a female with related children under 18 years, proportion of individuals within a street block that live below the poverty line, percent of residences heated by electricity in CSA, liquor outlet density (per 1,000 Residents) in a CSA, and the percentage of students suspended or expelled during school year all in the year 2014 [7].

Downscaling

Delta Method

The basic idea behind this method is to see the effect that each street block has on the predicted value at the CSA level after we remove it during aggregation. This is done in three steps; 1. Remove one of the blocks from the aggregated data and then get the predicted life expectancy at the CSA level that is due to the remaining street blocks; 2. Subtract this predicted value from the fitted value at the CSA when this street block is not removed; 3. Add the difference obtained to the observed life expectancy at the CSA that this street block belongs to and called that the predicted life expectancy for that street block[8].

[7] For more information on how the variables that were used in the GWR model were selected along with how the weights were defined, please see the GWR section of the supplement.

[8] see delta method section in supplement for more details

Street Block prediction/Transfer

The assumption here that the aggregate level relationships will be the same as the one at the street block level. To reduce the effect of misspecification, I centered and scaled both at the CSA level and then centered and scaled at the street block level. So both levels had a mean of zero and variance equal to 1. Then prediction at the street block level is

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_{n_s} \end{pmatrix} = \begin{pmatrix} x_1^T (X^T W(u_1, v_1) X)^{-1} X^T W(u_1, v_1) \\ \vdots \\ x_{n_s}^T (X^T W(u_{n_s}, v_{n_s}) X)^{-1} X^T W(u_{n_s}, v_{n_s}) \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_{54} \end{pmatrix}$$

where n_s is the number of street blocks.

Results

After dividing the datasets into a testing and training datasets. We used the two methods mentioned above and plotted it. After aggregating the street blocks to the CSA level, we found that the delta method gave estimates that were closer to the observed values at the CSA, figure 1. We also plotted the predicted life expectancy values and the variance of the prediction at the street block level, figure 2, for each CSA. In general we see that street blocks which are contiguous do not necessarily have similar life expectancy values if they they belong to different CSAs. Furthermore, even when they are within the same CSA they do not all have similar values and the change can be quite sudden. This is a bit comforting since blocks in Baltimore city that are within walking distances of one another are know to be quite heterogeneous. Note that in figure 2, when we examine both plots we see that street blocks which are downtown (middle of the plot) have a higher amount of uncertainty associated with the estimate, which seems reasonable as this region includes both Mount Vernon and the Inner harbor, which have a more transient population when compared to the other regions.

Discussion

As noted above we see a pattern in the predicted life expectancy values at the street block level, which is consistent with the nature of street blocks in Baltimore. That is, street blocks which abut one another do not necessarily have similar demographics and so we might expect this to be reflected in some way in the life expectancy values. However, there are a couple of shortcomings in this paper; 1. The datasets that were used were not collected at the same time period; 2. The method of interpolation also imposes a structure on the distribution of the datasets

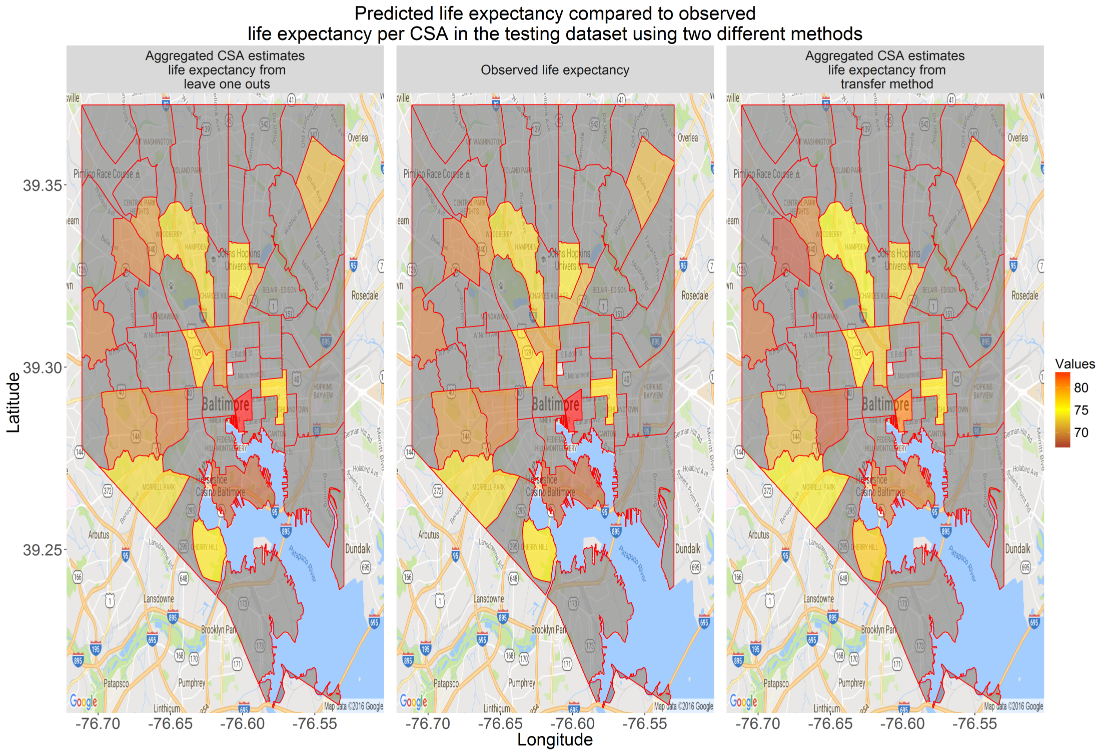

Figure 1: Predicted life expectancy compared to the observed life expectancy in the testing dataset using the two methods mentioned. Note that the grey areas represent CSAs that were used in the training dataset

at the street block level that could be biased; 3. For some variables, we had to assume that the value at the street block level corresponds with the value at the CSA level. This is a strong assumption, and may not be likely to be true depending on the variable we are talking about; 4. Each of the downscaling approaches rely unpon assumptions that may not be true. For instance in the delta method, it is assumed that there is a true underlying life expectancy value for each street block

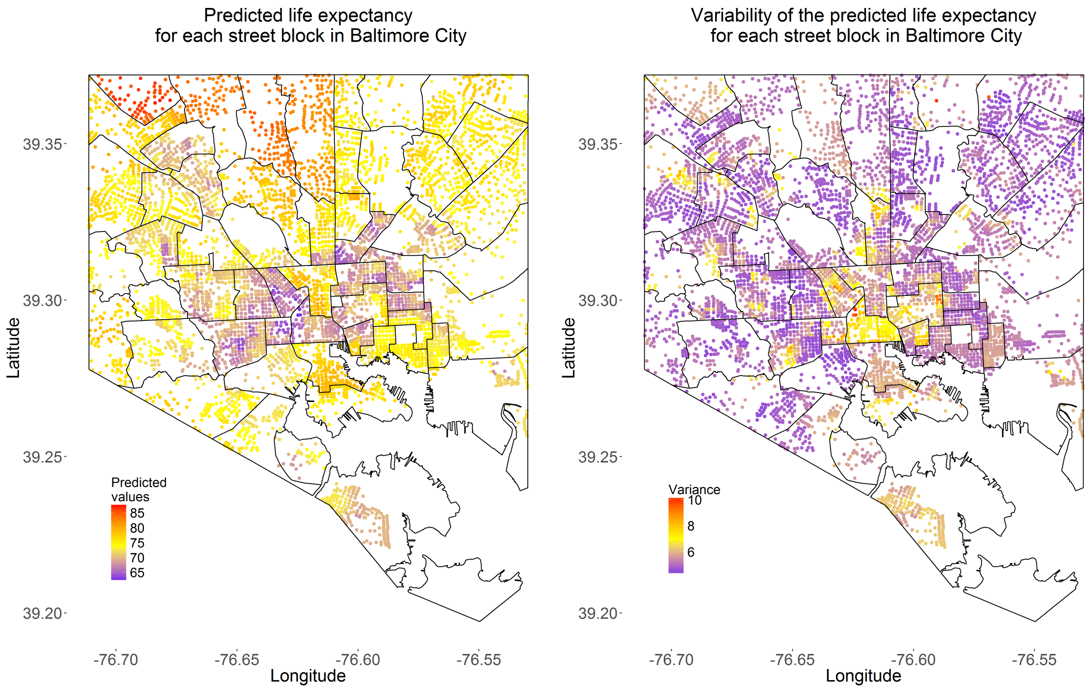

Figure 2: Predicted life expectancy estimates for each street block along with estimates of uncertainty within each CSA based on the street Block prediction approach

in a particular csa and each street block deviates from this randomly. While it may be true that each street block deviates randomly from the mean life expectancy in a particular cluster, we have no way knowing if this cluster will correspond to the CSA that the street block belongs to. However, even with the shortcomings mentioned above, we believe that predicted estimates still give a good suggestion as to what the distribution of life expectancy values at the street block level could be, and so might aid in decision making which involve policy changes.

Bibliography

[1] "An overarching health indicator for the Post-2015 Development Agenda". In: (2014). URL: http://www.who.int/healthinfo/indicators/hsi_indicators_SDG_TechnicalMeeting_December2015_BackgroundPaper.pdf.

[2] Ezra Haber Glenn. *acs: Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census*. R package version 2.0. 2016. URL: https://CRAN.R-project.org/package=acs.

[3] Isabella Gollini et al. "GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models". In: *Journal of Statistical Software* 63.17 (2015), pp. 1–50. URL: http://www.jstatsoft.org/v63/i17/.

[4] "Human Development Index". In: (2015). URL: http://hdr.undp.org/en/content/human-development-index-hdi.

[5] Fotheringham A. Stewart, Brunsdon Chris, and Charlton Martin. "Geographically Weighted Regression: The Analysis of Spatially Varying Relationships". In: *Wiley* (2002).

[6] Kyle Walker. *tigris: Load Census TIGER/Line Shapefiles into R*. R package version 0.3.3. 2016. URL: https://CRAN.R-project.org/package=tigris.