Downsample Life Expectancy on Single Block Level in Baltimore City

Yifan Zhou

1. Introduction

Health is directly related to health care, health behaviors and family health history, and it's also heavily influenced by socioeconomic position, race-ethnicity, and social cohesion[1]. Life expectancy is the average number of years a newborn can expect to live, assuming she or he experiences the currently prevailing rates of death throughout her or his lifespan[2]. Life expectancy is often used as a fator to reflect health condition.

According to 2011 Neighborhood Health Profile[3], there are lots of variations in life expectancy between different neighborhoods in Baltimore City. Even two nearby neighborhoods could have a huge gap in life expectany. For example, life expectancy is 77.1 in Inner Harbor/Federal Hill, while it's 63.9 in its nearby neighborhood Downtown/Seton Hill. Therefore, the inequality in Baltimore has been a big issue.

In this study, we focus on downsampling life expectancy on block level. Our goal is to develop a model for predicting life expectancy in Baltimore City down to single block resolution with estimates of uncertainty.

2. Material

2.1 Definition

For geographic information in Baltimore, we use Tiger shape files from Census 2010 data (available at the Maryland Department of Planning websit[4]), from which we obtained block ID, coordinates and footprint for each block. There are 13488 blocks defined in this way. We use the coordinate of latitude and longitude for the polygon center of each block as its location. To obtain coordinates of latitude and longitude, we use Google Maps API through R package "ggmap"[5]. "RANN"[6] package is also used to find the nearest location. Then we use "sp"[7,8] package to calculate distance between two locations, which returns the Euclidean distance between two polygon centers on map.

2.2 Data

We use life expectancy data in 2014 for Baltimore City on neighborhood level as the outcome variable (available at BNIA-J websit[9]). There are 55 neighborhoods in Baltimore City, we also use Census 2010' Tiger shape files to get the geographic information of those neighborhoods.

Since life expectancy is associated with many factors such as family disease history, individual health condition, environment and socioeconomic factors[10]. We consider those following factors as the potential predictors for life expectancy: Health environment, housing and development, financial condition, education, culture and art, public safety. Since it's impossible to directly obtain the those factors for each block, we generate independent variabes for each block in R program. Table 1 lists the independent variables used in our final model and how they are generated base on each block's location. All original datasets are available at "Open Baltimore"[11].

For independent variables on neighborhood level, it will cause trouble if we generate them using the same pipeline for generating block level variables, because the area of neighborhood is much larger than the area of block. In stead, we integrated the variables for blocks in the same neighborhood to obtain predictors

Table 1: Independent variables on block level

Variable	Description	Original Dataset
neighborhood	Indicator of which neighborhood the block belongs to	Tiger shape files from Census 2010
arrest rate	Number of people arrested in the surrounding area (220m)	BPD Arrests Data
shooting rate	Number of shootings in the surrounding area (220m)	BPD Part 1 Victim Based Crime Data
crime rate	Number of crimes in the surrounding area (2km)	BPD Part 1 Victim Based Crime Data
distance to library	Distance (km) to the nearest library	Baltimore City Public Libraries
distance to hospital	Distance (km) to the nearest hospital	Baltimore City Hospitals
liquor store density	Number of liquor stores in the surrounding area (1km)	Liquor Licenses
pubart density	Number of public arts in the surrounding area (2km)	Designated Landmarks, Museums, Monuments
vacant building density	Number of vacant buildings in the surrounding area (350m)	Vacant Builidngs
property tax	Average Real property tax of the neares 3 houses	Real Property Taxes
stressed	Indicator of the residential market condition: "0" for cluster A/B/C/D, means "Popular on the market", "1" for cluster E/F/G/H, means "Stressed on the market"	2014 Housing Market Typology

on neighborhood level. Since there is no available dataset containing popluation information for each block, we choose area of blocks as the unit to do the integration. For all numeric independent variables on neighborhood level, we calculate the weighted average of variables for blocks in each neighborhood, and the weight for each block is in proportion to the acre:

$$var_i^N = \sum_{\text{block j in neighboor i}} var_j \times \frac{\text{acre}_j}{\sum_{\text{block j in neighboor i}} \text{acre}_j} \quad , \quad i = 1, 2, ..., 55$$

For the indicator of residential market condition— stressed, we use its mode for blocks in each neighborhood as the independent variable on neighborhood level.

3. Method

We want to predict life expectancy for the 13488 blocks in Baltimore City. However, we could only obtain the outcome variable on neighborhood level. Therefore we couldn't build up the prediction model directly on block level. In this project, we first train a prediction model with neighborhood level training data, then we plug in the block level predictors into this model and obtain the prediction values for each block. Since the outcome variable is continuous with a relatively narrow range (from 67.16 to 89.62 years) and a relatively small standard deviation which is 4.53 years, we decide to fit a linear model.

3.1 Model

Appropriate data transformations would be helpful to improve model performance. We did log transformation for property tax, vacant building density and arrest rate since those variables are highly skewed. From the exploratory plots in Fig 1, we could see life expectancy has non-linear relationships with log property tax, log vacant building density, shooting rate, distance to hospital and distance to library. Since only the results of likelihood ratio tests for shooting rate don't supports the non-linear tranformation of shooting rate, we fianlly include non-linear terms for those variables except shooting rate. Also, we include the interaction of the residential market indicator, stressed, with public art density, liquor store density, non-linear terms of crime rate, distance to library and distance to hospital.

Then, we fit a full model and performe several likelihood ratio tests to drop non-significant terms and decide the fianl model. The final predition model is a linear regression model with life expectancy as the outcome, selected predictors of this model contain following terms: Indicator of neighborhood, log of arrest rate, shooting rate, dummy variable of stressed, public art density, natrual spline for log of property tax with knots at 6 and 7.2, natrual spline with 2 degrees of freedom for crime rate and log of vacant buildings density, natrual spline for distance to hospital with 3 degrees of freedom and their interaction terms with stressed, natrual spline for distance to library with 2 degrees of freedom and their interaction

terms with stressed.

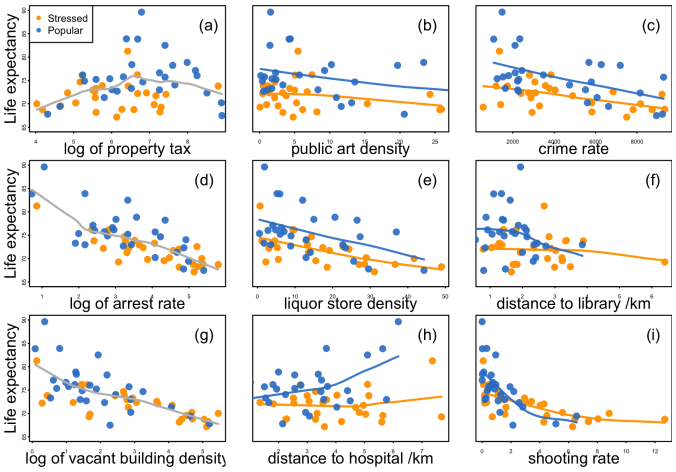

Fig 1: Relationship between life expectancy against predictors on neighborhood level. Solid lines are smooth curves fitted by Loess with gaussian family (span=0.5). Grey lines are based on all observations; Orange/blue indicates whether the residential market condition is "stressed" or not. **(a)** Life expectancy against log of property tax; **(b)** Life expectancy against public art density; **(c)** Life expectancy against crime rate; **(d)** Life expectancy against log of arrest rate; **(e)** Life expectancy against liquor store density; **(f)** Life expectancy against distance to library; **(g)** Life expectancy against log of vacant building density; **(h)** Life expectancy against distance to hospital; **(i)** Life expectancy against shooting rate;

3.2 Assumptions Checking

First, we check for the independence assumption by checking the correlation between two neighborhoods' residuals, and plot residuals against all the continuous predictors. Results shows there is no significant correlation between residuals themselve and predictors. Then, we check the constant variance assumption by plotting residuals against fitted values. Although the spread of the residuals is a little wide for fitted values larger than 80 which may be caused by the small number of training data, we don't think there is a violation of equal variance assumption. Q-Q plot demonstrates the normality and the plots of residuals against predictors and their smooth curve defence the linearity assumption. The adjusted r-squared is 0.851 for the fianl model which demonstrates the goodness of fit in this model. All those checking procedures show this prediction model is reasonable and reliable. (Results are shown in Appendix)

4. Result

we obtain the estimated life expectancy for each block using predictors on block level and the model fitted above. However, not all of the 13488 blocks are residential ares (they could be commercial or industrial areas, green space or institutional areas), therefore we treat the outcome variables in those blocks as NULL value, which are displayed by grey area on the map. Fig 2 (a) shows the estimated life expectancy down to single block level.

To measure the uncertainty of the estimated life expectancy for each block, we calculate the 95% confidence interval for predited values. Fig 2 (b) and (c) are the lower bound and upper bound of the 95% confidence interval for life expectancy, which also display the uncertainty of the esimated results.

In Fig 2, plenty of block spefic patterns could be eaily recognized, especially in downtown, East Balimore and West Baltimore areas. Even two nearby blocks would have very different estimated life expectancy. Meanwhile, the uncertainty of the estimated values varies a lot in those areas. And most of the blocks with large variation of life expectancy are located in West Baltimore and East Baltimore area.

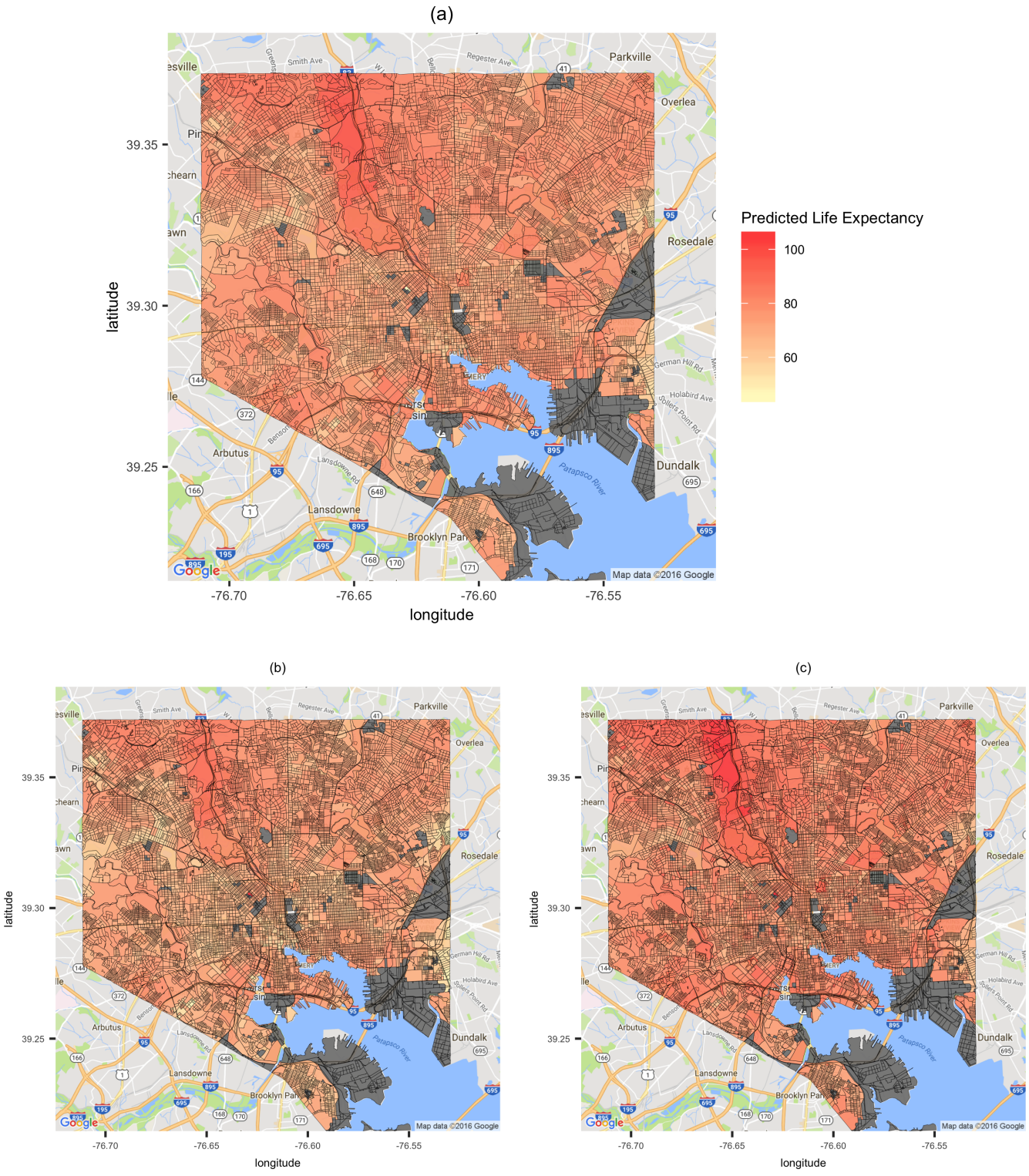

Fig 2: Prediction result of life expectancy in Baitimore City with 95% confidence interval. Grey areas are non-residential space. **(a)** Estimated life expectancy on block level; **(b)** Lower bound on the confidence interval of the estimated life expectancy; **(b)** Upper bound on the confidence interval of the estimated life expectancy;

There is one interesting result, we calculated the estimated life expectancy of the nearest block from 5 Johns Hopkins campus and list them on Table 2 below. Frome Table 2, we could see homewood campus

has the longest estimated life expectancy, 87 years with 95% confidence interval [82.8, 91.2]. The prediction result for Bloomberg School of Public Health is 72.5 years with confidence interval [69.1, 75.9], which is slightly lower than the mean of life expectancy among all blocks in Baltimore City (74.08 years). Those results are consistent with the campus saftey conditions. Carey Busniess School has the shortest estimated life expectancy with a large uncertainty, which is 65.13 years with 95% confidence interval [59.2, 71.1]. One possible reason is this block has a very large crime rates 7280, comparing with the mean crime rate 4170 and the 75% quantile 6160 among all blocks.

Table 2: Prediction results for 5 campus of Johns Hopkins University

Campus	Neighborhood	Life Expectancy	95 Confidence Interval	Address
Homewood Campus	Cross-Country/Cheswolde	87.0	[82.8, 91.2]	Charles St & 34th St N/B
School of Public Health	Oldtown/Middle East	72.5	[69.1, 75.9]	615 N Wolfe St
Peabody	Midtown	73.0	[68.4, 77.6]	18 E Mt Vernon Pl
Carey Business School	Harbor East/Little Italy	65.1	[59.2, 71.1]	100 International Drive
Bayview Medical Center,	Orangeville/East Highlandtown	68.6	[63.2, 74.0]	401 Anglesea St

We also calculate the weighted average of estimated life expectancy of blocks within each neighborhood and compare them with the original data on neighborhood level from Census 2010 (See Table 3 in Appendix). After integrating the estimated life expectancy, we could see Downtown/Seton Hill has the lowest weighted average life expectancy 68.9 years, and Mount Washington/Coldspring is the highest one with estimated life expectancy 93.9 years. This result is rational according to the previous discussion. For the uncertainty of predictions, Mount Washington/Coldspring has the largest weighted average of standard errors (4.07 years), and Hamilton has the smallest one (1.48 years).

5. Discussion

In this study, we build up an approach to downsample life expectancy into block level. Then we use this model to predict life expectancy in 13488 blocks in Baltimore City and measure the uncertainty using 95% confidence interval. We also check for the assumptions to verify the reliability of this model.

The patterns in Fig 2 shows the inequality in Baltimore is not only exists between different neighborhoods, it also exist among different blocks. Even for blocks on the same street, they may have a wide range of life expectancy. And the variation of the life expectancy between blocks is much larger in Downtown, East Baltimore and West Baltimore than other areas. These results are consistent with the realistic society in Baltimore City.

There are still some future works could be done in this research question. Since we could only obtain life expectancy data for 55 neighborhoods, the training datasets is relatively small for building a good model. One potential approach is to collect data from cities similar to Baltimore such as Philadelphia and Pittsburgh, then we could enlarge the training dataset and train the prediction model. The other approach is still using life expectancy data of the 55 neighborhoods, but we can use the data on different years and treat them as longitudinal data. In this study, we only use life expectancy data on 2014, and aggregate all data despite of the year. If we could obtain data on different year we could try mix effect model or other methods for longitudinal data analysis.

6. Reference

1. Woolf SH BP. Where health disparities begin: The role of social and economic determinants— and why current policies could make matters worse. Health Affairs: 2011.30:1852-1859

2. Wikipedia: https://en.wikipedia.org/wiki/Life_expectancy

3. Ames A, Evans M, Fox L, Milam A, Petteway R, Rutledge R. 2011 Neighborhood Health Profile. Baltimore City Health Department, December 2011

4. Department of Planning. Census 2010: http://www.mdp.state.md.us/msdc/S5_Map_GIS.Shtml

5. D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161.

6. Sunil Arya, David Mount, Samuel E. Kemp and Gregory Jefferis (2015). RANN: Fast Nearest Neighbour Search (Wraps Arya and Mount's ANN Library). R package version 2.5. https://CRAN.R-project.org/package=RANN

7. Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2), http://cran.r-project.org/doc/Rnews/.

8. Roger S. Bivand, Edzer Pebesma, Virgilio Gomez-Rubio, 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

9. Seema Lyer, Brandon Nida, Zak Bickel, et al. Jacob France Institute. BNIA-JFI: http://bniajfi.org/vital_signs/data_downloads/

10. Benjamin F. Evans, et al. Neighborhood Characteristics and Health in Baltimore, Maryland. Virginia Commonwealth University Center on Human Needs, 2012

11. Open Baltimore: https://data.baltimorecity.gov/