Predicting Life Expectancy on a Street Block Level in Baltimore, MD

Anton Kvit

October 28, 2016

Introduction:

Baltimore City, MD, is notorious for large discrepancies in the health of its population, which result in almost a 20 year life expectancy gap between neighborhoods only a few miles apart1. While there are many historical, economic, and social reasons that contribute to this phenomenon, perosnal income, as well as economic development of the surrounding area seem to play a major role2. Furthermore, life expectancy was found to be associated with race and educational attainment, 3, as well as neighborhood crime levels4.

Most health-related data in Baltimore is collected on a neighborhood or Combined Statistical Area (CSA) level, however, due to the city's distinct economic and racial divisions, this might be too coarse of a tool to isolate areas where health interventions are needed the most. Thus we developed a model attempting to predict life expectancy on a street block level, using information from CSA's as well as point locations throughout the city. Exploratory analysis showed that life expectancy in Baltimore is spatially dependent, which has to be accounted for, and that crime, race, and economic factors were all associated with life expectancy.

Methods:

Data Collection Data sources include: Baltimore Neighborhood Indicators Alliance, the official website of Baltimore City, and the Department of Planning of Maryland. Data includes information on CSA-level race, age, gender distributions, household income, households in poverty, Birthweight, elevated blood lead levels in children, life expectancy, mortality by age, liquor store density, fast food density, crime rates, and property values, among other variables. A shapefile of blocks in Maryland was obtained from the Department of Planning. The datasets and shapefiles were merged and cleaned using both R and ArcGIS 10.4.1, resulting in one shapefile with block-level data, and one shape file with CSA level data. Figure 1 contains Baltimore life expectancy values for 2011, and is representative of the format in which all the CSA data was obtained.

Exploratory Analysis Exploratory analysis was conducted to check the completeness of the data, to determine which variables are related to life expectancy, and to check for spatial autocorrelation. It was determined that in order to make variables as comporable as possible between CSA's and city blocks, they should be represented by deviation from their mean divided by its variance. For the same reason, crime and vacant builing counts were transformed into per CSA and per block densities. Using Moran's I correlograms, it was determined that life expectancy is in fact spatially autocorrelated, which should be accounted for. Nine blocks located in the Oldtown/Middle East area were found to contain no information from CSA's, due to a hole in the CSA shapefile. The missing values were assumed to be the same as the ones in the Oldtown/Middle East CSA. Baltimore life expetancy from 2011 was determined as most apropriate to build a prediction model, since that was the year for which most variables were available.

Statistical Modeling A generalized linear model was used to predict life expetancy at a street block level. In order to scale down from CSA level, both CSA and point-level data, including crime and vacant building density was used in the analysis. In order to account for spatial autocorrelation, geographic coordinates were included in the model. The final regression model used in the analysis included crime density, vacant house density, percent of households headed by women, racial diversity, and finally latitude and longitude of block

centroids. In order to create the model, 30 CSA's were randomly chosen as a training set, and 25 as a test set, the result of which was then applied to the street block data.

Reproducibility An R markdown document provides the code to obtain all the data, and conduct all the statistical analysis presented here. Additionally, the appendix provides all the steps taken in ArcGIS to spatially join necessary files in order to produce the final shapefiles used in analysis.

Results

The variables used in the final model included the precentage of households headed by females, racial diversity (defined as the percent chance that two people picked at random in the area will be of different race), crime density, vacant home denisty, and the latitude and longitude of street block centroids. Figure 2 displays the predicted life expectancy by street block. The confidence intervals for these predictions ranged from 11.15 to 30.67 years in width, and are displayed in Figure 3.

Discussion

There are several important limitations to the prediction model described above. Most importantly, down-sampling from CSA's to street blocks was accomplished by simply calculating the crime and vacant house densities within each block and within each CSA, and using these variables at both block and CSA levels. In order for this to work, we have to assume that the associations between these variables and life expetancy on a CSA level follows the same pattern as they do on a block level. In order to test this idea, one could "scale up" and compare CSA data with the same data within all counties of the state. Once the relationship between county and CSA has been established, one could extraploate this relationship down to the block level. Furthermore, important factors such as personal income and access to healthcare were not directly included in this analysis. Instead, City and State tax information was analyzed, since it was assumed that these variables might be higly related to income.

As we demonstrated above, while most data related to life expectancy is avaliable either on a CSA scale or on a point level scale, one could combine these two sources of data in order to predict life expetancy on a street block scale. Such an improvement in spatial resolution could help public health officials and policy makers target areas that need help the most with higher precision.

Acknowledgements

I greatly appreciate the advice of Kayode Sosina and Yifan Zhou regarding obtaining data and choosing the proper staistical analyses for this project.

Figures

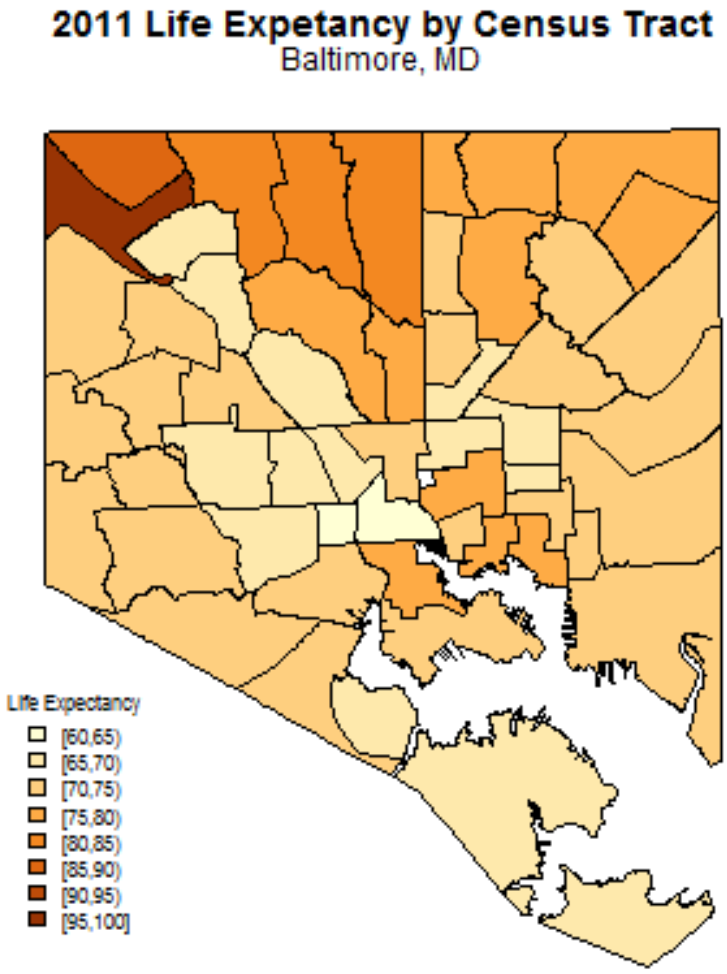

Figure 1: Baltimore, MD life expectancy by CSA in 2011.

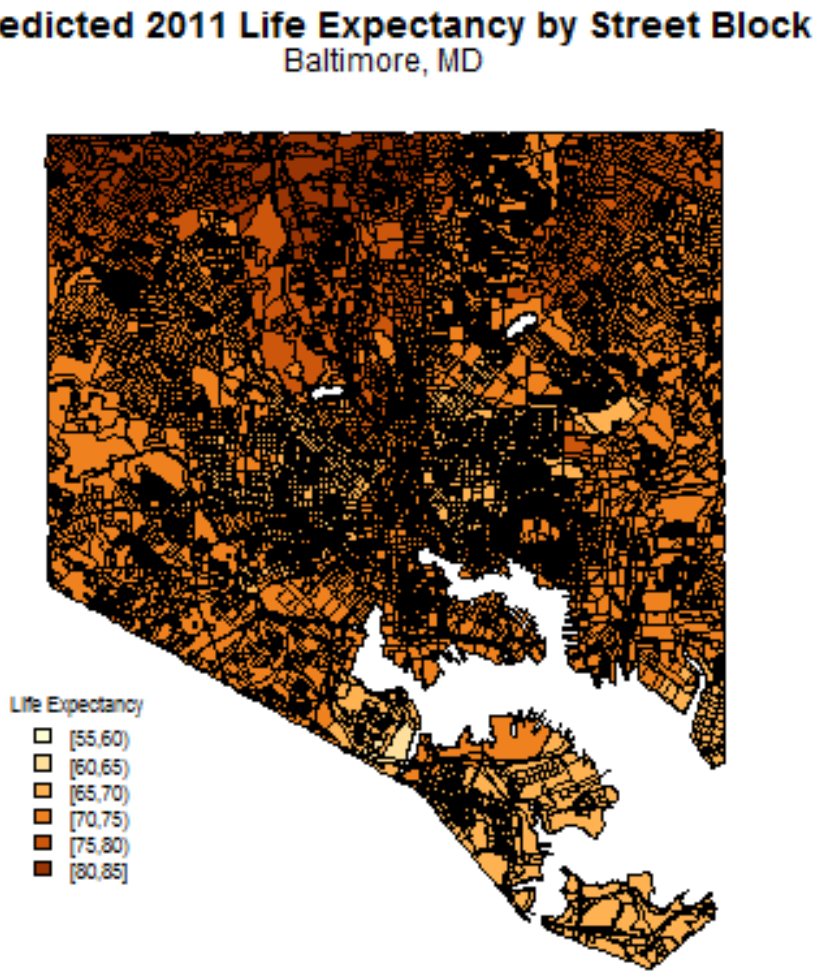

Figure 2: The predicted life expetancy by street block in Baltimore, MD.

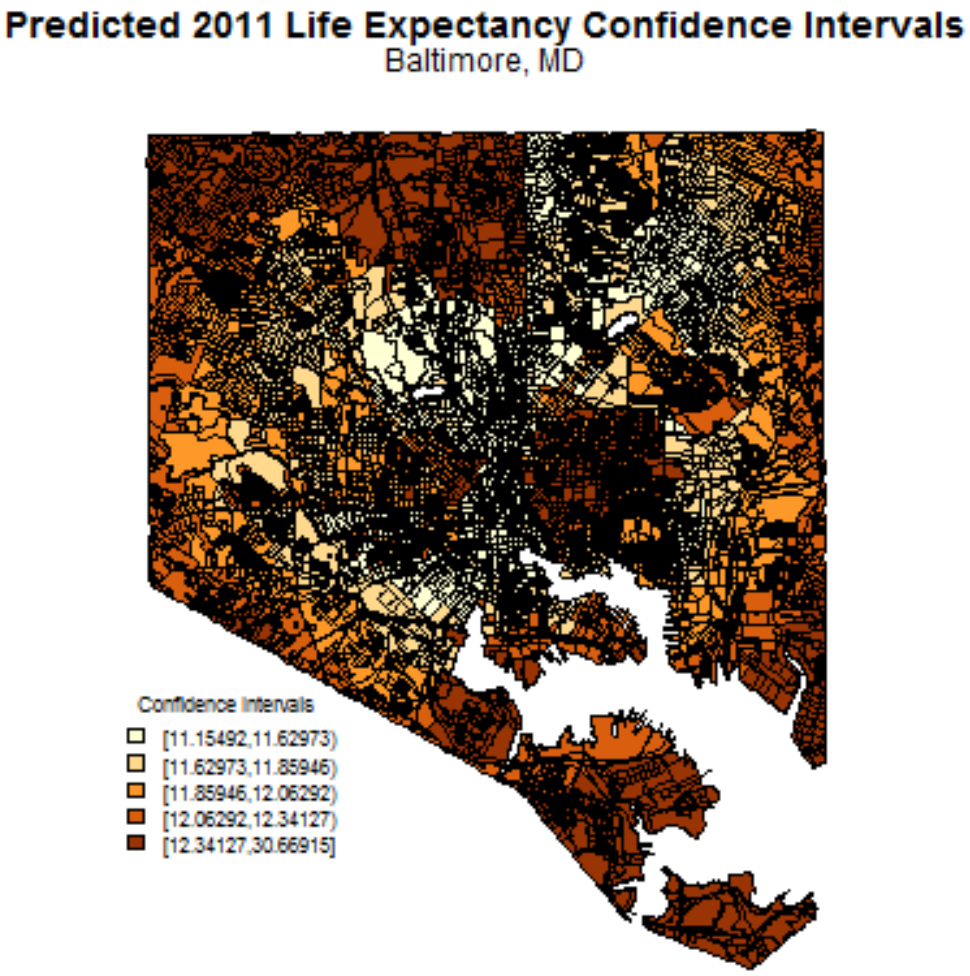

Figure 3: The confidence interval width for the prediction of life expectancy in each block. The width ranged from approximately 11 to approximately 31 years, with wider intervals representing higher prediction uncertainty.

References

1. Ames et al., Baltimore City 2011 Neighborhood Health Profile, 2011.
2. Chetty R, Stepner M, Abraham S, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA. 2016;315(16):1750-66.
3. Olshansky SJ, Antonucci T, Berkman L, et al. Differences in life expectancy due to race and educational differences are widening, and many may not catch up. Health Aff (Millwood). 2012;31(8):1803-13.
4. Redelings M, Lieb L, Sorvillo F. Years off your life? The effects of homicide on life expectancy by neighborhood and race/ethnicity in Los Angeles county. J Urban Health. 2010;87(4):670-6.
5. Neighborhood-Level Determinants of Life Expectancy in Oakland, CA., 2012, Virginia Commonwealth University, Richmond Vrigina.

Appendix

Merging Shapefiles in ArcGIS ArcGIS 10.4.1 for Desktop was used to merge shapefiles in order to obtain one shapefile with block-level data, and one shape file with CSA level data.

The following steps were taken:

1. Shapefiles Real_Property.shp, blk2010.shp, VS14_Health.shp, VS14_Census.shp, crime_.csv, estate_.csv, and vacant_.csv were uploaded into ArcMap. The projection for all the shapefiles was set to NAD 1983 Maryland State Plane (feet)

2. County 24510 (Baltimore County) was selected from the blk210 shapefile and saved as a separate shapefile (balt.shp)

3. Using the Spatial Join tool, crime point data (crime_.csv), estate tax point data (estate_.csv), and vacant building point data (vacant_.csv) was joined (many to one) to the street block shapefile (balt.shp), resulting in the shapefile balt_vac_crime_estate.shp

4. The Spatial Join tool was used to join (one to one) the CSA shapefile (VS14_Census.shp), and the Health shapefile (VS14_Health.shp) to the Blatimore blocks shapefile created in the previous step (balt_vac_crime_estate.shp)

5. The resulting shapefile was named balt_vac_crime_estate_cen_health_feet.shp, and saved in "Analyzed_Data/ArcGIS". It contained all the necessary data on the street block scale.

6. In order to get all the data on CSA level, the exact same steps were repeated, but without using the block-level shapefile. Thus, the CSA shapefile (VS14_Census.shp) was joined to the Health shapefile (VS14_Health.shp) and then joined to the the crime point data (crime_.csv), estate tax point data (estate_.csv), and vacant building point data (vacant_.csv). The resulting shapefile was named track_health_cens_crim_est_vac.shp and also saved in "Analyzed_Data/ArcGIS".

7. All the subsequent data cleaning and analysis was conducted using R 3.3.1