Supplemental Information: Predicting Life Expectancy in Baltimore at the block-level Resolution

Jacob Fiksel

October 28, 2016

Quantification of Location Data

Because the data for which we only have the latitude and longitude is not measured on the block or block group level it is not immediately obvious how to transform these measurements into quantitative variables that can be included in an analysis. The simple, and perhaps most naive solution, is to simply count the number of location points for each variable that occur in a block or block group. However, this means many blocks or block groups will be assigned a 0 for more sparse variables, like the location of fast food restaurants, even if they are in an area that is relatively more populated with fast food restaurants.

The approach we take is to first construct a binned 2D kernel density estimate over grid points spanning Baltimore City. We then interpolate the kernel density estimates for the grid points to the geographic center of each Census block (these steps are done using the *KernSmooth* package [1] and the *fields* package [2], respectively). Although this approach does not give exact kernel density estimates for the centers of each block, we believe that it is a relatively accurate measure of density that is easily implementable and computationally efficient.

Described in [3], the binned 2D kernel estimate over grid points is computed as follows.

1. We divide Baltimore City into a grid of points indexed by longitude and latitude with the points equally spaced in each dimension. If we index longitude and latitude by i (i=1,2), let (g_{j_1}, g_{j_2}) be the grid point at longitude j_1 and latitude j_2, such that j_1 and j_2 are within the range of longitude and latitude that contains Baltimore City. We also define M_1 and M_2 such that (g_1, g_1) represents the point at minimum longitude and latitude and (g_{M_1}, g_{M_2}) represents the point at the maximum longitude and latitude.

2. Use linear binning to obtain a count c_{j_1,j_2} for grid point (g_{j_1}, g_{j_2}). The computation is illustrated by the figure 1 below. Suppose A, B, C, D, E, and F are grid points separated by one unit in each direction (point A has coordinates (0,0) and point D

has coordinates (2,1)). For each event of interest, we note its coordinates (in our example, we consider only one event represented by a hollow circle at (0.5, 0.5)). For each event, we assign a weight for each grid point. First, only the four closest grid points (A, B, C, F) to the event are considered, and the other points are automatically given a count of 0. A vertical and horizontal line are then drawn to divide up the rectangle defined by these four points into four quadrants, such that both lines pass through the location of the event. The score of each quadrant is defined as the percentage of the total area of the rectangle taken up by that quadrant, and the weight for each grid point is the score of the diagonal quadrant. In our example, grid point C will get assigned the weight of the quadrant filled in by black, which is 0.25. If the event had occurred at the coordinates (0.1, 0.1), point C would be assigned a weight of .01, and if the event had occured at (0.9, 0.9), grid point C would be assigned a weight of .81. For each event, this procedure is repeated, and the sum of these weights for each grid point is c_{j_1,j_2}.

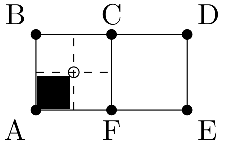

Figure 1: **Assigning weights to grid points with linear binning.** The event divides up the rectangle formed by the four closest grid points, A, B, C, and F, into four quadrants. Each of these four grid points is assigned a weight based on the percentage of the area of this rectangle taken up by the diagonal quadrant–for grid point C, it is given the percentage taken up by the shaded quadrant. In this example, there is unit distance between each grid point, and the event occurs at coordinates (0.5, 0.5), so grid point C is given weight 0.25.

3. Define h_1 and h_2 as the bandwidths for each dimension. Although the choice of bandwidth heavily influences kernel smoothing estimates, we arbitrarily use a bandwidth of .004° in each direction (corresponding to approximately 1/4 mile in each direction [4]). Letting K define the standard normal distribution and n the total number of grid points, the binned kernel density estimate at grid point (g_{j_1}, g_{j_2}) is:

$$\tilde{s}_k((g_{j_1}, g_{j_2})) = \frac{1}{nh_1h_2} \sum_{l_1}^{M_1} \sum_{l_2}^{M_2} K\left(\frac{g_{j_1} - g_{l_1}}{h_1}\right) K\left(\frac{g_{j_2} - g_{l_2}}{h_2}\right) c_{l_1,l_2} \tag{1}$$

Once a kernel density estimate is computed for each grid point, we then want to interpolate this density to the block centers, as computed by the *rgeos* package [5]. For each block center, we find the four closest grid points. The interpolated kernel density estimate for the block center is a weighted average of the kernel density estimates of the four grid points, with the weights being computed as described with the linear binning in step 2 above. Figure 2 demonstrates that the kernel density estimates accurately reflect whether or not a Census block is in an area that has had more shootings during 2014 and 2015.

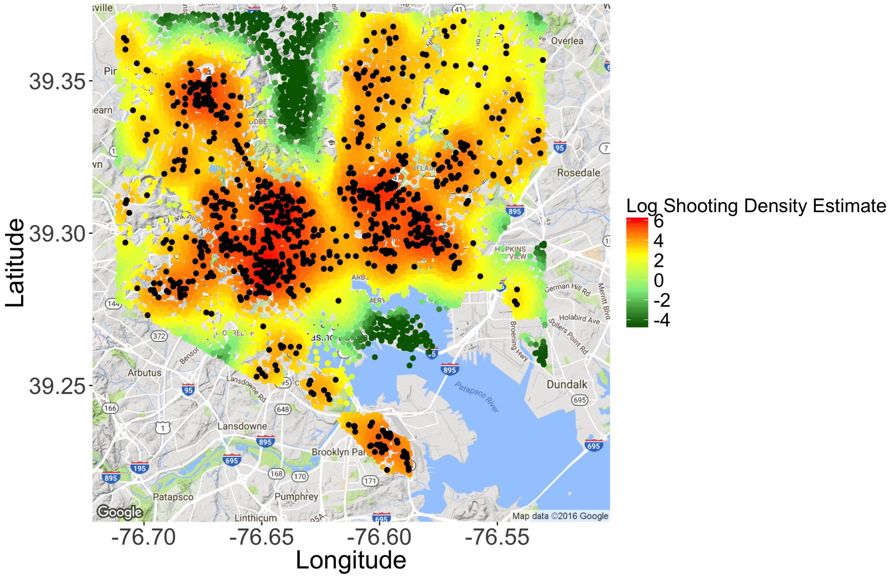

Figure 2: **Kernel density estimates allow location data to be quantified at Census blocks**. Each black dot represents the location of a shooting that occurred within Baltimore City during 2014 and 2015. The colored dots represent the centers of Census blocks, with green colored dots having a low log kernel density estimate of shootings, and red colored dots having a high log kernel density estimate of shootings. Although shootings have not occurred at every block, this figure shows the kernel density estimates at the Census block level reflects whether or not there is a high probability of a shooting at a given Census block

Linear model coefficient and standard error estimates

Coefficients	Estimate	Std. Error	P-value
(intercept)	48.07	19.11	**0.016**
income	2.68	1.86	0.157
segregation	-0.42	0.35	0.229
employment	-2.05	1.64	0.220
education	-2.67	0.84	**0.003**
vacancies	-0.83	0.71	0.247
shootings	0.18	0.54	0.746
burglaries	-0.61	1.50	0.689
fastfood	0.02	0.27	0.937
liquor	-0.13	0.54	0.815

Table 1: **Coefficient and standard error estimates for our final multivariate linear model**. Coefficients with p-values $< .05$ are in bold. In our model, we find our only statistically significant variable is the percentage of residents over 25 who have not received a high school degree or its equivalent. However, collinearity of predictor variables prevents meaningful interpretation of our results.

Correlation between predictor variables

	income	segregation	employment	education	vacancies	shootings	burglaries	fastfood	liquor
income	1.00	-0.70	0.85	-0.78	-0.63	-0.70	-0.33	-0.10	-0.21
segregation	-	1.00	-0.77	0.52	0.58	0.69	0.22	-0.14	0.07
employment	-	-	1.00	-0.68	-0.70	-0.72	-0.33	-0.14	-0.23
education	-	-	-	1.00	0.64	0.72	0.39	0.29	0.28
vacantcies	-	-	-	-	1.00	0.84	0.82	0.42	0.69
shootings	-	-	-	-	-	1.00	0.67	0.26	0.40
burglaries	-	-	-	-	-	-	1.00	0.55	0.80
fastfood	-	-	-	-	-	-	-	1.00	0.63
liquor	-	-	-	-	-	-	-		1.00

Table 2: **Pairwise correlations reveal collinearity between predictor variables**. Pairwise correlations each variable are listed in the entries. All variables are transformed with either the log or logit transformation, as described in Section 2.5, before the correlation is computed. Of interest is the high level of correlation between the percentage of residents over 25 who have not received a high school degree or its equivalent to the median household income, the percentage of people between 20-64 years of age that have worked in the past 12 months, and the density estimates of shootings and vacancies. Because we find the effect of education to be statistically significant, this may explain a large part of the variation in life expectancies that would have been attributable to other variables that have a high pairwise correlation with life expectancy.

Supplementary Figure S1

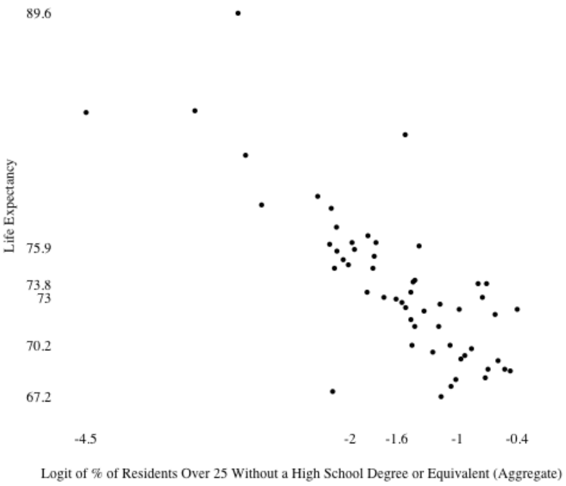

Figure 3: **CSA aggregate of education variable is linearly correlated with life expectancy**. The logit transformation of the CSA aggregate of the percentage of residents over 25 who have not received a high school degree or its equivalent versus true CSA level of life expectancy. We observe a linear relationship between the only statistically significant predictor and life expectancy at the CSA level, making standard linear regression a reasonable statistical model

Supplementary Figure S2

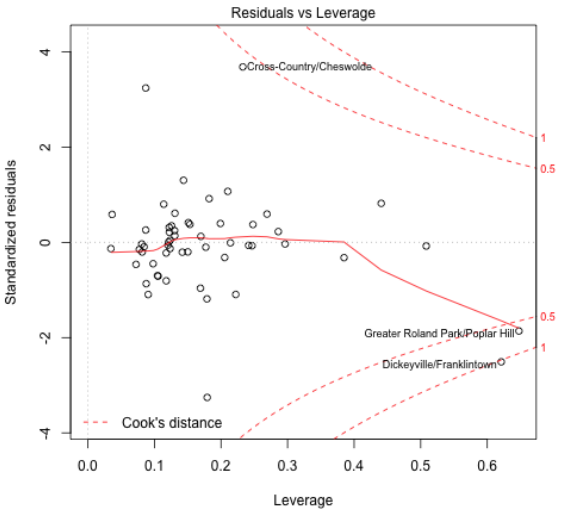

Figure 4: **CSAs at the borders of Baltimore City have high leverage and influence**. Leverage of the CSA level data in our linear model versus the standardized residuals. Cook's distance drawn in red dashed lines. Cross-Country/Cheswolde appears to have high influence, but low leverage. Greater Roland Park/Poplar Hill and Dickeyville/Franklintown both appear to be high leverage and high influence CSAs.

References

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

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

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

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

[5] 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.