Hipp and Boessen (2013) defined their “egohoods” starting with each block in a city surrounded by a buffer of fixed radius. The buffer around any block contains blocks, each of which have buffers in which the value of any characteristic is measured. The value for the block-based egohood is the sum or mean of the buffer values for all of the blocks falling within the buffer surrounding it. They rather dramatically described these as being “waves washing across the city.” They acknowledged that their approach “implies a social process with an inherent spatial decay function,” (p. 296) but that is as far as they went in pursuing this. However their egohood values are nothing more and nothing less than spatial kernel density estimates of values for the blocks using a distance-decay function implicitly determined by their methodology. Understanding this leads to the identification of certain problems and limitations of their methodology.

**Egohoods, Kernel Density Estimation, and Distance-Decay Functions**

Spatial kernel density estimation can assign a value to each point on a surface that is the sum or mean of all observations within a circular buffer around the point, with the values of the observations weighted by some declining function of distance. The “kernel” is the function that decreases in value from 1 at the center of the buffer at the point being estimated to 0 at the edge of the buffer. A variety of functions are used for spatial kernel density estimation.

The Hipp and Boessen approach uses blocks as the fundamental units. To illustrate how their method involves distance decay and is equivalent to kernel density estimation, a simple example will use a regular system of equal square grid cells in place of blocks. The grid cells are assumed to have a length of 1 unit. A buffer radius of 2.5 units is specified, so that a grid of 5 by 5 cells just encompassed the buffer around the middle cell. This is illustrated in the simple diagram, with the distances from the center of each cell to the center of the middle cell shown in each cell. The centers of all of the cells except the 4 corners are within the 2.5 mile buffer. Therefore, in the Hipp and Boessen procedure, the values of these cells within the buffer would be included in calculating the value of the buffer for the center grid cell. And the egohood value for the center cell includes the values of all of the buffers for those cells within the buffer surrounding the egohood block/cell at the center, the buffers around each of the cells within the shaded area.
Each grid cell will have a value derived from the buffer values of all of the cells included within its buffer. The egohood value for a cell is then the mean or sum of the buffer values for all of the cells included within the buffer around the cell to which the egohood value is to be assigned. As a result, the value of any cell can contribute to the final value multiple times, once for each buffer within which it is located.

An expanded diagram illustrates how this process, which assigns different weights to the cell values, can begin. The cells enclosed in the heavy black line are the cells within the buffer around the central cell for which the egohood value is to be calculated. The value of that buffer is the mean of those cell values, each given a weight of 1 as indicated by the numbers in those cells. That buffer value for that buffer will be included in the egohood value for the central cell.

But the buffer values for all of the other cells within that central buffer are also included in the egohood value. Start by adding the value for the buffer for the grid cell in the upper left part of the central buffer. That grid cell is shown in orange. The cells in yellow are those cells included in the buffer around the orange grid cell. It necessarily has the same pattern as the buffer for the central grid cell. All of the cells in that buffer, in yellow or orange, will be included in calculating the egohood value for the central cell, again, each with an equal weight of 1. So all of those cells in the yellow and orange area would have their weights increased by 1 as a result of including the value for this second buffer. The weights for those colored cells outside the central buffer will be increased from 0 to 1, and the weights for those colored cells within the central buffer will be increased from 1 to 2.

This process is repeated for the buffers for all of the remaining grid cells within the buffer around the central cell. All of the grid cells within the central buffer are within the buffer distance of the central cell so the central cell will be within the buffers around all of cells within the central buffer. There are 21 cells within the buffer, so the central cell will have a weight of 21. As one moves to cells away from the center, the number of buffers including the cell will decline and its weight will decline. But cells will be included in buffers around cells within the central buffer out to a distance of about twice the buffer radius, for buffers of cells at the periphery of the central buffer.
This diagram shows the weights for all of the grid cells. It is obvious how the weight given to the value in a cell in calculating the egohood value for the central cell decreases in a regular fashion with distance from the center. Therefore, the egohood value can be seen as a kernel density estimate of the value for the cell.

We can calculate the distance each cell is from the center and figure the total weight associated with each distance. These weights are plotted against distance to show the distance-decay function, with distances normalized to a buffer radius of 1 and weights normalized for a central weight of 1. This distance-decay function is not smooth, with the weights limited to a small number of integer values by the nature of the small array of grid cells for which they were calculated.

The next step is the derivation of the continuous distance decay function implied by the Hipp and Boessen buffer-in-buffer approach. This uses infinitely small grid cells or, more correctly, finds the limit as the size of the grid cells approaches zero. In the limit, the buffer and the system of grid cells have rotational symmetry. Therefore, the distance decay function will be identical for any line extending from the center of the buffer, so we only need to determine the function along one line.

We start with a buffer around the central grid cell with a radius of 1. For any area within the buffer and for any size grid cell, the number of grid cells and hence the number of buffers surrounding those grid cells will be proportional to the area within the buffer. This will hold in the limit as the size of the grid cells approach zero.
The buffers for all of the grid cells in the buffer will include the grid cell in the center. The number of all grid cells and buffers is then proportional to the area of the buffer. Since the radius of the buffer is 1, the area of the grid cell is $\pi$, and the number of buffers including the central grid cell and therefore the weight is proportional to $\pi$.

Now consider a point on the line, indicated by the vertical bar, at distance $s$ from the center, indicated by the black dot. The right circle is the original buffer around the central grid cell. The left circle around the point at the vertical bar also has a radius of 1. All of the cells within that circle would have buffers that would contain the point at the vertical bar at the center, but only those cells and buffers in the area of the intersection of the two circles fall within the original buffer and would contribute to the value of the egohood at the center. So the weight at the point of the vertical bar at distance $s$ from the center would be proportional to the area of the intersection of the two circles with centers separated by distance $s$, the slightly darker area. This continues to hold for locations farther away from the center along the line, outside of the original buffer, out to a distance of 2, twice the radius of the circles. In the diagram, the circle on the left is at a distance $s^*$ from the center. All of the cells and their buffers within the area of intersection will include the point on the vertical bar at the distance $s^*$, so the weight at that point would again be equal to the area of intersection.

The weight as a function of distance $s$ from the center will therefore be proportional to the area of intersection of two circles of radius 1 with their centers separated by the distance $s$. The formula for the area of intersection of two unit circles with centers at distance $s$ is as follows:

$$A = 2 \cos^{-1}\left(\frac{s}{2}\right) - \frac{s}{2} \sqrt{4 - s^2}$$

Since the weights are proportional to this area, this is the formula for the distance decay function. The weight at the center will be the area of the circle $\pi$, which will decline with $s$ to a value of 0 at $s = 2$. The weights are again normalized so the weight at the center, at distance 0, is equal to 1.
The chart graphs the distance decay function. This distance decay function is a smooth curve, similar to the discrete function calculated for the buffer within the 5 by 5 set of grid cells. It is the continuous function implied by the buffers-in-buffers method for estimating the egohood values.

Therefore, the Hipp and Boessen method of estimating values for their egohoods is nothing more and nothing less than spatial kernel density estimation of the values using this specific distance decay function for the kernel. The kernel has a radius which is twice the radius specified for the buffers.

**Problems and Limitations of the Egohood Approach**

The Hipp and Boessen buffers-in-buffers method is used for estimating egohood values for each block for both the independent variables and the dependent (crime) variables. The dependent variable is therefore a kernel density estimate of crime in the area surrounding the block within a distance of twice the buffer radius. The use of these smoothed crime values as the dependent variables raises major problems for the estimation of the models and their standard errors.

Spatial correlation is a common problem in spatial analyses due to the natural tendency of nearby observations in space to have similar values. This violates the assumption of independence of the observations for statistical estimation and can lead to the underestimation of the standard errors.

The Hipp and Boessen egohood measures of crime introduce spatial autocorrelation to an extreme extent by their construction. Nearby blocks will share nearly the same sets block and buffers, with very similar weights for the blocks. Or using the terminology of kernel density estimation, nearby blocks will have values calculated for nearly the same areas. The results will be extremely high correlations of egohood crime values for blocks located in close proximity, correlations that are the direct product of the manner in which the values have been estimated. This will likely produce underestimates of the standard errors for the models and heightened levels of significance.
Hipp and Boessen acknowledged that the egohood approach introduces spatial autocorrelation and responded by estimating spatial error models predicting the logs of crime rates. Given the extreme levels of spatial autocorrelation introduced in both the independent and dependent variables by their approach, it is questionable whether any available method for accounting for the spatial autocorrelation can fully eliminate the problem. In addition, given that the autocorrelation is introduced into the dependent variable by an explicit spatial lag process, the choice of spatial error over spatial lag models might be questioned.

More generally, since the egohood crime measures are simply kernel density estimates of crime, it is difficult to see the justification for using these as the dependent variable. The kernel density estimation smooths the crime pattern over space, eliminating some of the random variation. Thus it would not be surprising for this to lead to increased fit of models and higher levels of significance. Just as aggregating spatial data to larger units leads to better fit and greater significance, smoothing—which can be seen as a form of partial aggregation—would be expected to have a similar effect.

Using the kernel density estimates of crime as the dependent variable also introduces a logical paradox into the analysis. The egohood level of crime in a particular block includes not only crime in that block but crime in many other blocks, out to twice the buffer radius away, albeit with declining weights. So the independent variable for that block’s egohood are assumed to be affecting not only levels of crime in that block but also levels of crime in all of those other blocks. But each of those other blocks also has its own egohood with its own egohood independent variable characteristics. And those values are (also) assumed to affect levels of crime in that block. And if that block lies within the buffer of the first block, those independent variable values are also assumed to affect the levels of crime in the first block as well. So which values of the independent variables are affecting crime in any block? It seems that many of them are, to varying degrees. It is difficult to know what might be affecting what and how to interpret any results.

Hipp and Boessen compared their egohood approach with models they refer to as using individual social environments, with the independent variables estimated for buffers using a distance decay function and the dependent variables measured in a smaller area at the center of each buffer. Since the egohood values have been shown to be kernel density estimates using a distance decay function, the independent variables used in the two approaches are basically similar. The differences are in the distance decay function used in making the estimates and, for their direct comparisons, the sizes of the kernels used for the estimation. Hipp and Boessen chose to use buffer distances for the individual social environment estimates (the kernel radii) that are the same as the buffer distances used in creating the egohoods. But as shown above, the egohood buffer radius is one-half the radius of the kernel for the kernel density estimation. A strict comparison would have used twice the egohood buffer radius as the size of the kernel buffer for the individual social environments.
In describing their egogood method, Hipp and Boessen distinguish it from the individual social environment approach by saying that the latter only includes the effect of the buffer, while the egohoods consider the larger area beyond the buffer. Of course this is true if the buffer for the individual social environment kernel estimation is the same size as the egohood buffer. But given that both are using kernel density estimation and a distance decay function to estimate the values, and given that the size of the kernel for the egohood approach is twice the radius of the buffer, such a statement is quite irrelevant and nonsensical. Yes, the egohood estimate includes a larger area if the radius of the kernel is larger than that used for the individual social environment estimate. But if the latter is selected to be the same size as the egohood kernel, i.e., twice the egohood buffer radius—the egohood kernel radius—areas included would be identical.

One claim that Hipp and Boessen emphasized for their approach was that it somehow captured heterogeneity more effectively than alternative methods. Because the egohood value is equivalent to a kernel density estimate, it seems difficult to see how it could better reflect heterogeneity than any other kernel density estimate. The distance-decay function would be different from that used for other estimates, but it is unlikely that this would play a significant role in altering the extent to which the estimate incorporates heterogeneity.¹

Hipp and Boessen supported their claim that the egohood approach is better for identifying heterogeneity by pointing to the significance of the income inequality variable in their models. Two factors could be contributing to the significance in these models and not in the individual social environment models. The radii of the kernels implied by the egohood estimation—twice the buffer radius—are twice the radii of the buffers and kernels used in estimating the independent variables including income inequality in the individual social environment models. The egohood kernels therefore cover 4 times the area and the larger area is more likely to have greater heterogeneity, therefore creating greater variation in the income inequality variable. In addition, as discussed above, the egohood approach uses kernel density estimation for the dependent variable crime while the individual social environment models do not. This could be contributing to more spatial autocorrelation and greater underestimation of the standard errors, producing higher levels of significance.

**Units for Spatial Analysis of Crime or “What Is a Neighborhood?”**

With their creation of the term “egohood,” Hipp and Boessen obviously intend their units to be considered as a new, alternative definition of a neighborhood. Their focus is on the aspect of the traditional neighborhood definition as being those areas in which relatively

---

¹ One might expect to find greater heterogeneity in areas of the kernel at greater distances from the center. The area-of-circle-intersection function implied by the egohood method declines more rapidly with distance than many of the functions used for kernel density estimation, which are often highly convex downwards. The latter functions would therefore give relatively more weight to observations at greater distances from the center and might therefore actually do better at capturing heterogeneity.
higher levels of social interaction occur. They are discarding the notion of neighborhoods as non-overlapping units with fixed boundaries.

In creating their alternative concept of the neighborhood, Hipp and Boessen retained the assumption and practice from research going back at least to the Chicago School that a dependent variable in a neighborhood, such as crime, is affected by the values of the independent variables for social characteristics for the same neighborhood. This of course made sense when neighborhoods were bounded, non-overlapping areas. When one moves to considering the effects of overlapping areas, it is no longer clear that the areas for the measurement of the independent and dependent variables should be the same. Hipp and Boessen, however, do make that assumption and use the egohood values resulting from kernel density estimation for both the independent and dependent variables. This leads to increased spatial autocorrelation in the dependent variable and the paradoxes in causality discussed earlier.

The assumption is made here that using dependent variables measured using kernel density estimation or in any other way defined for overlapping areas is not appropriate. Instead, the dependent variable should be measured using a system of non-overlapping areas.

In discussing the individual social environment approach, with the buffers for the independent variables surrounding the smaller areas in which the dependent variables are measured, Hipp and Boessen describe the buffer as being the neighborhood. This is the area in which the social interaction is assumed to occur that affects crime in the smaller area.

Alternatively, with the individual social environment approach, the smaller, inner areas that are not overlapping and in which crime is measured could be considered to be the neighborhoods. The assumption would now be that social interaction affecting crime occurs beyond the boundaries of the neighborhood. This would reduce the focus on the interaction component of the traditional neighborhood definition and retain the aspect of neighborhoods as being bounded, non-overlapping areas.

Note that the smaller, non-overlapping units could be assumed to be areas in which higher levels of interaction occurred than the levels of interaction with the areas outside in the larger units. Hipp and Boessen actually argued that blocks constitute such units in justifying their use of blocks as the basic elements for their approach. Social characteristics of the smaller units could then also be considered to be relevant to levels of crime along with the social characteristics of the larger units. (In this case, one might consider the social characteristics of that portion of the larger unit outside the smaller unit in place of the social characteristics of the entire area.)

Then if the larger, overlapping units are to be considered the neighborhoods (following Hipp and Boessen), the smaller units might be called the inner neighborhoods, with the remainder being the outer neighborhoods. Or the smaller units could be considered to
be the neighborhoods, with the areas outside being called the extended neighborhoods (or just the areas outside!).

Perhaps best would be to discard the stretching of the idea of neighborhood beyond its original meaning and just describe the areas being used in the analysis without using the term “neighborhood.”

So we have a system of smaller contiguous, non-overlapping areas in which the dependent variable such as crime is measured and in which other characteristics of those areas might or might not be included as independent variables. This is combined with a system of larger, overlapping units surrounding each of those smaller units, the characteristics of which are thought to also affect levels of the dependent variable in the smaller units.

Each system of units could be either systems of uniform, regular units or irregular units defined in some other fashion. The individual social environment approach as implemented by Hipp and Boessen used blocks—irregular units—as the smaller, non-overlapping units, and circular buffers around the blocks—regular units—as the larger, overlapping units. But the smaller, non-overlapping units could be regular grid cells, surrounded by the overlapping circular buffers. This is exactly what we did in earlier papers for which I was a coauthor—land use, sex offenders, and foreclosures, with multiple buffers with different radii in the second foreclosure paper (Payton, Stucky, and Ottensmann 2016; Stucky and Ottensmann 2009, 2016; Stucky, Ottensmann, and Payton 2012). Finally, one could use irregular units for both the smaller, non-overlapping areas and for the larger, overlapping areas. The immediate example from our research would be census tracts as the smaller areas and the tracts plus the contiguous tracts as the larger, overlapping areas in a subsequent paper on crime and income inequality (Stucky, Peyton, and Ottensmann 2016).

The choice of units would depend upon both theoretical concerns and methodological considerations that may in part depend upon the nature of the data. For the sex offenders paper, since distance was the measure restricting the location of sex offenders with respect to certain facilities and areas, circular buffers using distance were dictated by the problem being addressed. And having the original sex offender data as point data made possible the direct estimation of values for the buffers. Likewise, having the foreclosures as point data made the use of regular circular buffers obvious. And the second foreclosure paper, considering the effect of distance on the foreclosure-crime relationship, necessarily required a series of regular circular buffers of varying size. In all of these cases, since the crime data were also point data, regular grid cells could be used for the smaller, non-overlapping areas.

The crime and income inequality data presented a very different situation. For this research, socioeconomic status, ultimately income, was to be used to create a variety of different measures of level and variation. Here, the nature of the data played a major role in dictating the units to be used. The income data would be coming from the American Community Survey and were available at both the census tract and block
group levels. Because of the critical nature of the income data—we would be using the counts of households in each of the income categories—sampling error was a significant concern. For this reason, the first choice was made to use census tracts as the source the income data (and then the other, related data) to minimize the sampling error. Relatively complex methods were to be used in creating some of the measures such as the Gini coefficients, which required estimation of the mean income level for each income category. For this reason, any consideration of estimating the counts for the income categories for a regular system of units was immediately rejected, as the error introduced by such estimation would have had unknown effects on the measures ultimately estimated. Therefore, the census tracts were used as the smaller, non-overlapping areas, with the tracts plus the queen-contiguous tracts being used as the larger areas.

References